Analyse d'incertitude d'un modèle de culture : démarche et illustration sur deux cas d'étude

François Brun, Nathalie Keussayan, Arnaud Bensadoun, Jacques-Eric J.-E. Bergez, Bernard Lacroix, Philippe P. Debaeke, Luc Champolivier, Jean-Pierre Palleau, Emmanuelle Mestries, Daniel D. Wallach

12èmes Journées Européennes
AGRO-INDUSTRIE ET METHODES STATISTIQUES (AGROSTAT 2012)

12th European Symposium on STATISTICAL METHODS FOR THE FOOD INDUSTRY

Paris, France
28, 29 février et 1, 2 mars 2012
February 28, 29 th and March 1st, 2nd 2012

Organisées à / *Organized at*
Institut des Sciences et Industries du Vivant et de l'Environnement, AgroParisTech

Pour le / *For the*
Groupe Agro-Industrie de la Société Française de Statistique

Bienvenue à AGROSTAT 2012 !

C'est avec grand plaisir que les Comités Organisateur et Scientifique des Journées Européennes Agro-Industrie et Méthodes Statistiques vous accueillent dans cette belle ville de Paris. Nous sommes heureux d'avoir pu continuer notre effort d'ouverture vers nos collègues européens. Ainsi, cette année, tous nos conférenciers invités sont européens, et non français ! Ceux-ci, nous en sommes sûrs, permettront de mieux faire connaître l'existence de notre Congrès en dehors de la France. Cette année, les journées sont organisées en l'honneur de André Kobilinsky, Directeur de Recherche à l'INRA, contributeur majeur au domaine des plans d'expériences tant sur l'aspect théorique que sur l'aspect applicatif dans le secteur de l'agro-alimentaire.

Nous remercions le directeur général d'AgroParisTech - le nom officiel est l'Institut des sciences et industries du vivant et de l'environnement - le Professeur Gilles Trystram, de nous accueillir au sein de cette prestigieuse Grande École, dont le site rue Claude Bernard abrita dès 1882 l'Institut National Agronomique, berceau historique de l'enseignement et de la recherche agronomique de haut niveau en France.

Nous remercions aussi la Société Française de Statistique (SFdS) pour avoir permis l'organisation de cet évènement. Plusieurs partenaires et exposants ont aussi accepté d'y apporter leur soutien financier, ce sont : Addinsoft France, Adria Developpement, AgroParisTech, Biosystemes, Coheris, Danone, la Société Française de Biométrie, le département MIA de l'INRA et Sigma Plus. Tout particulièrement nous remercions chaleureusement la société Addinsoft France qui a proposé de décerner un prix XLSTAT (1000 euros) destiné à valoriser la qualité du travail de recherche d'un jeune statisticien (doctorant ou post-doctorant).

Un numéro spécial du Journal de la SFdS sera consacré à des articles scientifiques issus des communications présentées, sélectionnées par le Comité Scientifique après le Congrès.

Nous souhaitons enfin à tous les participants des échanges scientifiques fructueux, mais également bien évidemment un agréable séjour à Paris, séjour qui sera tout particulièrement marqué par la visite de ce si magnifique musée qu'est le musée d'Orsay.

Isabelle Albert, Douglas Rutledge et Jean-Pierre Gauchi
Pour les Comités Organisateur et Scientifique d'AGROSTAT 2012

Welcome to AGROSTAT 2012 !

It is with great pleasure that the Organizing and Scientific Committees of the European Symposium on Statistical Methods for the Food Industry welcome you to the beautiful city of Paris. We are pleased to have been able to continue our efforts of opening up towards our European colleagues. Thus, this year, all of our guest speakers are European, and none are French ! We are sure that they will increase the renown of our Congress outside the France. This year, the Congress is organized in honor of André Kobilinsky, Research director at INRA, a major contributor to the field of experimental designs - both the theoretical aspects and their application in the agri-food sector.

We thank Professor Gilles Trystram, the Director General of AgroParisTech - the official name of which is the Institute of life sciences and industries and of the environment - for welcoming us to this prestigious "Grande École", whose campus rue Claude Bernard housed the "Institut National Agronomique", historic cradle of high-level education and agronomic research in France since 1882.

We also thank the French statistical society (SFdS) for its contribution in the organization of this event. Several sponsors and exhibitors have also agreed to provide financial support, these are : Addinsoft France, Adria Developpement, AgroParisTech, Biosystemes, Coheris, Danone, la Société Française de Biométrie, le département MIA de l'INRA and Sigma Plus. Especially we warmly thank the Addinsoft France company which has offered the XLSTAT award of 1000 euros to recompense the quality of the research work of a young statistician (PhD student or post-doc).

A special issue of the Journal of the SFdS will be devoted to a number of scientific articles to be selected by the Scientific Committee after the Congress from among the communications presented.

Finally, to all participants we wish fruitful scientific exchanges, but also of course a pleasant stay in Paris, a stay which will be particularly marked by the visit of the particularly beautiful "Musée d'Orsay".

Isabelle Albert, Douglas Rutledge and Jean-Pierre Gauchi
For the Organizing and Scientific Committees of AGROSTAT 2012

Comité scientifique/Scientific Committee

Président d'honneur/Honorary chairman : André Kobilinsky (INRA, France)

Président/Chairman : Jean-Pierre Gauchi (INRA, Jouy-en-Josas, France)

- Christophe Boulais (Danone, France)
- Philippe Courcoux (ONIRIS, Nantes, France)
- Marie-Laure Delignette-Muller (VetAgro Sup, Lyon, France)
- Jean-Baptiste Denis (INRA, Jouy-en-Josas, France)
- Bernadette Govaerts (STAT-Université catholique de Louvain, Belgique)
- Francois Husson (Agrocampus, Rennes, France)
- Jan Van Impe (Biotec, Katholieke Universiteit, Belgique)
- Joachim Kunert (Dortmund University, Germany)
- Riccardo Leardi (Genoa University, Italy)
- Hervé Monod (INRA, Jouy-en-Josas, France)
- Robert Sabatier (Université de Montpellier 1, France)
- Pascal Schlich (INRA, Dijon, France)
- Michèle Sergent (UPCAM, Marseille, France)
- Arthur Tenenhaus (Supélec, Gif-sur-Yvette, France)

Comité organisateur/Organizing Committee

Présidents/Chairmen :
- Isabelle Albert (INRA, France)
- Douglas N. Rutledge (AgroParisTech, France)

- Sophie Ancelet (IRSN, Fontenay-aux-Roses, France)
- David Blumenthal (AgroParisTech, Massy, France)
- Delphine Bouveresse (INRA/AgroParisTech, Paris, France)
- Christophe Cordella (INRA/AgroParisTech, Paris, France)
- Clémence Rigaux (INRA, Paris, France)
- Anne Saint-Eve (AgroParisTech, Grignon, France)
- Eric Teillet (AgroParisTech, Massy, France)

Équipe administrative/Administrative team :

Association ADEPRINA - *Contacts :* Sylvain Lisembard, Dominique Lefrançois
Adresse postale/ Mailing address : ADEPRINA, 16 rue Claude Bernard, 75005 Paris - France
Téléphone/Phone : +33 (0)1 44 08 18 37 - *Fax :* +33 (0)1 44 08 18 70
Adresse email/e-mail address : adeprina@agroparistech.fr

Site Web/Web Site : http ://www.chimiometrie.fr/agrostat2012/index.html

Christophe Cordella

Création, couverture des actes/Proceedings compilation, graphic design :

Sophie Ancelet, Isabelle Albert, Jean-Pierre Gauchi, Martial Guisnet (AgroParisTech)

TABLES DES MATIÈRES / *TABLES OF CONTENTS*

Mercredi 29 janvier 2012 / *Wednesday, January 29th, 2012*

Vendredi 2 mars 2012 / *Friday March 2nd, 2012*

Session 1 : Plans d'Expérience / *Experimental Designs*

Design of experiments with very low average replication

Rosemary A. Bailey

Queen Mary, University of London.
E-mail: *r.a.bailey@qmul.ac.uk*

Abstract
Trials of new crop varieties usually have very low average replication. Thus one possiblity is to have a single plot for each new variety and several plots for a control variety, with the latter well spread out over the field. A more recent proposal is to ignore the control, and instead have two plots for each of a small proportion of the new varieties.

Variation in the field may be accounted for by a polynomial trend, by spatial correlation, or by blocking. However, if the experiment has a second phase, such as making bread from flour milled from the grain produced in the first phase, then that second phase usually has blocks. The optimality criterion used is usually the A criterion: the average variance of the pairwise differences between the new varieties. I shall compare designs under the A criterion when the average replication is much less than two.

Keywords: design of experiments, replication, crop varieties.

Sur le compromis entre l'effort expérimental et le contenu des informations dans le design optimal de l'expérimentation

On the trade-off between experimental effort and information content in optimal experiment design

Dries Telen, Filip Logist, Eva Van Derlinden & Jan Van Impe

BioTeC & OPTEC - Chemical and Biochemical Process Technology and Control
Department of Chemical Engineering, Katholieke Universiteit Leuven
W. de Croylaan 46, B-3001, Leuven, Belgium
E-mail : *jan.vanimpe@cit.kuleuven.be*

Résumé
Dans la microbiologie prédictive, des modèles mathématiques et dynamiques sont développés pour décrire l'évolution microbienne en fonction des conditions environnementalles. Après la sélection un modèle acceptable, les valeurs fiables des paramètres sont nécessaires pour obtenir des prédictions valides du modèle. Pour obtenir des estimations précises des paramètres du modèle, quelques expérimentations devraient être effectuées. Les techniques du design optimal des expérimentations pour l'estimation des paramètres sont indispensables pour limiter l'effort expérimental. Une question importante concernant le design optimal de l'expérimentation est le schéma optimal d'échantillonnage. Une publication récente illustre que le schéma optimal d'échantillonnage rend un schéma bang-bang. Dans cette contribution, l'optimisation du schéma d'échantillonnage en utilisant une approche bang-bang est implémentée. Un deuxième point de ce travail concerne le compromis entre l'effort expérimental et le contenu informative dans le design optimal de l'expérimentation.
Mots-clés : le design optimal de l'expérimentation, optimisation multi-objectif, l'estimation des paramètres

Abstract
In predictive microbiology, dynamic mathematical models are developed to describe the microbial evolution under time-varying environmental conditions. Next to an acceptable model structure, reliable parameter values are necessary to obtain valid model predictions. To obtain these accurate estimates of the model parameters, labour and cost intensive experiments have to be performed. Optimal experiment design techniques for parameter estimation are invaluable in order to limit the experimental burden. An important issue in the optimal experiment design process is the sampling scheme. Recent work illustrates that identifying the sampling decisions results in bang-bang control of the weighting function in the Fisher information matrix. A second point addressed in this work will be the trade-off between the amount of time an experimenter has available for measurements on the one hand, and the information content on the other hand.
Keywords : optimal experiment design, multi-objective optimisation, parameter estimation

1 Introduction

Dynamic models that predict the microbial behaviour in food products can play an important role in food quality and safety. In these models, the effect of temperature, pH, water activity and preservatives are important elements. When a suitable model structure is chosen, reliable parameter estimates have to be obtained for the different microbial strains. However, performing experiments are cost and labour intensive: one has to take samples and subsequently perform an analysis. To reduce this experimental effort, optimal experiment design (OED) is developed. In optimal experiment design for parameter estimation some scalar function of the Fisher information matrix [1, 2] is employed as objective function in the resulting dynamic optimisation problem. Design of an optimal experiment can focus on different aspects. An important aspect is the sampling scheme, i.e., when does the experimenter has to take a sample. Recent work [3] indicates that the control of the weighting function in the Fisher information matrix results in bang-bang control. In this contribution, optimisation of the sampling scheme using this bang-bang approach is implemented with respect to the accurate estimation of the parameters of the Cardinal Temperature Model with Inflection. In this way the most interesting parts of the experiments can be identified. A second point addressed in this work is the trade-off between the amount of time an experimenter has available for measurements on the one hand, and the amount of information available in the experiment on the other. These are two different and conflicting objectives. When these objectives are combined in a multi-objective approach, a set of optimal solutions, i.e., the so-called Pareto set is obtained. This allows the experimenter to make a decision based on his/her preferences.

The paper is structured as follows. In section 2, the employed mathematical techniques and the mathematical model are discussed. Section 3 presents the obtained results. Following section 3, the conclusion is formulated.

2 Model and Methods

The techniques of optimal experiment design are discussed in the first part of this section. In the second part, the specific methods of multi-objective optimisation are described. This section concludes with the description of the case study.

2.1 Optimal experiment design

In optimal experiment design for parameter estimation, some scalar function of the Fisher information matrix is used as the objective function. This matrix is defined as:

$$\mathbf{F}(\mathbf{p}) = \int_0^{t_f} \left(\frac{\partial \mathbf{y}(\mathbf{p},t)}{\partial \mathbf{p}}\right)^T \Big|_{\mathbf{p}=\mathbf{p}^*} \mathbf{Q} \left(\frac{\partial \mathbf{y}(\mathbf{p},t)}{\partial \mathbf{p}}\right) \Big|_{\mathbf{p}=\mathbf{p}^*} dt \tag{1}$$

The true values $\mathbf{p}^*$ are unknown so the Fisher matrix depends on the current best estimate. Two parts constitute the Fisher information matrix: the inverse of the measurement error variance-covariance matrix, $\mathbf{Q}$ and the sensitivities of the model output to small variations in the model parameters, $\frac{\partial \mathbf{y}(\mathbf{p},t)}{\partial \mathbf{p}}$. The latter can be found as the solution to:

$$\frac{d}{dt}\frac{\partial \mathbf{y}(\mathbf{p},t)}{\partial \mathbf{p}} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}(\mathbf{p},t)}{\partial \mathbf{p}} + \frac{\partial \mathbf{f}}{\partial \mathbf{p}} \tag{2}$$

An interesting property of the Fisher information matrix is that under the assumption of unbiased estimators and uncorrelated Gaussian noise, the inverse of $\mathbf{F}$ approximates the lower bound of the parameter estimation variance covariance matrix, which is the Cramér-Rao lower bound. The most widely used scalar functions are [1, 2, 4]:

A-criterion: min[trace($\mathbf{F}^{-1}$)] A-optimal designs minimise the mean of the parameter estimation errors. The geometrical interpretation is the minimisation of the enclosing frame around the joint confidence region. In order to decrease the computational effort, the problem is reformulated as maximising the trace of $\mathbf{F}$.

D-criterion: max[det(F)] D-optimal designs minimise the geometric mean. Geometrically, this is minimising the volume of the joint confidence region.

E-criterion: max$[\lambda_{min}(\mathbf{F})]$ E-optimal designs aim at minimising the largest parameter error. Uncertainty regarding other parameters is thus neglected. This corresponds to minimising the length of the largest uncertainty axis of the joint confidence region.

In [3] an additional control function $w(t) \in [0, 1]$ is introduced in the Fisher information matrix. This function controls whether the experimenter has to measure are not. When the measurement time is not constrained to a certain value, the maximal information content is found by measuring continuously during the experiment time. As performing experiments is cost and labour intensive the following, additional objective is taken into consideration: **min** W in which $\frac{dW}{dt} = w(t)$ with $W(0) = 0$. This objective represents the experimental effort in the sense that it represents the amount of hours one has to spend. If one is not interested in obtaining an complete overview using a multi-objective approach, this objective can also be formulated as a constraint. Where $W(t_f) \leq \alpha$ with α a predetermined number of measurement hours .

2.2 Multi-objective optimisation

The general formulation of a dynamic optimisation problem involving multiple objectives can be described as:

$$\min_{\mathbf{x}(t),\mathbf{u}(t),\mathbf{p},t_\mathrm{f}} \{J_1(\mathbf{z}), \ldots, J_m(\mathbf{z})\} \tag{3}$$

subject to

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}, t) \quad t \in [0, t_\mathrm{f}] \tag{4}$$

$$\mathbf{0} = \mathbf{b}_c(\mathbf{x}(0), \mathbf{p}) \tag{5}$$

$$\mathbf{0} \geq \mathbf{c}_p(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}, t) \tag{6}$$

where $\mathbf{x}$ are the state variables, $\mathbf{u}$ the time-varying control inputs, including the additional controls whether to measure or not and $\mathbf{p}$ the model parameters. The vector $\mathbf{f}$ represents the dynamic system equations (on the interval $t \in [0, t_\mathrm{f}]$) with initial conditions given by the vector $\mathbf{b}_c$. The vector $\mathbf{c}_p$ indicates path inequality constraints on the states and controls. Furthermore, the vector $\mathbf{y}$ is introduced, which contains the measured outputs. These are usually a subset of the state variables $\mathbf{x}$. The decision variables can be grouped together as $\mathbf{z} = [\mathbf{x}(t), \mathbf{u}(t), t_f)]$. Note that $\mathbf{p}$ is no longer a degree of freedom, but considered known. All vectors $\mathbf{z}$ that satisfy Equations (4) to (6) form the set of feasible solutions S.

As concept for optimality in multi-objective optimisation, *Pareto optimality* is used (see, e.g., [5]). A solution is called Pareto optimal if there exists no other feasible solution that improves at least one objective function without worsening another. The decision variables $\mathbf{z}$

belong to the feasible set $\mathcal{S}$, defined previously, and the vector of all individual cost functions is defined as $\mathbf{J} = [J_1(\mathbf{z}), \ldots, J_m(\mathbf{z})]^{\mathrm{T}}$.

Methods for generating Pareto sets can be classified into two classes: (i) methods converting the MOO problem into a series of parametric single objective optimisation (SOO) problems (e.g., weighted sum, ...), and (ii) methods tackling directly the MOO problem (e.g., stochastic evolutionary algorithms [6]). However, since methods of the latter class require a repeated evaluation of the objectives (and, thus, also of the underlying models), they can become time consuming for the class of systems under study. Therefore, only deterministic techniques from the first class will be studied, because they allow using fast, gradient-based techniques.

2.2.1 Weighted Sum (WS)

The (convex) weighted sum is in practice the most widely used technique for combining different objectives. The resulting parametric SOO problem is the following :

$$\min_{\mathbf{z}\in\mathcal{S}} J_{ws} = \sum_{i=1}^{m} w_i J_i(\mathbf{z}) \tag{7}$$

with $w_i \geq 0$ and $\sum_{i=1}^{m} w_i = 1$. By consistently varying the weights w_i an approximation of the Pareto set is obtained. However, despite its simplicity, the weighted sum approach has several intrinsic drawbacks [7]. A uniform distribution of the weights does not necessarily results in an even spread on the Pareto front and points in non-convex parts of the Pareto set cannot be obtained.

2.2.2 Normal Boundary Intersection (NBI)

This method has been proposed by [8] to mitigate the above mentioned drawbacks of the WS. NBI tackles the MOO problem from a geometrically intuitive viewpoint. It first builds a plane in the cost space $\mathcal{J}_f$ which contains all convex combinations of the individual minima, i.e., the *convex hull of individual minima* (CHIM), and then constructs *(quasi-)normal* lines to this plane. The rationale behind the method is that the intersection between the (quasi-)normal from any point $\mathbf{J}_p$ on the CHIM, and the boundary of the feasible cost space closest to the origin is expected to be Pareto optimal. Hereto, the MOO objective problem is reformulated as to maximise the distance λ from a point $\mathbf{J}_p$ on the CHIM along the quasi-normal through this point, without violating the original constraints. Technically, this requirement of lying on the quasi-normal introduces additional equality constraints, resulting in the following formulation:

$$\max_{\mathbf{z}\in\mathcal{S},\lambda} \lambda \tag{8}$$

$$\text{subject to : } \boldsymbol{\Phi}\mathbf{w} - \lambda\boldsymbol{\Phi}\mathbf{e} = \mathbf{J}(\mathbf{z}) - \mathbf{J}^* \tag{9}$$

where $\boldsymbol{\Phi}$ is the $m \times m$ pay-off matrix in which the i-th column is $\mathbf{J}(\mathbf{z}_i^*) - \mathbf{J}^*$, with $\mathbf{z}_i^*$ being the minimiser of the i-th objective J_i and $\mathbf{J}^*$ being the utopia point, which contains the minima of the individual objectives $J_i(\mathbf{z}_i^*)$. $\mathbf{w}$ is a vector of weights such that $\sum_{i=1}^{m} w_i = 1$ with $w_i \geq 0$, and $\mathbf{e}$ is a vector containing all ones. Now, $\boldsymbol{\Phi}\mathbf{w}$ describes a point in the CHIM and $-\boldsymbol{\Phi}\mathbf{e}$ defines the (quasi-)normal to the CHIM pointing towards the origin. When the points on the CHIM are selected with an equal spread (via a uniform distribution of $\mathbf{w}$), also an equal spread on the Pareto frontier in the objective space is obtained.

2.2.3 Normalised Normal Constraint (NNC)

NNC as introduced by [9], employs similar ideas as NBI, but combines them with features of the ε-constraint method [10]. This ε-constraint method minimises the single most important objective function J_k, while the $m-1$ other objective functions are added as inequality constraints $J_i \leq \varepsilon_i$, which are interpreted as hyperplanes reducing the feasible criterion space. After normalisation of the objectives, NNC also first constructs a plane through all individual minima (called here, the *utopia hyperplane*). Then NNC minimises a selected (normalised) objective $\overline{J}_k$, given the original constraints, and while additionally reducing the feasible space by adding $m-1$ hyperplanes through a selected point $\overline{\mathbf{J}}_p$ in the utopia plane. These hyperplanes are chosen perpendicular to each of the $m-1$ *utopia plane vectors*, which join the individual minimum $\overline{\mathbf{J}}(\mathbf{z}_k^*)$ corresponding to the selected objective $\overline{J}_k$, with all other individual minima $\overline{\mathbf{J}}(\mathbf{z}_i^*)$. Hence, this approach leads to an additional set of inequality constraints:

$$\min_{\mathbf{z}\in\mathcal{S}} \overline{J}_k \tag{10}$$

$$\text{subject to : } (\overline{\mathbf{J}}(\mathbf{z}_k^*) - \overline{\mathbf{J}}(\mathbf{z}_i^*))^{\mathrm{T}}(\overline{\mathbf{J}}(\mathbf{z}) - \overline{\mathbf{J}}_p) \leq \mathbf{0}$$

$$i = 1 \ldots m, i \neq k. \tag{11}$$

As in NBI, evenly distributed points on the utopia plane $\overline{\mathbf{J}}_p$ can be selected by a uniform variation of a vector $\mathbf{w}$, which also ensures an even spread on the Pareto set.

2.3 Predictive microbial growth model for *E. coli*

In this case study optimal experiments for the parameters of the Cardinal Temperature Model with Inflection [11] are designed. This is a secondary model to the model of Baranyi and Roberts [12]. This latter model describes the cell density as a function of time whereas the former incorporates the dependency on temperature. The model equations of the Baranyi and Roberts model are:

$$\frac{dn}{dt} = \frac{Q}{Q+1}\mu_{max}(T(t))[1-\exp(n-n_{max})] \tag{12}$$

$$\frac{dQ}{dt} = \mu_{max}(T(t))Q \tag{13}$$

in which n [ln(CFU/ml)] is the natural logarithm of the cell density, n_{max} the maximum value for n, Q [-] the physiological state of the cells. In the current optimal experiment design the state Q is omitted, because the duration of the microbial lag phase determined by the prior and actual experimental conditions, which is modelled through Q, cannot be predicted. The model equations thus reduce to a logistic growth model. The temperature dependency described by the Cardinal Temperature Model with Inflection is given by:

$$\mu_{max} = \mu_{opt}\gamma(T) \tag{14}$$

with

$$\gamma(T) = \frac{(T-T_{min})^2(T-T_{max})}{(T_{opt}-T_{min})[(T_{opt}-T_{min})(T-T_{opt})-(T_{opt}-T_{max})(T_{opt}+T_{min}-2T)]} \tag{15}$$

The values of the parameters for the model are shown in Table 1. The initial condition is $n(0) = 7$ ln(CFU/ml). The end time is fixed to 38 h. For model validity reasons the dynamic

temperature profiles are constrained to:

$$15°C \leq \quad T(t) \quad \leq 45°C \tag{16}$$

$$-5°C/h \leq \quad \Delta T(t)/\Delta t \quad \leq 5°C/h \tag{17}$$

The simplest approach is to estimate the four parameters from one single experiment. However as an alternative dividing the four in the six possible combinations is proposed in [13].

Table 1: Parameter values used for the design of the optimal experiments for predictive growth model.

T_{min}	11.33 °C	T_{opt}	40.85 °C
T_{max}	46.54 °C	μ_{opt}	2.397 1/h
n_{max}	22.55 ln(CFU/mL)	σ_n^2	3.27×10^{-2} ln(CFU/mL)^2

3 Results

3.1 Maximising information content

In order to estimate four parameters, six possible 2 by 2 combinations can be used, these six criteria are individually optimised. The six criteria as used in [13] are: $J_1 = D_{T_{max},\mu_{opt}}$, $J_2 = D_{T_{max},T_{min}}$, $J_3 = D_{T_{max},T_{opt}}$, $J_4 = D_{T_{min},\mu_{opt}}$, $J_5 = D_{T_{min},T_{opt}}$, $J_6 = D_{T_{opt},\mu_{opt}}$. The individual maxima are shown in Table 2. These values are obtained using a multiple shooting [14] approach using the ACADO-toolkit [15]. The integrator tolerance of the RungeKutta78 is set to 10^{-6}, the KKT-tolerance of the resulting SQP-problem is also to 10^{-6}. In this case, the assumption is that there measurements are performed continuously during the 38 hours of the experiment. These results are used as initialisation in the next section as they represent the upper bound of available information content. Compared with the results obtained in [13], these results are similar. Furthermore, the effect of increasing the number of discretisation intervals is clearly visible in Table 2. The third column depicts the results for a different control action every 15 minutes whereas the second column requires a control action every 2 hours. By a finer control discretisation slightly better results are obtained, however to limit computational time in the multi-objective approach, 19 control intervals will be chosen in the subsequent simulations.

Table 2: Values for the D-criterion for all the possible combinations for 19 en 152 discretisation intervals.

Combination	19 intervals	152 intervals
$T_{max} - \mu_{opt}$	8.41×10^5	8.45×10^5
$T_{max} - T_{min}$	9.38×10^4	1.12×10^5
$T_{max} - T_{opt}$	7.40×10^4	8.49×10^4
$T_{min} - \mu_{opt}$	1.80×10^6	1.81×10^6
$T_{min} - T_{opt}$	7.76×10^4	7.76×10^4
$T_{opt} - \mu_{opt}$	1.58×10^6	1.62×10^6

3.2 The trade-off between information content and measurement burden

In order to study the trade-off between information content and the measurement burden a multi-objective optimisation approach is employed, where the additional objective is the minimisation of the experimental effort. The extra decision variable introduced in this case, is $w(t)$, which indicates whether to measure or not. Each of the six described criteria are combined with the criterion which aims at minimising the measurement time. These trade-off experiments are designed using the more advanced approach, i.e., NNC. As software tool, the multi-objective extension of the ACADO-toolkit [16] is employed. For each of the combination, the number of Pareto points is set to 21. The resulting Pareto fronts for each of the criteria are displayed in Figure 1. From the Pareto fronts one can infer that this does not result in a linear or even a convex relation. Convex parts of Pareto fronts can not be found by the WS, so the advanced methods like NNC and NBI have to be used. These Pareto fronts allow the experimenter to select a specific experiment with a given amount of work. Based on the Pareto front he or she then knows how much information to expect from the experiment. The decision when to

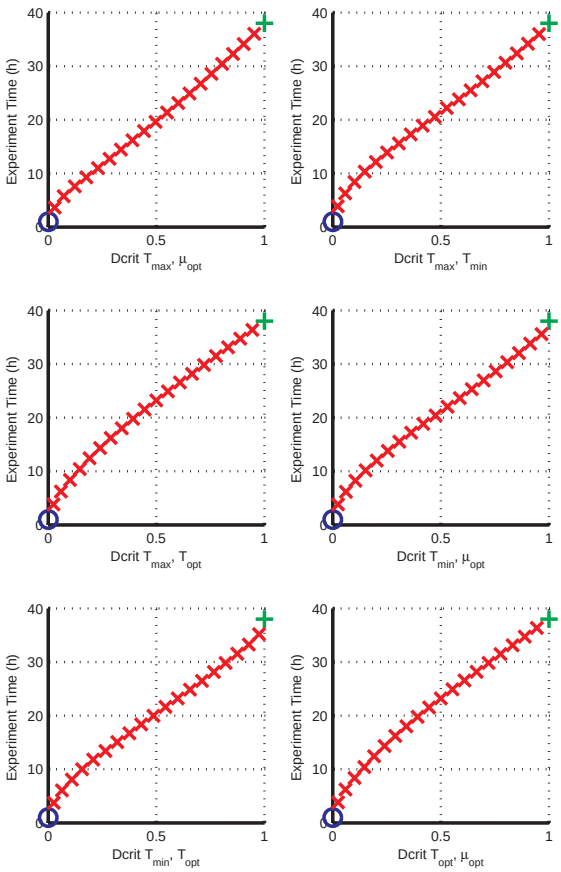


Figure 1: Pareto fronts for the six different criteria in relation to measurement time

measure, also influences the designed experiment. This is illustrated in Figure 2. Here one can see the designed experiments for $J_6 = D_{T_{opt},\mu_{opt}}$ and when one has to measure. The case with $w = [1\ 0]^T$, corresponds with measuring continuously during the experiment. Note the different weights in the figure indicate how points are selected on the CHIM or utopia plane and not the classic WS approach. From Figure 2, it is clear the designed temperature profile, can be influenced when the experimental effort is taken into consideration. In the case when there is a continuous measuring, there is a small temperature increase and subsequently decrease between 12 and 24 hours into the experiment. When the amount of measurement hours are decreased the temperature jump decreases and is spread over more hours in such a way that the increasing and decreasing parts fall in the measured time window, except for the increasing part of weight combination $w = [0.25\ 0.75]^T$.

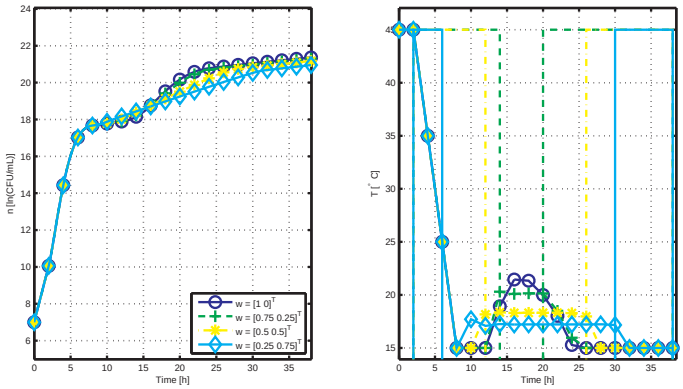


Figure 2: Evolution of cell concentration and temperature profile for maximising information content and minimising experimental effort

Conclusion

In this paper, a new approach to study the trade-off between experimental effort and information content is proposed. Both objectives are solved using advanced multi-objective optimisation methods, which allow a systematic evaluation of all different alternatives. Based on the obtained Pareto fronts, the experimenter can assess which experiment to perform based on available time and expected information content. Furthermore, it is illustrated that these experiments change when less measurement time is available. In future work, simulation studies will be performed to check if still reliable parameter estimates are obtained.

Acknowledgements

Work supported in part by Projects, OT/10/035, PFV/10/002 OPTEC (Centre-of-Excellence Optimisation in Engineering) and KP/09/005 SCORES-4CHEM knowledge platform of the Katholieke Universiteit Leuven, by the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian Federal Science Policy Office and by industry project ACCM. Dries

Telen has a Ph.D grant of the institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Jan Van Impe holds the chair Safety Engineering sponsored by the Belgian chemistry and life sciences federation essenscia.

References

[1] F. Pukelsheim. *Optimal design of Experiments*. John Wiley & Sons, Inc., New York, 1993.

[2] E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Springer, Paris, 1997.

[3] S. Sager. Sampling decisions in optimum experimental design in the light of Pontryagin's maximum principle. *SIAM Journal on Control and Optimization*, 2012. (submitted).

[4] G. Franceschini and S. Macchietto. Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, 63:4846–4872, 2008.

[5] K. Miettinen. *Nonlinear multiobjective optimization*. Kluwer Academic Publishers, Boston, 1999.

[6] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley, Chichester, London, UK, 2001.

[7] I. Das and J.E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14:63–69, 1997.

[8] I. Das and J.E. Dennis. Normal-Boundary Intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, 8:631–657, 1998.

[9] A. Messac, A. Ismail-Yahaya, and C.A. Mattson. The normalized normal constraint method for generating the Pareto frontier. *Structural & Multidisciplinary Optimization*, 25:86–98, 2003.

[10] Y.Y. Haimes, L.S. Lasdon, and D.A. Wismer. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1:296–297, 1971.

[11] L. Rosso, J.R. Lobry, and J.P. Flandrois. An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model. *Journal of Theoretical Biology*, 162:447–463, 1993.

[12] J. Baranyi and T.A. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 23:277–294, 1994.

[13] E. Van Derlinden, K. Bernaerts, and J. Van Impe. Simultaneous versus sequential optimal experiment design for the identification of multi-parameter microbial growth kinetics as a function of temperature. *Journal of Theoretical Biology*, 264:347–355, 2010.

[14] H.G. Bock and K.J. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. In *Proceedings of the 9th IFAC world congress, Budapest*. Pergamon Press, 1984.

[15] B. Houska, H.J. Ferreau, and M. Diehl. ACADO Toolkit - an open-source framework for automatic control and dynamic optimization. *Optimal Control Applications and Methods*, 32:298–312, 2011.

[16] F. Logist, B. Houska, M. Diehl, and J.F. Van Impe. Fast pareto set generation for nonlinear optimal control problems with multiple objectives. *Structural and Multidisciplinary Optimization*, 42:591–603, 2010.

Planification d'expériences optimales pour la modélisation des procédés alimentaires

Design of optimal experiments to model food processes

Daniel Goujot[1,2], Xuân Meyer[3] & Francis Courtois[2,1]

[1] *INRA, UMR1145 Ingénierie Procédés Aliments, 1 avenue des Olympiades, F-91300 Massy, France*
E-mail : *daniel.goujot@agroparistech.fr*

[2] *AgroParisTech, UMR1145 Ingénierie Procédés Aliments, 1 avenue des Olympiades, F-91300 Massy, France*
E-mail : *francis.courtois@agroparistech.fr*

[3] *Université de Toulouse, Laboratoire de Génie Chimique, UMR CNRS/INPT/UPS 5503 BP 34038, 4 allée Emile Monso, 31030 Toulouse cedex 4, France*
E-mail : *xuan.meyer@ensiacet.fr*

Résumé
Il existe des verrous expérimentaux et calculatoires dans l'identification de *paramètres* inconnus de modèles de procédés alimentaires, particulièrement lorsque les modèles sont basés sur des systèmes dynamiques non-linéaires en les paramètres. Le but de la planification expérimentale est d'optimiser les expériences de façon à faciliter cette identification.
L'approche originale de la planification séquentielle utilisée dans Goujot *et al.* (2012) est ici détaillée. Elle combine des algorithmes d'optimisation testant plusieurs initialisations pour les inconnues, un reconditionnement systématique de toutes les grandeurs optimisées. La dérivation en les paramètres des équations différentielles du modèle permet de calculer les sensibilités de manière semi-analytique.
Cette approche est validée dans quatre contextes dont:
(i) retrouver des plans optimaux pharmacocinétiques tels que publiés par Pronzato (2008).
(ii) retrouver des barèmes optimaux de fermentation tels que publiés par van Derlinden *et al.* (2008).
(iii) valider expérimentalement une démarche séquentielle visant à identifier un modèle de séchage sur un pilote de laboratoire à Massy Goujot *et al.* (2012).
Mots-clés : expériences optimales, planificiation d'expériences, reconditionnement, reparamétrisation, dérivation de systèmes dynamiques, critère A-optimal, critère D-optimal, critère E-optimal

Abstract
The identification of model parameters in food processes is both an experimental and a numerical challenge. The aim of the design of experiments is to optimize (tune) the experiments at the lowest experimental cost in order to get best identification results.
In this work, an original approach and implementation of the sequential experiment design is presented. It combines multistart optimization, systematic reconditioning by hand of all unknowns being optimized or integrated, direct-differentiation method to compute sensitivities.
This approach is validated over four applications, and was able to:
(i) find again the optimal pharmacokinetic experiment from Pronzato (2008).
(ii) find again the optimal fermentation experiments from van Derlinden *et al.* (2008).
(iii) compute three optimal experiments drying rice with varying temperature Goujot *et al.* (2012).
Keywords : optimal experiments, reconditioning, reparameterization, direct-differentiation method, A-optimal criterion, D-optimal criterion, E-optimal criterion

1 Introduction

The many improvements of the implementation of the sequential experiment design for model based on differential equations used in Goujot *et al.* (2012) are detailed, justified and illustrated. It combines multistart optimization, systematic reconditioning by hand of all unknowns being

optimized or integrated, direct-differentiation method to compute sensitivities. This approach has been recently implemented in a Matlab Toolbox published by INRA Bertrand & Cordella (2011).

1.1 Principle of experiment design (DoE)

Given a number of available experimental data, and a computer *model* which reflects perfectly the physics of the experiments, the identification consists in estimation of values and confidence intervals for the unknown *parameters* of the model (*e.g.* heat & mass transfer coefficients). A classical identification setting minimizes the norm of the *residual* vector, which is the difference between data and its value predicted by the model, divided by the standard error of the measurement. At the end of the computations, the parameter estimates are issued with their relative uncertainties.

As a general assumption, the residual is a projection of a vector of independent normalized Gaussian random variables. The projection is orthogonal to the linear subspace tangent to the *manifold* of values generated by the model for all possible parameter vectors. The parameter *confidence region* is the parameter set whose corresponding part of the manifold lies in a distance of at most $\chi_2(N)$ from its identified value, where N is the number of parameters.

The principle of **sequential design of optimal experiments for models nonlinear in parameters** is to optimize a subsequent experiment.The free parameters (*protocol*) of this experiment are optimized in order to diminish some *norm* on the confidence region over the estimated (model related) parameters. Some examples of norms have been published: Vila & Gauchi (2007) used the region volume, and provides a numerical implementation of so-called X-optimality; Dette *et al.* (2003) prefered the region diameter.

Gauchi & Pázman (2006) maximized the size of a "norm" of the covariance of the predictions generated when the parameter vectors follows a normal Gaussian uncentered law. In these cases, the generalized A- and D-optimality criteria correspond to the trace and determinant of this covariance matrix. Pázman & Pronzato (1992) also defined the E-optimality in this context.

In this work, above cited designs were not used, for two reasons:

- According to Table 1, the optimization of numerical approximation of these criteria accumulate billions of model calls. It is very expensive, in our context, since there is no way to avoid the numerical resolution of ODEs at each model call.
- The confidence regions estimated are not close to the parameter boundaries. In this context the loss of information due to use of linear approximation detailed in part 1.2, when compared to generalized methods of design of experiments (Gauchi & Pázman, 2006, p1144) seen on last line of Table 1, are close to 14% which is negligible.

1.2 Linearization of experiment design

Most DoE applications assume that the parameter vector is a gaussian vector centered on its true value, whose covariance matrix is given by *Fisher*'s *information matrix*. This simplification was not used in works cited in part 1.1. *Fisher*'s *information matrix* is the inverse of the *Gramm* matrix of the set of derivatives of the model predictions with respect to each model parameter. For instance, the first coefficient of this *Gramm* matrix is the sum of squares of model prediction derivatives with respect to the first model parameter. It is not directly related to the variance coefficient of the first parameter, although *Schur* reduction proves that their

Table 1: For each class of models: Comparative efficiencies in term of typical time duration of designed experiments (***: shortest (best); **: 10-20% more; *: at least 200% more), frequency in DoE applications found in publications (+++: majority; ++: uncommon; +: barely).

Class of model			solved by numerical ODE	others, non-linear in parameter	others, non-polynomial in protocol	other models (simplest)
Typical execution time of compiled model			10ms	10μs	1μs	10ns
Type of design of experiment (DoE)	Type of optimizer	Typical model calls				
N-factor M-level	Table lookups	0	+++/*	+++/*	+++/*	+++/***
A-, D-, E-optimality	Local	10^4	++[1]/**	++/**	++/***	***[2]
	Global[3]	10^6	+/**	+/**	***[2]	***[2]
Generalized, X-...	Custom[4]	10^9	***[5]	+[6]/***	***[2]	***[2]

[1]Goujot *et al.* (2012); Shin *et al.* (2007); Bäumler *et al.* (2007) [2]no interest because above DoE does the job [3]Moles *et al.* (2003) [4]Gauchi & Pázman (2006) [5]no known publication [6]Pázman & Pronzato (1992); Vila & Gauchi (2007)

product is greater than 1. The A-, D- and E-optimalities consist in minimizing respectively the trace, the determinant and the spectral radius of this information matrix. When the model is linear in the parameters, these criteria correspond to their generalized counterparts presented in section 1.1. Hence the identified parameter are non-biased only when models are linear in their parameters.

It is generally admitted that A-optimal criterion is smoother than the E-optimal criterion (Uciński, 2005, p. 293), hence it was preferred in this work.

2 Setup

Some papers are dealing with very specific DoE exchange-type algorithm known as Fedorov, like trying to design sampling locations Uciński (2005) or to identify one single concentration Casey *et al.* (2007) with a dedicated methodology. In the general case as in Goujot *et al.* (2012), this class of algorithms cannot be used: the protocol vector cannot be split in two parts to avoid correlated effects on the predictions.

In this work, a fair amount of time was dedicated to design the most generic, yet fast, minimization algorithms. As a matter of fact, solution came from some -negligible- linear approximation combined with smart parameterization and multistart search. For instance, 20 different initial guesses are used for the protocol vector hence minimizing the risk for local minimum lockout.

2.1 Smart parameterization

The toolbox optimizes the design of experiments under the condition that the parameterization of experiments are systematically well conditioned, which means that:

1. **Little correlation.** Protocol elements should be chosen in order to minimize coupled effects on model predictions. The search for optimization directions is way more efficient.

For instance, a reaction rate depending on temperature T is parameterized by its values at extremal conditions $(k_{T\,\min}, k_{T\,\max})$ which limits magnitude differences and disturbing cross-couplings found with classical Arrhenius parameterization (k_0, E_a).

2. **Simple boundaries.** The total protocol region should be parallelipipedic (simple boundaries on each protocol number) and each dimension should have a domain of variation whose order of magnitude is close to one. This allows algorithms to converge faster since they naturally stick in the autorized domain and do not waste time to deal with domain constraints.

3. **Smoother search.** The set of couples of protocol vectors that give the same predictions should have the smallest dimension and norm as possible. For example, if temperature profile T depending on time t is to be linearly interpolated between four control values $(0, T_0), (t_1, T_1), (t_2, T_2), (t_3, T_3)$, it should be parameterized by t_1/t_3, $(t_2 - t_1)/(t_3 - t_1)$, t_3, T_0, T_1, T_2, and T_3, with 0 and 1 being the minimum and maximum of first two parameters. The parameterization by the t_i and T_i with easy boundaries would for example have the couple $([(0, T_0), (t_1, T_1), (t_2, T_2), (t_3, T_3)], [(0, T_0), (t_2, T_2), (t_1, T_1), (t_3, T_3)])$ which contains two protocol vectors giving the same predictions. As any permutation would apply, the number of local minima in the search space would be six time larger.

4. **Local identifiability.** The parameterization of protocol should be chosen to ensure significant influence on model predictions, on the whole domain of variation. Otherwise, the optimizer may get locally stuck on a flat hyper surface. As a corollary remark, it is better to make experiment design for model discrimination between several simple reactions pathways, than to make experiment design for a single model with complex reaction pathways containing all possible reactions.

5. **Consistent parameterization** Two protocol vectors which are close (difference with order of magnitude smaller than one) should give model predictions which are qualitatively similar. This is because the algorithm is making even efforts on the whole search domain, and may miss something depending on additional efforts in a small subregion of the search domain. For example, if a limitant diffusivity coefficient may have small values, it should be replaced by its logarithm; if this diffusion coefficient is not so limitant below a certain small threshold, the hyperbolic sinus of the ratio between the diffusion coefficient and this threshold might give interesting results.

6. **Readability.** As a matter of fact, the above described model parameterization should be done with publishing in mind. Hence, each protocol element should be compatible with a clear, easy to understand, English expression. This way, there would be no need to transform the set of parameter confidence intervals for communication purposes. For instance, a diffusion coefficient is better than a diffusivity parameter, even in a compartment model.

The parameter vector should be reconditioned for similar reasons; it makes the confidence regions more natural, and the identification optimization more efficient and faster.

2.2 Computer implementation

Implemented as a MATLAB Toolbox, the overall calculation process typically requires the computer resources to run about 10^4 simulations.

In the toolbox, the direct-differentiation method Bock (1981); Leis & Kramer (1988)(Uciński, 2005, p. 29) provides a more stable *Fisher*'s information matrix than the finite difference differentation. This eliminates a lot of local minima for the optimizer. The sensitivities are hence computed by numerical integration of exact differential equations. One may consider it as a semi-analytical calculation of sensitivies.

Additionnaly, since the state variables of the (model) differential system and their derivatives in the parameters have been respectively divided by their maxima and passed through an hyperbolic sinus, the time integration of the above exact differential equations is the arc hyperbolic sinus of the sensitivities.

In the computation of *Fisher*'s information matrix, the inversion was also tuned to avoid good D-optimal scores for near-singular information matrix. The technique takes into account the relative precision on the *Gramm* matrix. Without this technique, if an information matrix is close to singular, then the computation of its determinant would be very sensitive to computation roundings used to compute this information matrix. The technique is detailed in Goujot *et al.* (2012).

3 Application cases

In all cases, a valid simulation model, with plausible initial guess for parameters, is required. In addition, the experiment design is efficient when the pilot plant has quick transient regimes (limited inertia) and/or when there is a plausible model for the inertia. Ideally, the online instrumentation should be designed to minimize uncertainties on measurements, ensuring structural identifiability of the model.

This Matlab toolbox has been validated over 3 reactors, coming from literature. Last application was effectively tested in a real experimental work.

3.1 Pharmacokinetic application

The toolbox was able to find the optimal experiment protocol for a pharmacokinetic problem described and solved by van Derlinden *et al.* (2008).

Diffusion takes place in the circular system of an animal. The accummulated concentration of an injected drug is measured eight times in the peripherical part. The toolbox found that optimal sampling times were [1.0003, 1.0004, 10.4131, 10.4273, 67.3876, 70.8813, 718.9642, 719.7539] (values in minutes) which is very close to the D-optimal protocol [1, 1, 10, 10, 74, 74, 720, 720] proposed by Pronzato (2008), as a validation case for constant conditions and varying sampling times. Even better, it has a D-optimal criterion which is 0.31% smaller.

3.2 Fermentation application

The toolbox was able to find the optimal temperature profile of a biological reactor of a fermentation problem described and solved by van Derlinden *et al.* (2008). The concentration of Eschericha Coli is known at every moment (continuous observation), and its activation is not measured directly.

Table 2 shows that the presented methodology is able to compute independently the six D-optimal protocols proposed by van Derlinden *et al.* (2008), as a validation case for experiments with non-constant temperature profiles.

Table 2: Results of proposed methodology in fermentation case, compared to (van Derlinden *et al.*, 2008, Table 2).

D-optimality criterion		initial temperature (°C)		final temperature (°C)		duration of initial temperature (h)		temperature decrease (°C/h)	
published	this work	published	this work	published	this work	published	this work	published	this work
$6.815{\cdot}10^5$	$8.48{\cdot}10^5$	45	44.983	15	15.042	3.835	3.0785	5	4.6126
96949	$1.12{\cdot}10^5$	45	44.999	15.94	15.2	2.854	3.5405	5	4.9986
66129	81980	45	44.935	16.62	15.833	2.938	3.2592	5	4.8999
$1.798{\cdot}10^6$	$1.8{\cdot}10^6$	40.85	40.987	16.91	16.952	1.78	1.7432	5	4.9981
76796	76740	31.15	31.174	16.87	16.871	4.382	4.2245	2.388	2.2608
$1.528{\cdot}10^6$	$1.55{\cdot}10^6$	45	44.979	16.67	16.751	2.801	2.6751	5	4.9956

3.3 Experimental dryer

As published in Goujot *et al.* (2012), a validated rice drying model from Courtois *et al.* (2001) is considered.

The inlet air temperature, relative humidity and velocity are controlled by three independent PID controllers, under the supervision of a computer. Product (rice here) is weighted every minute. The current mean grain moisture content can be computed from this weight; it is the only product related variable that is measured online.

The coefficients of the protocol vector are the following:

- the air temperature set points (divided by $T_{\max}$ for normalization) of each of the five time segments.
- the set points of water addition related to their maximum available value (saturation, or maximum capacity reached) for each of the five time segments.
- the duration of each first four time segments divided by the remaining time of current experiment.

Minimal value is $T_{\min}/T_{\max}$ for first 5 protocol coefficients, and 0 for last 9 protocol coefficients. Maximal value for all protocol coefficients is 1. The control of air drying conditions is done by PID control set points, according to the vector protocol. Hence, it is necessary to inject, in the model, the pilot inertias in response of set point changes to mimic correctly in the simulation what will happen in real.

The computation time of an A-optimal protocol is three hours on a cluster of 12 processors (about a day on one processor).

The proposed methodology allows the replacement of nine drying experiments with static conditions by three A-optimally designed drying experiments with dynamic conditions. The experimentation time and the amount of used product are also divided by three while the confidence intervals on the identified parameters were quite as effective.

4 Conclusion

An original approach, and its implementation, for the optimal sequential design of experiments has been described. It is available for the experiment design of any model relying on the

numerical time integration of ODEs.

Fine parameterization of the parameter and protocol vectors allows for the use of standard local optimization algorithms while working models with no analytical solution. Furthermore, global or proprietary optimizers were avoided (as opposed to some other published works Rodriguez-Fernandez *et al.* (2006); Balsa-Canto *et al.* (2007); Shin *et al.* (2007); Bäumler *et al.* (2007). In addition, standard local optimizers were proved sufficient for the DoE of a nonlinear model and confronted both to published cases and to real experiments.

This DoE implementation or dynamic systems give results comparable to optimal designs already published in litterature, in the context of fixed profile by Pronzato Pronzato (2008), and in the context of variable non-constant temperature profile by vanDerlinden et al. van Derlinden *et al.* (2008).

Based on a benchmarking on the drying case Goujot *et al.* (2012), A-optimality was found to be somehow better than E-optimality and far better than the D-optimality. Suprisingly, it was observed, in the same context, that first experiment reaches directly the right parameter neighborhood (same order of magnitude).

An ongoing work in collaboration with *Laboratoire de Génie Chimique de Toulouse* (ENSIACET) is in progress to apply this methodology to a tubular reactor.

This toolbox is currently implemented in SAISIR Bertrand & Cordella (2011) toolbox (in upcoming release) which is published by INRA.

References

Balsa-Canto, E., Rodriguez-Fernandez, M., & Banga, J. R. (2007). Optimal design of dynamic experiments for improved estimation of kinetic parameters of thermal degradation. *Journal of Food Engineering*, 82(2), 178–188.

Bertrand, D. & Cordella, C. B. Y. (2011). SAISIR Package®. Free toolbox for chemometrics in the MATLAB ®, Octave or Scilab environments.

Bock, H. G. (1981). Numerical Treatment of Inverse Problems in Chemical Reaction Kinetics. In K. H. Ebert, P. Deuffhard, & W. Jager, editors, *Modelling of chemical reaction systems: proceedings of an international workshop*, volume 18 of *Springer series in Chemical Physics*, (pp. 102–125), Heidelberg, September 1-5, 1980: Springer-Verlag.

Bäumler, C., Matzopoulos, M., & Urban, Z. G. E. (2007). Enhanced methods optimize ownership costs for catalysts. *Hydrocarbon Processing*, 86(6), 71–78.

Casey, F. P., Baird, D., Feng, Q., Gutenkunst, R. N., Waterfall, J. J., Myers, C. R., Brown, K. S., Cerione, R. A., & Sethna, J. P. (2007). Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Systems Biology*, 1(3), 190–202.

Courtois, F., Abud Archila, M., Bonazzi, C., Meot, J. M., & Trystram, G. (2001). Modeling and control of a mixed-flow rice dryer with emphasis on breakage quality. *Journal of Food Engineering*, 49(4), 303–309.

Dette, H., Melas, V. B., Pepelyshev, A., & Strigul, N. (2003). Efficient design of experiments in the Monod model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3), 725–742.

Gauchi, J. P. & Pázman, A. (2006). Designs in nonlinear regression by stochastic minimization of functionals of the mean square error matrix. *Journal of Statistical Planning and Inference*, 136(3), 1135–1152.

Goujot, D., Meyer, X. M., & Courtois, F. (2012). Identification of a rice drying model with an improved sequential optimal design of experiments. *Journal of Process Control*, 22(1), 95–107.

Leis, J. R. & Kramer, M. A. (1988). Algorithm 658: ODESSA - an ordinary differential equation solver with explicit simultaneous sensitivity analysis. *ACM Transactions on Mathematical Software (TOMS)*, 14, 61–67.

Moles, C. G., Mendes, P., & Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13(11), 2467–2474.

Pronzato, L. (2008). Optimal experimental design and some related control problems. *Automatica*, 44(2), 303–325.

Pázman, A. & Pronzato, L. (1992). Nonlinear experimental design based on the distribution of estimators. *Journal of Statistical Planning and Inference*, 33(3), 385–402.

Rodriguez-Fernandez, M., Mendes, P., & Banga, J. R. (2006). A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Bio Systems*, 83(2-3), 248–265.

Shin, S. B., Han, S. P., Lee, W. J., Im, Y. H., Chae, J. H., Lee, D.-i., Lee, W. H., & Urban, Z. G. E. (2007). Optimize terephthaldehyde reactor operations. *Hydrocarbon Processing*, 86(4), 83–90.

Uciński, D. (2005). *Optimal measurement methods for distributed parameter system identification*. CRC Press.

van Derlinden, E., Venken, L., Bernaerts, K., & van Impe, J. F. (2008). Optimal dynamic experiment design as a tool for accurate estimation of microbial growth cardinal temperatures. In *FOODSIM 2008, June 26-28*, (pp. 102–109), University College Dublin, Ireland: EUROSIS-ETI Bvba.

Vila, J.-P. & Gauchi, J.-P. (2007). Optimal designs based on exact confidence regions for parameter estimation of a nonlinear regression model. *Journal of Statistical Planning and Inference*, 137(9), 2935–2953.

Génération automatique de plans factoriels réguliers : la librairie R PLANOR

Automatic generation of regular factorial designs : the PLANOR R library

Hervé Monod[1,2] & André Kobilinsky[1] & Annie Bouvier[1]

[1] *INRA, UR341, Unité MIA-J, Jouy-en-Josas*
E-mail : *herve.monod@jouy.inra.fr*

[2] *Isaac Newton Institute, Cambridge*

Résumé
Les plans d'expériences sont d'une utilité reconnue pour l'expérimentation en agro-alimentaire. Dans cet article, nous nous intéressons aux plans factoriels dits réguliers, construits par des relations de définition entre facteurs codées sous forme d'une *matrice clé*. Après avoir illustré cette classe de plans à travers un plan fractionnaire et un plan en carré latin, nous présentons la librairie R PLANOR, qui permet de les générer par un algorithme backtrack. Nous décrivons ensuite l'utilisation de la librairie sur quelques exemples d'application.
Mots-clés : plan d'expériences factoriel, plan factoriel régulier, algorithme backtrack

Abstract
Factorial designs are a most useful tool for experimentation in agro-food research and development. In this paper, we are interested in regular factorial designs, constructed by defining relationships between factors that are coded in a *key matrix*. After illustrating this class of designs through a fractional factorial design and a latin square, we present the PLANOR R library, which allows to generate them by a backtrack algorithm. We then show how the library can be used on a few examples of application.
Keywords : factorial experiment, regular factorial design, backtrack algorithm

1 Introduction

Initiés par Fisher et Yates dans les années 1930 (Yates, 1933, 1937 ; Fisher, 1942), les plans d'expériences factoriels se sont développés dans tous les domaines d'application, en particulier en agro-alimentaire. Nous nous intéressons ici aux plans factoriels réguliers, dont la construction algébrique est directement héritée des travaux de Fisher et Yates. Cette classe de plans recouvre un grand nombre de plans d'expériences utilisés en pratique pour étudier plusieurs facteurs simultanément et pour contrôler des effets blocs (Kobilinsky, 1997). Ils vérifient des propriétés d'orthogonalité qui les rendent optimaux dans un sens très large en terme de précision d'estimation des paramètres.

Dans les années 90, une méthode algorithmique a été développée par André Kobilinsky pour générer automatiquement des plans factoriels réguliers sous des conditions très générales,

en prenant compte de diverses contraintes fréquemment rencontrées en pratique (Kobilinsky, 2005). Cette méthode est issue du principe de la matrice clé proposé par Franklin et Bailey (1977) et Patterson et Bailey (1978). Elle produit des plans dans lesquels les effets factoriels issus d'une décomposition canonique précise sont soit mutuellement orthogonaux, soit complètement confondus.

Initialement implémentée sous la forme du logiciel PLANOR programmé en APL et utilisable sous Windows (Kobilinsky, 2005), la méthode a maintenant été implémentée sous la forme d'une librairie R. Nous en détaillons les principes et présentons quelques exemples d'application.

2 Plans factoriels réguliers

2.1 Un exemple de fraction régulière

Considérons le cas d'un plan pour quatre facteurs A, B, C, D tous à deux modalités notées 0 et 1. Supposons qu'un plan de taille 8 seulement soit envisageable, Il faut donc sélectionner $N = 8 = 2^3$ traitements (a, b, c, d) parmi les 16 possibles. La solution classique est de faire un plan complet sur A, B, C et de déterminer les modalités de D par l'équation $d = a + b + c \pmod 2$, appelée *relation de définition*. On obtient le plan du Tableau 1.

A	B	C	D
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

TABLE 1 – Plan fractionnaire et sa matrice X pour 4 facteurs en 2^3 unités expérimentales.

Si on analyse ce plan avec un modèle $\mathcal{M}$ d'analyse de la variance comprenant les effets principaux et les interactions entre deux facteurs, on obtient la matrice X du modèle linéaire donnée dans le Tableau 2. Les colonnes de X sont associées aux effets factoriels notés 1 (moyenne générale), A, ..., D (effets principaux), $A.B$, ..., $C.D$ (interactions entre deux facteurs). Elles sont soit mutuellement orthogonales, soit confondues (par exemple, les colonnes AB et CD sont égales). Du point de vue statistique, les effets principaux sont estimables et on peut démontrer que le plan est optimal pour la précision de leurs estimateurs. Les interactions entre deux facteurs sont confondues deux à deux et donc non estimables individuellement, mais le plan garantit qu'elles n'ont aucun risque de biaiser l'estimation des effets principaux.

Pour construire ce plan, on a associé les traitements aux unités par l'équation $t = \Phi u \pmod 2$, où $u = (u, v, w)^T$ représente une unité, $t = (a, b, c, d)^T$ représente le traitement appliqué à u, et $\Phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. On obtient $\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \Phi \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} u \\ v \\ w \\ u+v+w \end{pmatrix}$, et l'on retrouve

$$X = \begin{pmatrix} 1 & A & B & C & D & AB & AC & AD & BC & BD & CD \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

TABLE 2 – Fractional design and its X matrix for 4 factors in 2^3 experimental units.

ainsi la relation de définition $d = a + b + c$.

La transposée de Φ, notée ici Φ^*, est appelée *la matrice clé*. Elle contient d'une part l'information pour construire le plan (les relations de définition), et d'autre part l'information sur les confusions d'effets. Notons, comme il est d'usage, les effets factoriels sous la forme $A^a B^b C^c D^d$ (par exemple, $A^0 B^0 C^0 D^0 = 1$ dénote la moyenne, $A^1 B^0 C^1 D^0 = AC$ dénote l'interaction entre A et C, *etc.*). On montre que deux effets factoriels $A^a B^b C^c D^d$ et $A^{a'} B^{b'} C^{c'} D^{d'}$ sont confondus si et seulement si le vecteur $(a - a', b - b', c - c', d - d')^T$ appartient au noyau de la matrice clé Φ^*. Par exemple, on retrouve la confusion entre les interactions AB et CD par le fait que $\Phi^*((1, 1, 0, 0) - (0, 0, 1, 1))^T$ est égal au vecteur nul dans $(\mathbb{Z}_2)^3$.

2.2 Un exemple de plan en blocs

Le plan en carré latin est parfois utilisé en agro-alimentaire, par exemple pour des expériences de dégustation. Il ne nécessite pas de logiciel sophistiqué pour sa construction, mais il offre un exemple simple et complémentaire du précédent d'un plan factoriel régulier.

Supposons que l'on veut étudier un facteur A à 5 modalités, par exemple des produits à comparer. Supposons que l'on dispose pour conduire les observations, de 25 unités expérimentales réparties en 5 lignes et 5 colonnes, où lignes et colonnes représentent, par exemple, les dégustateurs et les ordres de présentation. On note L et C les facteurs lignes et colonnes, à 5 modalités. Le plan en carré latin du Tableau 3 a été construit en identifiant les modalités de L, C, A aux éléments (l, c, a) de $(\mathbb{Z}_5)^3$ et en prenant comme matrice clé $\Phi^* = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 3 \end{pmatrix}$, c'est-à-dire en utilisant la relation de définition $a = l + 3c \pmod 5$.

Dans ce plan, les trois facteurs sont mutuellement orthogonaux, au sens que tous les couples de modalités (l, c), (l, a), (c, a) sont présentes un même nombre de fois. Cependant, le plan est incomplet (il nécessiterait 125 unités expérimentales) et les effets factoriels sont confondus par groupes de 5 contrastes mutuellement orthogonaux. Un de ces groupes est par exemple constitué des contrastes notés L^2, $L^3 C^3 A^4$, $L^4 C^1 A^3$, $L^5 C^4 A^2$, $C^2 A^1$. Il comprend un degré de liberté de l'effet principal du facteur L, trois degrés de liberté de l'interaction $L.C.A$, et un degré de liberté de l'interaction $C.A$.

Du point de vue statistique, les effets principaux des trois facteurs sont estimables dans le modèle additif supposant les interactions négligeables. Concrètement, le plan permet de comparer les 5 produits tout en contrôlant les effets des dégustateurs et des ordres de présentation.

Produit A		Colonne				
		$C0$	$C1$	$C2$	$C3$	$C4$
	$L0$	0	3	1	4	2
	$L1$	1	4	2	0	3
Ligne	$L2$	2	0	3	1	4
	$L3$	3	1	4	2	0
	$L4$	4	2	0	3	1

TABLE 3 – Plan en carré latin construit par la relation de définition $a = l + c \pmod 5$.

2.3 Généralisation

Les deux exemples précédents font partie des plans d'expériences dits orthogonaux, avec confusions d'effets soit totales soit nulles. Ils s'analysent tous deux avec des méthodes basées sur le modèle linéaire et l'analyse de variance. Du point de vue mathématique, ils sont encore plus proches, puisqu'on peut décrire leur construction avec exactement le même formalisme.

De façon plus générale, un plan factoriel régulier possède les caractéristiques suivantes (Pistone et Rogantin, 2008 ; Kobilinsky *et al.*, 2011) :

- les modalités des facteurs sont identifiées aux éléments d'un groupe T produit de groupes cycliques (tels que $\mathbb{Z}_2$, $\mathbb{Z}_5$, etc.) ;
- il y a N unités expérimentales et ces unités sont elles aussi identifiées aux éléments d'un groupe U produit de groupes cycliques ;
- le traitement t affecté à l'unité u est défini par la relation $t = \Phi u + t_0$, où Φ est un morphisme de groupe, c'est-à-dire vérifie $\Phi(u + u') = \Phi(u) + \Phi(u')$ pour toute paire d'unités u et u' appartenant à U, et t_0 est un élement de T (sauf mention du contraire, on utilise dans cet article le vecteur nul pour t_0).

La définition générale des plans factoriels réguliers inclut les cas où les nombres de niveaux des facteurs sont quelconques et diffèrent d'un facteur à l'autre. Il n'est pas possible, dans ce court article, de rentrer dans les détails techniques rencontrés dans le cas général. Mais il faut retenir que comme dans les exemples, les matrices Φ et Φ^* contiennent l'information essentielle, d'une part pour construire un plan factoriel régulier, d'autre part pour en déterminer les effets mutuellement confondus et en déduire quels termes factoriels sont estimables pour un modèle donné. La recherche d'une matrice clé Φ^* adaptée aux objectifs du plan d'expériences est donc au cœur des calculs effectués par PLANOR.

3 Présentation de PLANOR

3.1 Principes d'utilisation

Dans les deux exemples ci-dessus, nous sommes partis d'un plan d'expériences construit avec une matrice clé, puis nous avons déduit de cette matrice clé les termes estimables pour un modèle donné. En pratique, le choix du plan d'expériences se pose généralement de façon inverse : en fonction des facteurs que l'on veut étudier et du modèle que l'on prévoit d'appliquer, on cherche un plan d'expériences permettant d'estimer tous les termes factoriels qui nous intéressent.

PLANOR a été conçu pour répondre à ce besoin. Ses entrées sont :

- la liste des facteurs et de leurs modalités (facteurs à étudier et facteurs blocs) ;
- la taille du plan d'expériences recherché ;
- le modèle $\mathcal{M}$ prévu pour l'analyse et les termes factoriels $\mathcal{E}$ que l'on veut estimer dans ce modèle.

À partir de ces spécifications, PLANOR recherche une ou plusieurs matrices clés solutions, c'est-à-dire permettant de construire un plan factoriel régulier de la taille souhaitée et tel que les termes factoriels $\mathcal{E}$ soient estimables si on applique le modèle $\mathcal{M}$. Des fonctions permettent ensuite de construire le plan d'expériences recherché.

Par ailleurs, PLANOR dispose d'atouts supplémentaires, qui seront illustrés dans la Section 4. En particulier il est possible

- de spécifier des contraintes entre facteurs, appelées contraintes de hiérarchie ;
- de spécifier plusieurs couples (modèle, effets à estimer) ;
- de randomiser le plan d'expériences obtenu pour des structures en blocs très générales.

3.2 L'algorithme principal

L'algorithme implémenté dans PLANOR est complexe et sa description détaillée fait l'objet d'un article en cours de préparation (Kobilinsky *et al.*, 2011). Son principe est de déterminer, en fonction du modèle et des termes à estimer, les effets factoriels qui ne doivent pas appartenir au noyau de Φ^* (c'est-à-dire être confondus avec la moyenne générale).

Cette première étape conduit à une décomposition du problème de recherche selon les différents nombres premiers impliqués dans les nombres de modalités des facteurs. Puis la recherche d'une matrice clé vérifiant ces contraintes est entreprise pour chaque nombre premier, par un algorithme backtrack. Cet algorithme consiste à construire Φ^* colonne par colonne, en explorant pour chaque colonne toutes les possibilités jusqu'à ce qu'une solution soit trouvée. Si aucune solution n'est possible pour la colonne j, l'algorithme revient en arrière pour tester la solution suivante pour la colonne $j-1$. L'algorithme s'arrête s'il a trouvé le nombre de solutions demandé par l'utilisateur ou s'il a exploré sans succès toutes les possibilités. Bien entendu, pour certains problèmes à grand nombre de facteurs, une exploration complète n'est pas envisageable. Actuellement, la recherche doit alors être interrompue manuellement par l'utilisateur.

3.3 Outils de post-traitement

En pratique, la recherche du plan d'expériences le mieux adapté à une situation concrète est généralement l'objet d'une étude approfondie des contraintes et des différentes possibilités : des compromis doivent être trouvés entre la taille du plan et le nombre de facteurs ou de modalités de ces facteurs ; plusieurs modèles sont envisagés ; etc. Il faut alors tester et comparer plusieurs jeux d'entrées et plusieurs solutions.

Pour accompagner cette démarche pragmatique, plusieurs classes d'objets R sont définies par la librairie PLANOR pour décrire les objets générés lors de la recherche d'un plan. En particulier, une telle classe existe pour les matrices clés et PLANOR dispose de fonctions permettant d'étudier plus finement les confusions d'effets associées à un objet de cette classe (fonctions `print`, `summary`, `alias`). D'autres fonctions permettent de construire et randomiser le plan d'expériences à mettre en place à partir de la matrice clé. Par ailleurs l'environnement R permet facilement à l'utilisateur averti d'effectuer ses propres développements.

4 Applications

Nous illustrons l'utilisation de PLANOR en reprenant les deux exemples de la Section 2. Par souci de concision, certains arguments ou options ne sont pas détaillés et certaines sorties brutes sont simplifiées.

4.1 Fraction régulière

Commençons par l'exemple de la Section 2.1. Sous PLANOR, la recherche de la matrice clé est effectuée en utilisant la fonction `planor.designkey`. Par défaut la recherche s'arrête dès qu'une solution a été trouvée. Le résultat est stocké dans un nouvel objet appelé ici `ABCD.key`. Les confusions d'effets peuvent être déterminées en appliquant la fonction `alias` à cet objet.

```
> library("planor",lib="/home/hmonod/R")
> ABCD.key <- planor.designkey(factors=c("A","B","C","D"),
+                              nlevels=2,
+                              model=~(A+B+C+D)^2,
+                              estimate=~A+B+C+D,
+                              nunits=8,
+                              base=~A+B+C)

Determination of ineligible factorial terms
Determination of ineligible pseudofactorial terms
Independent searches for prime(s) :  2
Key-matrix search for prime p = 2
There are  3  predefined columns
First visit to column  4
The search is closed: max.sol =  1 solution(s) found

> print(ABCD.key)

--- Solution  1  for prime  2  ---

  A B C D
A 1 0 0 1
B 0 1 0 1
C 0 0 1 1

> alias(ABCD.key, model=~(A+B+C+D)^2)

--- Solution  1  for prime  2  ---

UNALIASED TREATMENT EFFECTS
A ; B ; C ; D

ALIASED TREATMENT EFFECTS
A B = C D
A C = B D
A D = B C
```

Pour construire le plan d'expériences, on choisit maintenant d'appliquer une randomisation totale des unités, en appliquant la fonction `planor.design` à l'objet `ABCD.key` contenant la matrice clé.

```
> ABCD.design <- planor.design(ABCD.key, randomize = ~UNITS)@design

> print(ABCD.design)

  UNITS A B C D
1     1 2 1 2 1
2     2 2 1 1 2
3     3 1 1 1 1
4     4 2 2 1 1
5     5 2 2 2 2
6     6 1 2 2 1
7     7 1 2 1 2
8     8 1 1 2 2
```

4.2 Plan en carré latin

Pour le carré latin vu en Section 2.2, nous employons, pour exemple, la fonction `planor.factors` pour déclarer les facteurs. Par ailleurs, l'exemple est légèrement modifié afin de montrer que PLANOR n'est pas limité à des facteurs dont le nombre de modalités est un même nombre premier. On recherche donc un carré latin en 6 lignes et 6 colonnes, pour 6 traitements correspondant au produit de deux facteurs A et B à 2 et 3 modalités respectivement.

```
> set.seed(123)
> cl.fact <- planor.factors(factors=
+                           list(Ligne=c("L0","L1","L2","L3","L4","L5"),
+                                Colonne=c("C0","C1","C2","C3","C4","C5"),
+                                A=1:2,
+                                B=1:3))
> cl.key <- planor.designkey(factors=cl.fact,
+                            model=~Ligne + Colonne + A*B,
+                            nunits=36,
+                            base=~Ligne + Colonne)

Determination of ineligible factorial terms
Determination of ineligible pseudofactorial terms
Independent searches for prime(s) :  2 3
Key-matrix search for prime p = 2
There are  2  predefined columns
First visit to column  3
The search is closed: max.sol =  1 solution(s) found
Key-matrix search for prime p = 3
There are  2  predefined columns
First visit to column  3
The search is closed: max.sol =  1 solution(s) found
```

```
> print(cl.key)

********** Prime  2  design **********

--- Solution  1  for prime  2  ---

          Ligne_1 Colonne_1 A
Ligne_1         1         0 1
Colonne_1       0         1 1


********** Prime  3  design **********

--- Solution  1  for prime  3  ---

          Ligne_2 Colonne_2 B
Ligne_2         1         0 1
Colonne_2       0         1 1
```

La recherche de la matrice clé a été décomposée entre une recherche relative au nombre premier 2 et une recherche relative au nombre premier 3, comme attesté par les retours en cours d'exécution et par la commande `print(cl.key)`. Mais au final, ces étapes techniques sont à peu près transparentes pour l'utilisateur uniquement intéressé par le plan d'expériences. Dans la mesure où une solution est trouvée pour la matrice clé, celui-ci peut passer directement à la construction du plan d'expériences randomisé avec la fonction `planor.design`. Pour un carré latin, il doit randomiser indépendamment les lignes et les colonnes.

```
> cl.design <- planor.design(cl.key, randomize=~Ligne+Colonne)@design
```

Après quelques manipulations sous R reproduites en annexe, on obtient le carré latin du Tableau 4.

Traitements		Colonne					
		$C0$	$C1$	$C2$	$C3$	$C4$	$C5$
Ligne	$L0$	$A2\,B2$	$A1\,B1$	$A1\,B3$	$A2\,B3$	$A1\,B2$	$A2\,B1$
	$L1$	$A2\,B3$	$A1\,B2$	$A1\,B1$	$A2\,B1$	$A1\,B3$	$A2\,B2$
	$L2$	$A1\,B1$	$A2\,B3$	$A2\,B2$	$A1\,B2$	$A2\,B1$	$A1\,B3$
	$L3$	$A1\,B3$	$A2\,B2$	$A2\,B1$	$A1\,B1$	$A2\,B3$	$A1\,B2$
	$L4$	$A2\,B1$	$A1\,B3$	$A1\,B2$	$A2\,B2$	$A1\,B1$	$A2\,B3$
	$L5$	$A1\,B2$	$A2\,B1$	$A2\,B3$	$A1\,B3$	$A2\,B2$	$A1\,B1$

TABLE 4 – Plan en carré latin construit par PLANOR pour deux facteurs traitements A et B à 2 et 3 niveaux, en 6 lignes et 6 colonnes.

4.3 Plan de type split-plot

Il arrive que des contraintes expérimentales ne permettent de faire varier un facteur d'intérêt qu'entre des groupes contigus d'unités expérimentales. C'est le cas du split-plot (ou expérience en

parcelles divisées), bien connu dans le domaine des plans d'expériences agronomiques (Dagnelie, 2003). Pour analyser de telles expériences, il faut tout d'abord utiliser la randomisation appropriée à la structure en blocs des unités, puis appliquer un modèle mixte ou modèle d'analyse de variance décomposée en strates (Bailey, 2008, chapitre 8).

Ce type de plans d'expériences dépasse le cadre de cette introduction aux plans réguliers, mais il fait partie de ce que PLANOR permet de rechercher. Un exemple dans le domaine de la désinfection de surfaces en agro-alimentaire (Brouillaud-Delattre *et al.*, 1994) est détaillé dans Kobilinsky (2005) et Kobilinsky *et al.* (2011). Dans cet exemple, les unités expérimentales sont des surfaces specimens circulaires réparties en lignes-colonnes sur des plaques. Pour des raisons pratiques, certains facteurs d'intérêt ne peuvent varier qu'entre plaques, ou qu'entre lignes ou colonnes de ces plaques.

5 Discussion

Les plans factoriels réguliers regroupent de nombreux plans d'expériences orthogonaux, utilisés en pratique dans de nombreux domaines d'application, pour des expérimentations réelles mais aussi pour des expérimentations purement numériques (Lurette *et al.*, 2009 ; Courcoul *et al.*, 2011). Il existe aussi, bien sûr, de nombreuses situations dans lesquelles un plan orthogonal ne convient pas. D'autres approches des plans d'expériences doivent alors être mobilisées, telles que les plans optimaux ou les plans pour surfaces de réponses.

Les exemples donnés dans cet article sont de construction relativement simple, à la portée de statisticiens ou d'expérimentateurs connaissant la théorie des plans d'expériences. Cependant, il est difficile voire impossible de déterminer les relations de définition dans des situations plus complexes. Il existe dans ce cas des résultats théoriques (Mukerjee et Wu, 2006), mais ils ne sont pas forcément faciles à identifier par le non spécialiste et ne correspondent pas toujours au problème précis que l'on veut résoudre. La librairie R PLANOR offre un moyen alternatif et souple pour explorer les solutions possibles. Elle permet également d'aller au bout de la démarche en construisant et en randomisant le plan d'expériences.

Addendum : la librairie R PLANOR est en libre accès sur le site de l'unité MIA-Jouy de l'INRA : `http://w3.jouy.inra.fr/unites/miaj/public/logiciels/planor/`
Il est prévu de la rendre accessible sous peu sur le CRAN. Par ailleurs, des applications très concrètes de la 1ère version de PLANOR ont fait l'objet de témoignages lors d'une journée organisée en l'honneur d'André Kobilinsky, le 5 mai 2011, par l'unité MIA-Jouy :
`http://w3.jouy.inra.fr/unites/miaj/public/matrisq/jbdenis/jbd/11_05_02/ak.html`

Bibliographie

Bailey, R. A. (2008). *Design of Comparative Experiments*. Cambridge, University Press.

Brouillaud-Delattre, A., Kobilinsky, A., Cerf, O. (1994). Méthode de mesure de l'efficacité des procédés de nettoyage et de désinfection des surfaces ouvertes. *Lait*, **74**, 79-88.

Courcoul, A., Monod, H., Nielen, M., Klinkenberg, D., Hogerwerf, L., Beaudeau, F., Vergu, E. (2011). Modelling the effect of heterogeneity of shedding on the within herd *Coxiella burnetii* spread and identification of key parameters by sensitivity analysis. *Journal of Theoretical Biology*, **284**, 130-141.

Dagnelie, P. (2003). *Principes d'expérimentation. Planification des expériences et analyse de leurs résultats.* Gembloux, Presses Agronomiques.

Fisher, R. A. (1942). The theory of confounding in factorial experiments in relation to the theory of groups. *Annals of Eugenics*, **11**, 341-353.

Franklin, M., Bailey, R.A. (1977). Selection of Defining Contrasts and Confounded Effects in Two-level Experiments. *Applied Statistics*, **26**, 321-326.

Kobilinsky, A. (1997). Les plans factoriels. *In :* Plans d'expériences : applications à l'entreprise (Droesbeke, J.J., Fine, J. et Saporta, G. éds.), pp. 69-209 (Chapitre 3). Technip, Paris.

Kobilinsky, A. (2005). *PLANOR : program for the automatic generation of regular experimental designs. Version 2.2 for Windows.* Technical Report. MIA Unit, INRA Jouy en Josas.

Kobilinsky, A., Bouvier, A., Monod, H. (2012). *PLANOR : an R library for the automatic generation of regular fractional factorial designs.* Technical Report. MIA Unit, INRA Jouy en Josas.

Kobilinsky, A., Monod, H., Bailey, R.A. (2011). Automatic generation of regular factorial designs. *In prep.*

Lurette, A., Touzeau, S., Lamboni, M., Monod, H. (2009). Sensitivity analysis to identify key parameters influencing Salmonella infection dynamics in a pig batch. *Journal of Theoretical Biology*, **258**, 43-52.

Mukerjee, R., Wu, C.F.J. (2006). *A Modern Theory of Factorial Designs.* Springer :Berlin.

Patterson, H., Bailey, R.A. (1978). Design keys for factorial experiments. *Applied Statistics*, **27**, 335-343.

Pistone, G. Rogantin, M.-P. (2008). Indicator function and complex coding for mixed fractional factorial designs. *Journal of Statistical Planning and Inference*, **138**, 787-802.

Yates, F. (1933). The principles of orthogonality and confounding in replicated experiments. *Journal of Agricultural Science*, **23**, 108-145.

Yates, F. (1937). *Design and Analysis of Factorial Experiments.* London, Imperial Bureau of Soil Science.

Annexe : commandes R pour la mise en forme du carré latin de la Section 4.2

```
> ## 1. Lignes du data.frame contenant le plan
> ##    reordonnees par lignes et colonnes du carre latin
> cl.design <- cl.design[order(cl.design$Ligne,cl.design$Colonne),]
> ## 2. Agregation des colonnes du plan associees a A et B
> toprint <- paste("A",cl.design$A," B", cl.design$B, sep="")
> ## 3. Mise sous forme de matrice 6x6 du plan
> toprint <- matrix( toprint, nrow=6, ncol=6, byrow=TRUE )
> ## 4. Impression a l'ecran
> print(toprint)
```

Session 2 : Analyse de Risque I / *Risk Analysis I*

Observations, sensitivity and Bayesian inference in QMRA

Jukka Ranta[1,2], Antti Mikkelä[1], Pirkko Tuominen[1]

[1] *Finnish Food Safety Authority Evira*
E-mail : *jukka.ranta@evira.fi*

[2] *Helsinki University*
E-mail : *jukka.ranta@helsinki.fi*

Abstract
In QMRA, observable data are often results of microbial testing methods, typically applied in official control programme sampling schemes but also in other types of sampling. The sensitivity of detection may critically depend on underlying unknown conditions. For statistical analysis towards estimation of risk, Bayesian methods have been used to account for variations in overall test sensitivity and combining evidence from various data sources.
Keywords : Bayesian inference, microbial, test sensitivity, risk assessment, QMRA, food safety

1 Introduction

Quantitative food safety assessment models aim to describe the stochastic chain of events from the origin of the hazard to the outcome. However, substantial uncertainties are involved. Therefore, data based probabilistic inference is challenging but essential for the full assessment of all evidence and the associated uncertainties. Observed data represent our evidence we have about a current situation in a food production chain, e.g. egg production, under risk of e.g. Salmonella. Bayesian inference is a way of computing probabilities of the unknowns, based on available evidence. Some of the evidence is related to background information, some to the actual observable situation. In Bayesian inference, the former becomes prior, the latter will constitute our data. We aim to compute probabilities conditional to all data, e.g. various reported testing results in a control programme. However, the information in those samples can largely rely on sensitivity of the whole testing scheme which can also change over time. This overall test sensitivity is key parameter to be modeled. Examples from Quantitative Microbial Risk Assessment (QMRA) projects for Salmonella are given.

1.1 Unidentifiable parameters arising in QMRA

Effectively, data will constrain any Bayesian model so that unknown model parameters and variables become estimated not only from the most directly related data but also by indirect data from connected variables. The exact probabilistic mechanism of such holistic inference is determined by the connections in the whole dependency structure of the model. This is an advantage when evidence is fragmented and composed of multiple data sources, each contributing differently to the total evidence. In this presentation, the problem involves two basic elements to be modeled and then combined: the unknown prevalence to be estimated and the unknown sensitivity to be accounted for. Clearly, problems of parameter identifiability occur if nothing

more could be assumed of either quantity and if the only data would be observable testing results which are dependent on both parameters.

2 The starting point

In a typical diagnostic problem, we aim to estimate the unknown prevalence q based on the observed positives x in a sample of size n, but accounting for imperfections in the testing method. The testing method is defined by its sensitivity and specificity. Expert opinion or literature is usually available to provide estimates of both. In a Bayesian analysis, informative prior distributions would then be constructed to express such knowledge. In the following, we shall consider only the case where sensitivity is uncertain, and specificity is assumed to be 100%. This leads to the two-dimensional posterior distribution of

$$\pi(q, p \mid x, n) \propto \binom{n}{x} (qp)^x (1 - qp)^{n-x} \pi(q)\pi(p)$$

where the prior density of q is uniform, and the prior of p is Beta-density, reflecting plausible values for sensitivity. This is the simplest example where we take advantage of some prior information about p, i.e. the nuisance parameter, but assume no prior information of the parameter of interest, q. Other versions of a related problem may assume no observations x, but independent binomial data for q and p, thus allowing direct Monte Carlo sampling of both parameters and hence any functions of these. For example, apparent prevalence $\theta = qp$, similarly to the example in [1]. However, in a risk assessment context we can assume only data that are dependent of both unknown parameters. In this case, a Gibbs sampler can be constructed, or we can take advantage of WinBUGS/OpenBUGS software for the Markov Chain Monte Carlo sampling (MCMC). Marginal posterior density of q is then easily obtained from the MCMC sample. In a risk assessment context, the overall sensitivity parameter is often more complicated than could be expressed by the beta-density.

2.1 Overall annual sensitivity: Salmonella testing for cattle herds

As control programmes provide data from several sources, not all based on statistically rigorously designed and implemented sampling schemes, it occurs that informative data appear in an incomplete form which makes the analysis more challenging. As an example, consider estimating true herd prevalence for a Salmonella risk assessment. Herd level sampling can occur for different reasons. To simplify, there are tests done for clinical symptoms (CS) and tests for other reasons (NCS). The latter may be assumed to represent nearly independent sampling, but the former is selective for Salmonella. Estimation requires modeling of annual overall sensitivity, including probability to be selected in either type of testing. Moreover, selective sampling is mostly based on individual samples whereas the other is based on pooled samples, with variations in pool size k. The model for overall sensitivity becomes the sum $P(\mathrm{CS} \setminus \mathrm{NCS})p_\mathrm{f} + P(\mathrm{NCS} \setminus \mathrm{CS}) \sum_{k=1}^{K} (1 - (1-p_\mathrm{w})^k) p_\mathrm{f} P(k) + P(\mathrm{CS} \cap \mathrm{NCS}) \Big[1 - (1-p_\mathrm{f}) \Big(1 - \sum_{k=1}^{K} (1 - (1-p_\mathrm{w})^k) p_\mathrm{f} P(k)\Big)\Big]$. This is a function of several other parameters, each connected to separate additional (annual) data which can be pulled together in the complete Bayesian model [2].

2.2 Dynamic sensitivity: Salmonella testing times for laying hens

Salmonella control programmes for laying hens specify several sampling times over the egg laying period. The sensitivity at any testing event is hard enough to quantify, but it is likely to change over time too. Expert opinions can be drawn to assess the sensitivity, but they are also uncertain and can depend on several other unknown factors. Although the sensitivity of the sampling type and the associated laboratory method may be assumed fairly high for designed conditions, the overall sensitivity of detecting an infected flock is more complicated function of e.g. within flock infection process. After beginning of infection, there can be a short period during which the new infection might not be detectable yet. After this, the detection probability can peak, possibly to decline later again. Here, we sketch only simple scenarios for sensitivity, and then compute results under each scenario. The subleties of detection problems have been discussed e.g. in an EFSA report on Salmonella in laying hens [3].

For the flock population infections, the two-state Markov process model, [4], for the latent infection I_t assumes that flocks can move from susceptible state to infected state and back, with intensities λ (for infection) and μ (for recovery). However, in the context of Salmonella infections, it is known that recovery of an entire flock can be slow, even slower than its lifetime in production. Therefore, μ is likely to be small. Moreover, in a low prevalence production conditions where Salmonella is rare, we can expect also flock infections to be rare. Thus, λ is likely to be small too. According to the Salmonella control programme, detected infected flocks are to be destroyed. The observed history of testing results for any flock is therefore a series of negative tests, possibly (but rarely) ending with a positive test. Our interest is to quantify the true prevalence while accounting for complicated overall sensitivity of the testing.

Let τ_0 denote the unknown time of *last* infection before standing infected, $I_t = 1$, at testing time t. In a very simplified model, detection probability would jump to its maximum value, p, after the onset of infection, and drop to zero after some more time:

$$P(\oplus_t \mid \tau_0, I_t = 1) = \begin{cases} 0 & \text{, if } t - \tau_0 < d_1 \\ p & \text{, if } d_1 < t - \tau_0 < d_2 \\ 0 & \text{, if } t - \tau_0 > d_2 \end{cases}$$

Taking into account the uncertainty about the start of infection τ_0, $(0 < \tau_0 < t)$ we calculate the integral

$$p_t = P(\oplus_t \mid I_t = 1) = \int_0^t P(\oplus_t \mid \tau_0, I_t = 1)\pi(\tau_0 \mid I_t = 1)\mathbf{d}\tau_0,$$

where, approximately,

$$\pi(\tau_0 \mid I_t = 1) = \frac{e^{-(t-\tau_0)\mu}}{(1 - e^{-\mu t})/\mu}.$$

The integration results to:

$$p_t = \frac{pe^{-\mu t}}{1 - e^{-\mu t}}\left(e^{\mu \max\{t-d_1,0\}} - e^{\mu \max\{t-d_2,0\}}\right).$$

With several testing times, by time t, we also know the number of negative test results so far, and their timing. The likelihood becomes increasingly difficult, but approximately this could

be built into the likelihood of the latent variable $\pi(I_t = 1 \mid \tau_0)$ by setting increasing weights for time periods between the observed testings:

$$\pi(I_t = 1 \mid \tau_0) = e^{-(t-\tau_0)\mu}(1-p)^{K_t - i + 1} \quad \text{if } \tau_0 \in [t_{i-1}, \min(t_i, t)], i = 1, \ldots, K_t + 1$$

where $K_t - i + 1 \in [0, K_t]$ is the number of tests done between τ_0 and current time t. This is a simplification because in place of the constant p we should have a sensitivity that depends on the age: p_t. This would be obtained after integrating over the unknown starting time of infection τ_0, by using some formulation of sensitivity as a function of duration of infection. With the above simple step function, the computations become cumbersome. Therefore, continuous functions may be preferred. Several qualitatively different functions may be useful, depending on the epidemiological assumptions we are willing to make. One might be based on the curve of a Gaussian function, reflecting a peak sensitivity at some optimal duration of infection $d = t - \tau_0$. This requires two parameters to be specified: the optimal duration d^*, and the steepness of the Gaussian function σ:

$$p(t - \tau_0) = p(d) = pe^{-0.5(d-d^*)^2/\sigma^2}$$

Another option is to assume the sensitivity can only increase as a function of duration of infection. Hence, one might choose:

$$p(d) = p(1 - e^{-ad}).$$

This requires only one parameter, a, to specify the steepness. One more option is to assume the sensitivity can initially increase to its maximum value, but eventually start declining again:

$$p(d) = p(1 - e^{-ad})e^{-a\max(0, d-d^{**})}.$$

This requires two parameters, for the steepness and the lag time after which the decline will start. The goal is to keep the number of parameters as small as possible, to enable expert elicitation that would be based on robust epidemiological assumptions. As it may still be difficult to elicit the parameters, we may resort to considerations of worst case and best case assumptions. Whatever the choice, these continuous functions can then be used for solving the sensitivity as a function of age, p_t, by integration over the unknown time of infection τ_0. For example, with the simple Gaussian function:

$$p_t = P(\oplus_t \mid I_t = 1) = \int_0^t pe^{-0.5(t-\tau_0-a)^2/\sigma^2} \frac{e^{-(t-\tau_0)\mu}}{(1-e^{-\mu t})/\mu} \mathbf{d}\tau_0$$

$$= \frac{p\mu}{1-e^{-\mu t}} e^{-0.5(a^2-(a-\mu\sigma^2)^2)/\sigma^2} \sqrt{2\pi}\sigma[\Phi((t-a+\mu\sigma^2)/\sigma) - \Phi((-a+\mu\sigma^2)/\sigma)].$$

Finally, when accounting for the past negative tests, the density of τ_0 is solved by numerical integration, and then the sensitivity function as a second numerical integration. For example, to compute the sensitivity at a time point between the first and the second testing time, $t \in [t_1, t_2]$, assuming the Gaussian function, we obtain

$$\pi(\tau_0 \mid I_t = 1) \propto \begin{cases} e^{-(t-\tau_0)\mu}(1 - pe^{-0.5(t_1-\tau_0-a)^2/\sigma^2}) & \text{if } \tau_0 \in [0, t_1] \\ e^{-(t-\tau_0)\mu} & \text{if } \tau_0 \in [t_1, t], t < t_2 \end{cases}$$

The normalizing constant for this is obtained by numerical integration, (not simulation), available in OpenBUGS. Having this computed, the sensitivity is obtained by another integration to get

$$p_t = \int_0^t p e^{-0.5(t-\tau_0-a)^2/\sigma^2} \pi(\tau_0 \mid I_t = 1) \mathbf{d}\tau_0,$$

again, by using the numerical integration tool in OpenBUGS. In this way, we avoid convergence problems and slower simulation of the unknown times of infection τ_0 for each flock. On the other hand, we need to set a tolerance limit for the errors in the numerical integration and a too small value would slow down the computations again.

Having computed the sensitivity as a function of age, we return to our original inference problem which involves the uncertain sensitivity and the uncertain prevalence. In place of the prevalence parameter, we now have the probability of the latent infection at a given time. Given the past history of negative results, the probability of positive detection at testing time t_k can be written as

$$P(\oplus_{t_k}) = P(D_{t_k} = 1 \mid H^0_{t_{k-1}}) = p_{t_k} P(I_{t_k} = 1 \mid H^0_{t_{k-1}}) = p_{t_k}(p_{01}(1 - \rho_{t_{k-1}}) + p_{11}\rho_{t_{k-1}}),$$

where p_{ij} is the transition probability from state i to state j over two testing times. The probability for hidden infection ρ is solved recursively, [5] and the sensitivity p_{t_k} is computed at testing times t_k. In total, data from a Salmonella control programme can be modeled as a product of binomial distributions each with a probability parameter of the form above. The posterior distribution is simulated by OpenBUGS for the intensity parameters μ, λ, the peak sensitivity p, and the initial probability of infection v, with partially informative priors. Posterior distributions of other quantities of interest can be derived from these by inspection of the corresponding MCMC samples. Some example scenarios will be computed to provide estimates in a low prevalence situation where all the test results from a control programme are negative; yet we need to quantify the microbial risk due to a possibly nonzero true prevalence. Eventually, different sampling designs of possible new control programmes could be compared.

References

[1] Berger, J. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, 1(3), 385-402.

[2] Ranta, J., Tuominen, P., Maijala, R. (2005). Estimation of true salmonella prevalence jointly in cattle herd and animal populations using Bayesian hierarchical modeling. *Risk Analysis*, 25(1), 23-37.

[3] EFSA Panel on Biological Hazards (BIOHAZ); Scientific Opinion on a quantitative estimate of the public health impact of setting a new target for the reduction of Salmonella in laying hens. EFSA Journal 2010; 8(4):1546. [86 pp.]. doi:10.2903/j.efsa.2010.1546. Available online: www.efsa.europa.eu

[4] Karlin, S., Taylor, H.M. (1975): *A First Course in Stochastic Processes*. Second Edition. Academic Press, San Diego.

[5] Nagelkerke, N.J., Chunge, R.N., Kinoti, S.N. (1990): Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine*, 9, 1211-1219.

Association d'approches déterministes et stochastiques appliquée à la chaine du froid des produits alimentaires

Combined deterministic and stochastic approaches applied to the food cold chain

Onrawee Laguerre[1], Minh Hong Hoang[1], Evelyne Derens[1], Graciela Alvarez[1], Denis Flick[2]

[1] Irstea, Refrigeration Process Engineering, 92600 Antony, France
onrawee.laguerre@irstea.fr
[2] AgroParisTech, INRA, Cnam: UMR 1145, Food and Process Engineering, F-91300 Massy, France

Résumé

Plusieurs enquêtes ont montré que la température de produits dans les 3 dernières étapes de la chaîne du froid est problématique: meuble frigorifique de vente, transport par les consommateurs après l'achat et réfrigérateur domestique. Cette étude a été effectuée pour proposer une méthodologie de prédiction de l'évolution de la température du produit et de la charge microbienne tout au long de la chaine du froid. Les modèles déterministes développés prennent en compte le transfert de chaleur par convection, conduction et rayonnement. Ceux-ci ont été combinés avec des modèles stochastiques pour tenir compte de différents paramètres aléatoires: la température ambiante, le réglage du thermostat, la position et le temps de séjour du produit dans les équipements etc. Des lois de distribution ont été développées pour ces paramètres aléatoires par ajustement sur des données d'enquête. Les valeurs échantillonnées de ces lois ont été utilisées comme paramètres d'entrée du modèle de transfert de chaleur permettant de prédire les évolutions de température d'un grand nombre de produits le long de la chaîne du froid. Ces évolutions de température ont été utilisées dans un modèle de croissance microbienne pour prédire les évolutions de la charge de *Listeria monocytogenes*. Des analyses statistiques des résultats ont été réalisées permettant de connaître le pourcentage de produits ayant différents niveaux de contamination.

Mots-clés : chaine du froid, sécurité des aliments, température, charge microbienne, prédiction.

Abstract

Several surveys have shown temperature abuses in the last 3 links of the cold chain: display cabinet, transport by consumer and domestic refrigerator. This work was carried out to predict the product temperature and microbial load evolutions in these steps. Deterministic models were used to take into account the heat transfer by convection, conduction and radiation inside the equipments. They were combined with stochastic models to take into account different sources of randomness: ambient temperature, thermostat setting, product position and residence time in the equipments etc. Distribution laws were developed to fit the survey data of these random parameters. The sampling values were used as input parameters of the heat transfer models to predict the temperature evolutions of a large number of products along the cold chain. Then, these temperature evolutions were applied to a growth model of *Listeria monocytogenes* to predict the bacterial contamination. Statistical analysis of

simulation results was carried out enabling the distribution of product at various contamination levels.

Keywords : cold chain, food safety, temperature, microbial load, prediction

1. Introduction

In refrigeration of food products, temperature control along the cold chain is essential to maintain the product quality. To provide safe food products of high organoleptic quality, attention must be paid to every aspect of the cold-chain from the production until the consumption. As the product is present in several equipments during the supply-chain, it is difficult to control and maintain the temperature all along the cold chain.
The objective of this study is to develop a prediction methodology for the product time-temperature history and the evolution of microbial load along the cold chain. This methodology takes into account different sources of randomness in the cold chain: ambient temperature (may vary due to the seasons, presence/absence of air conditioning system), thermostat setting (may vary due to the behaviour of operator/consumer), product position and residence time in the equipments (may vary due to the strategy of logistic management) etc. In the present paper, the methodology was applied to the 3 last links of the cold chain: refrigerated display cabinet, transport by consumer after purchase, domestic refrigerator where temperature abuse was often observed (Figure 1).

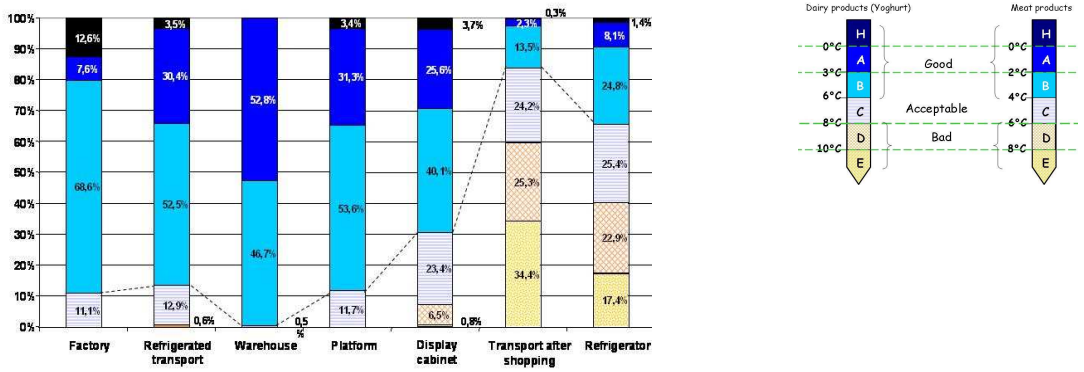


Figure 1: Percentage of dairy and meat products in different temperature ranges according to the Cemagref and Ania survey (2004).

2. Deterministic models

2.1 Simplified steady state heat transfer models for load

Simplified steady state heat transfer models were developed to predict the temperature of the load (T_{load}) located at different positions in the display cabinet and domestic refrigerator. The models take into account the heat transfer by convection, conduction and radiation.

2.1.1 Domestic refrigerator (Ref)

The proposed simplified heat transfer model of loaded refrigerator represents the main phenomena observed by CFD simulation (Laguerre et al., 2007): circular airflow in the cavity, temperature

stratification along the height. Figure 2 presents a diagram of simplified airflow and heat transfer in steady state (Laguerre et al., 2010). During the flow, air exchanges heat with cold wall, with bottom load, with warm wall and with top load, simultaneously. There is also radiation between cold wall and load and between warm wall and load. Finally, there is conduction in the door and the side walls and convection with the external air.

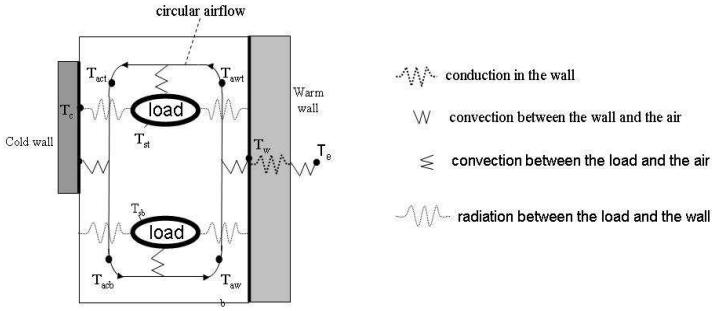


Figure 2: Simplified model of heat transfer and airflow in loaded refrigerator.

The static domestic (without a fan, free convection) and ventilated refrigerators (forced convection) are studied in this work. In the static type, there is about 4°C temperature difference between top and bottom (Laguerre and Flick, 2004) while this difference is lower in a ventilated appliance.

The load temperatures depend on the position and on two random parameters: the air temperature in the kitchen, T_{ext} and the thermostat setting, T_{th} which depends on consumer habits. In order to take into account the non uniformity of product temperature, 2 load positions are considered: top and bottom.

The load temperatures can be calculated from T_{ext} and T_{th} by following relations (Laguerre and Flick, 2010):

- For static refrigerator with two load positions : *l*=1 top, *l*=2 bottom

$$T_{load.1} = 0.0723T_{ext} + 0.9277T_{th}$$
$$T_{load.2} = 0.0077T_{ext} + 0.9923T_{th} \quad (1)$$

- For ventilated refrigerator with two load positions : *l*=1 top, *l*=2 bottom

$$T_{load.1} = 0.0343T_{ext} + 0.9657T_{th}$$
$$T_{load.2} = 0.0147T_{ext} + 0.9853T_{th} \quad (2)$$

2.1.2 Display cabinets (DC)

In this work, only the open front vertical display cabinet is studied because it is mostly used for keeping chilled food in supermarket (Gac and Gautherin, 1987). In order to take into account the non-uniformity of product temperature, 4 load positions are considered: rear top, rear bottom, front top and front bottom (Figure 3). The details of the development and validation of the thermal model were presented in Laguerre *et al.*, 2012.

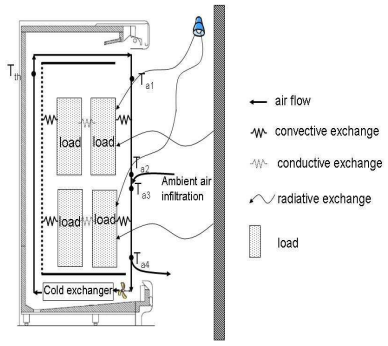


Figure 3. Heat transfer model considering 4 load positions in an open vertical display cabinet.

The load temperatures depend on the position and two random parameters: the air temperature in the store T_{ext} and the radiation temperature T_{rad}. The load temperature can be calculated from the equipment parameters by linear relations (Laguerre *et al.*, 2011):

- product placed in rear part (DC/rear), *1*=top, *2*=bottom

$$\begin{aligned} T_{load.1} &= 0.0027T_{ext} + 0.0430T_{rad} + 1.4655 \\ T_{load.2} &= 0.0244T_{ext} + 0.0435T_{rad} + 1.4814 \end{aligned} \quad (3)$$

- product placed in front part (DC/front), *1*=top, *2*=bottom

$$\begin{aligned} T_{load.1} &= 0.0117T_{ext} + 0.1831T_{rad} + 1.3534 \\ T_{load.2} &= 0.1040T_{ext} + 0.1889T_{rad} + 1.4178 \end{aligned} \quad (4)$$

2.2 Simplified transient heat transfer model for product of interest

A simplified transient heat transfer model was developed to predict the temperature evolution of the product *T(t)* which circulates along different links of the cold chain (called product of interest). This model takes into account the temperature of the load located near the product of interest (T_{load}), product properties: weight (m), surface area (A) and specific heat (C), the product initial temperature (T_0), and the heat transfer coefficient between the products and surrounding air (by natural or forced convection).

Inside display cabinet, shopping basket and domestic refrigerator, using a lumped thermal model, the mean temperature of the product of interest tends to the neighbouring load temperature (Figure 4). The temperature evolution of the product of interest can be expressed by exponential equation below:

$$T(t) = T_0 + \left(T_{load,l,k} - T_0\right) exp\left(-\frac{H_{l,k}t}{mc}\right) \quad (5)$$

$$\Rightarrow ln\left(T^*\right) = -\left(\frac{t}{\tau_{l,k}}\right) \quad (6)$$

where $\tau_{l,k} = \frac{mc}{H_{l,k}}$, $T^* = \frac{T(t) - T_{load,l,k}}{T_0 - T_{load,l,k}}$, $H_{l.k}$ = heat transfer conductance

l=position index, *l=1* (top), *l=2* (bottom),
k= link index, k=1 display cabinet with rear load, k=2 display cabinet with front load, k=3 shopping basket, k=4 static domestic refrigerator, k=5 ventilated domestic refrigerator.

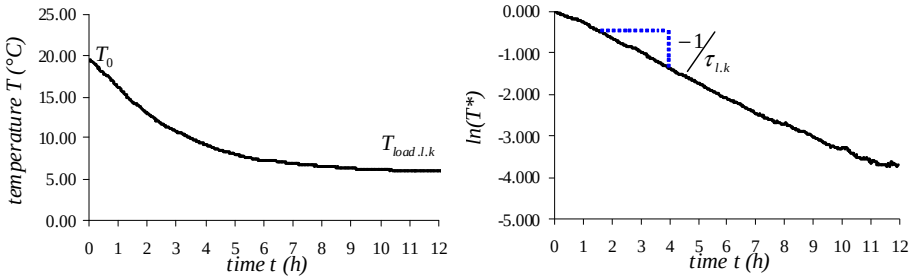


Figure 4. Temperature evolution of product of interest
(a) temperature *T* vs time *t* - (b) $ln(T^*)$ vs time *t*

The heat transfer conductance, $H_{l.k}$ was measured experimentally for each position *l* of each link k. Packaged meat (*m*=0.250 kg and *C*=3500 $J.kg^{-1}K^{-1}$) was used to determine $H_{l.k}$,. During experiment, the product was placed at a position *l* of a link *k* and a thermocouple was placed at the product's centre to monitor the evolution of product temperature. Using the eq.6, $ln\left(T^*\right)$ versus t was traced (Figure 4b) and the slope was used to calculate $H_{l.k}$. for each link (Table 1).

$H_{l.k}$ $(W.K^{-1})$	*k*=1 DC/rear	*k*=2 DC/front	*k*=3 Shopping basket*	*k*=4 Static Ref	*k*=5 Ventil Ref
l=1	0.26	0.35	0.15 (non insulated basket) 0.09 (insulated basket):	0.12	0.13
l=2	0.30	0.26	-	0.13	0.13

* one product position is considered in shopping basket

Table 1: Heat transfer conductance $H_{l.k}$ $(W.K^{-1})$ of pre-packaged meat (0.250 kg) in display cabinet, shopping basket and domestic refrigerator.

2.3 Evolution of microbial load N

As a primary model, a simple first order growth was assumed:

$$\frac{dN(t)}{dt} = \mu N(t) \qquad (7)$$

$$\Rightarrow ln\left(\frac{N(t)}{N_0}\right) = \int_0^t \mu.dt' \quad (8)$$

where *N(t)* is the microbial load at time instant t, N_0 is the initial load (CFU/g) and.μ is the specific growth rate (s^{-1}).
The dependence of the specific growth rate μ on the temperature is described by the square root model of Ratkowsky *et al.* (1982):

$$\sqrt{\mu} = b(T - T_{\min}) \quad (9)$$

where T_{min} is the minimum temperature under which there is no bacterial growth and *b* is a coefficient. For *Listeria monocytogenes*, Duh and Schaffner (1993) reported that T_{min}=0°C and *b*=0.00035 $s^{-1/2}K^{-1}$
More sophisticated model proposed by Baranyi et al (1993) and Zwietering et al (1996) could also be used with the same approach.

3. Stochastic models

The distribution laws of the random input parameters of each link were developed to fit the survey data (Table 2). The normal law was used to represent the temperature distributions, the exponential law for residence time distributions. According to Gac and Gautherin (1987), it appears that the radiation temperature is correlated to the ambient air temperature. In supermarkets, 2 cases are generally presented. First case: the radiation temperature is close to that of air when grocery shelves are placed oppositely to the refrigerated display cabinet. Second case, the radiation temperature is averagely 6°C lower than the air temperature when 2 display cabinets are placed oppositely. Therefore, Bernoulli law was chosen with two equally probable values of $T_{rad} - T_{ext}$: $0°C$ and $-6°C$. Investigations in different countries showed that the vast majority of people did not use any means of food protection during transportation: 87.3% in the UK (Evans, 1992); 84.5% in Slovenia (Jevsnik *et al.*, 2008) and 81.4% in New Zealand (Gilbert *et al.*, 2007). So, it is assumed that 84% (mean of these 3 values) of the shopping baskets do not have insulation and 16% are thermally insulated. Bernoulli law was chosen to represent the possibility for the consumer to use the insulated/non insulated basket.

4. Simulations

The logistic chain of meat is shown in Figure 5. The probability of product to be placed at the front (0.2) and the rear (0.8) of a display cabinet was obtained using expert data. The probability of product to be placed in a static (0.68), ventilated refrigerator (0.28) and directly consumed (0.04) was obtained by Diouris and Mahé (2007). There is an equal probability (0.5) to be at the top and the bottom of display cabinet and domestic refrigerator.
The sampling values (by Monte Carlo) of the input random parameters were applied to the steady state heat transfer model enabling the calculation of the load temperature, T_{load} and to the transient heat transfer model enabling the calculation of the product of interest temperature, *T(t).*
The simulation was repeated for numerous product items (10^5) to predict the temperature evolutions in different cold chain equipments until domestic refrigerator. The time-temperature history of the products was applied to a growth model of *Listeria monocytogenes* to predict the bacterial load evolution until consumption. The analysis of the numerical results was carried out and shown in paragraph 5 (Results and discussion).

Link	Random input parameter	Distribution law	Reference
Display cabinet	T_{ext}	N(16.5°C; 1.8°C)	Lindberg *et al.* (2010)
	T_{rad} -T_{ext}	Bernoulli p(0°C)=0.5 p(-6°C)=0.5	Gac and Gautherin (1987)
	Residence time (Δt_{DC}, days)	Exp(3.8)	Cemagref and ANIA (2004)
Shopping basket	T_{ext}	N(17.2°C, 5.8°C)	Cemagref and ANIA (2000)
	H	Bernoulli p(0.15)=0.84 p(0.09)=0.16	Evans, (1992) UK survey Jevsnik *et al.* (2008) Slovenia survey Gilbert *et al.* (2007) New Zealand survey.
	Residence time (Δt_{SB}, days)	Exp(0.05)	Cemagref and ANIA (2004)
Domestic refrigerator	T_{ext}	N(16.7°C, 3.1°C)	Hunt and Gidman (1982)
	T_{th}	N(6.0°C, 2.3°C)	Laguerre et al. (2002)
	Residence time (Δt_{Ref}, days)	Exp(2.8)	Cemagref and ANIA (2004)

Table 2: Distribution law of random input parameters
Normal distribution: N(mean, standard deviation)
Exponential distribution: Exp(mean)
Bernoulli distribution: probability of a random parameter, p(random parameter)

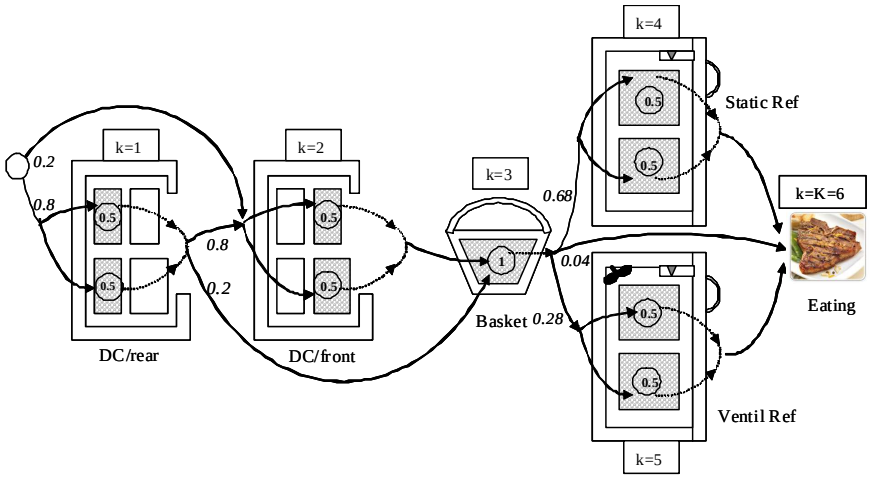


Figure 5**.** Logistic chain: probability of product transfer from an equipment to another (*italic*), probability of product position (**bold** in circle).

5. Results and discussion

The total residence time, *ts*, is the sum of Δt_{link} of the 3 links. It is interesting to compare *ts* with the shelf life of the product. For pre-packaged meat, the shelf life of a product at the end of production is 21 days (Legrand et al., 2010). According to the ANIA investigation, the mean residence time of the product in the upstream links (factory, refrigerated transport and platform) is 3 days. Thus, the shelf life of the product for the final 3 links (display cabinet, shopping basket and refrigerator) is 18 days on the average. Figure 6 presents the cumulative distribution of *ts* obtained from the simulation and the survey. The residence time of 50% of the products is less than 5.5 days and it is less than 18 days for 97% of products.

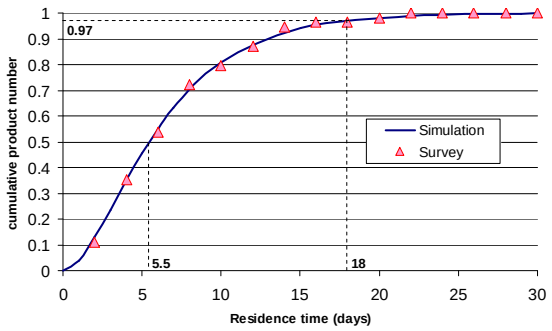


Figure 6. Cumulative distribution of total residence time, *ts*, at the end of the cold chain (display cabinet, shopping basket and domestic refrigerator)
ts ≤ 5.5 days for 50% of product items
≤ 18 days for 97% of product items
(number of product items *I*=100 000)

Figure 7 presents the cumulative distribution of the overall Decimal Increase: $DI = log(N_F / N_I); F = final; I = initial$ in the microbial load of the cold chain for 10^5 product items. It can be observed that for 50% of the products, the microbial load is multiplied by less than 3.31, which is reasonable. The microbial load increases over 200-fold for 5% of products, and this can be critical for consumers. An overall relative increase in the microbial load (N_F / N_I) of more than 10^4 is obtained for 1% of products, which may lead to food poisoning. These products are statistically representative because they represent 1000 items.

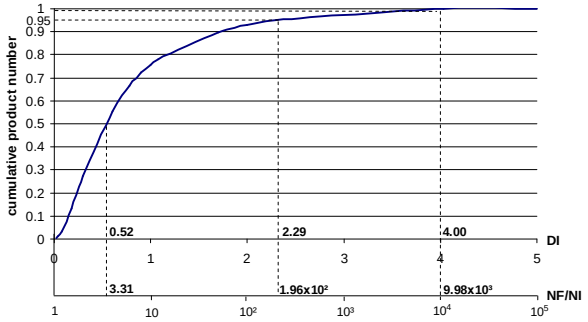


Figure7. Cumulative distribution of the overall decimal increase in the microbial load of the cold chain *DI* (display cabinet, shopping basket and domestic refrigerator)
DI= log (N_F/N_I)≤ 0.52 or $N_F/N_I \leq 3.31$ for 50% of product items
≤ 2.29 or $N_F/N_I \leq 1.96x10^2$ for 95% of product items
≤ 4.00 or $N_F/N_I \leq 9.98x10^3$ for 99% of product items
(number of product items *I*=100 000).

6. Conclusion

A methodology combining deterministic heat transfer models for the refrigeration equipments and stochastic models taking into account various sources of randomness in the cold chain was proposed. It enables the prediction of the temperature evolution of a large number of products. Theses temperature evolutions can then be used to estimate microbial growth. Thus, this approach can contribute to evaluate food safety along the cold chain. It can also be used as a complementary approach of survey which is of high cost and time consumption.

Bibliographie

Cemagref and ANIA (2004). La chaine du froid du fabricant au consommateur: résultats de l'audit ANIA/Cemagref. *Revue Générale du Froid & du conditionnement d'air*, 1042, 29-36.

Baranyi, J., Roberts, T.A., McClure, P.J., 1993. A non-autonomous differential equation to model bacterial growth. Food Microbiology. 10, 43-59.

Diouris, A., Mahé, C. (2007). Réfrigérateurs et congélateurs domestiques: consommer moins et conserver mieux tout en préservant l'environnement. *Revue Générale du Froid & du conditionnement d'air*, 1079, 41-46.

Duh, Y. H., Schaffner D. W. (1993). Modeling the effect of temperature on the growth rate and lag time of *Listeria innocua* and *Listeria monocytogenes*. *Journal of Food Protection*, 56, 205-210.

Evans, J. (1992). Consumer handling of chilled foods: Perception and practice. *International Journal of Refrigeration*, 15(5), 290-298.

Evans, J.A., Scarcelli, S., Swain, M.V.L. (2007). Temperature and energy performance of refrigerated retail display and commercial catering cabinets under test conditions. *International Journal of Refrigeration*, 30, 398-408.

Gac, A. and Gautherin, G. (1987). Le Froid dans les Magasins de Vente de Denrées Périssables, pyc Edition, Paris.

Gilbert, S. E., R. Whyte, G. Bayne, S. M. Paulin, R. J. Lake, P. van der Logt. (2007). Survey of domestic food handling practices in New Zealand. *International Journal of Food Microbiology*. 117(3), 306-311.

Hunt, D.R.G., Gidman, M.I., (1982). A national field survey of house temperatures. *Building and Environment*, 17(2), 107-124.

Jevsnik, M., Hlebec, V., Raspor, P. (2008). Consumers' awareness of food safety from shopping to eating. *Food Control*. 19(8), 737-745.

Laguerre, O., Derens, E., Palagos, B. (2002). Study of domestic refrigerator temperature and analysis of factors affecting temperature: a French survey. *International Journal of Refrigeration*, 25(5), 653-659.

Laguerre, O., Flick, D. (2004). Heat transfer by natural convection in domestic refrigerators. *Journal of Food Engineering*, 62, 79–88.

Laguerre, O., Ben Amara, S., Moureh, J., Flick D. (2007), Numerical simulation of airflow and heat transfer in domestic refrigerators. *Journal of Food Engineering*, 81, 144-156.

Laguerre, O., Flick, D., (2010). Temperature prediction in domestic refrigerator: association of deterministic and stochastic approaches. *International Journal Refrig*eration, 33, 41-51.

Laguerre, O., Derens, E., Flick, D. (2011). Temperature prediction in a refrigerated display cabinet: deterministic and stochastic approaches. *Electronic Journal of Applied Statistical Analysis (EJASA), 4(2)*, 191-202.

Laguerre, O., Hoang, M.H., Flick, D. (2012). Heat transfer modelling in a refrigerated display cabinet: the influence of operating conditions, *Journal of Food Engineering*, 108,353–364.

Legrand, I., Recoules, E. (2010). Emploi du monoxyde de carbone pour le conditionnement d'UVC de bœuf sous atmosphère modifiée. 13èmes Journées « Sciences du Muscle et Technologies des Viandes».

Lindberg, U., Axell, M., Fahlén, P. (2010). Vertical display cabinet without and with doors - a comparison of measurements in a laboratory and in a supermarket. *Sustainability and the Cold Chain*. 29-31, 3/2010, Cambridge, UK.

Ratkowsky, D.A., Olley, J., McMeekin, T.A., Ball, A. (1982). Relationship between temperature and growth rate of bacterial cultures. *Journal of Bacteriol*ogy, 149, 1-5.

Zwietering, M.H., De Koos, J.T., Hasenack, B.E., De Wit, J.C., Van't Riet, K. (1991). Modeling of bacterial growth as a function of temperature. *Applied and Environmental Microbiology*, 57, 1094 -1101.

ACKNOWLEDGEMENT

The research leading to this result has received funding from European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 245288.

Derivation of a variables sampling plan based on a Poisson-Gamma model representing within-batch and between-batch variability in low microbial counts in food

Ursula Gonzales-Barron[1] & Francis Butler

[1] *UCD School of Biosystems Engineering, University College Dublin. Dublin 4, Ireland.*
E-mail: *ursula.gonzalesbarron@ucd.ie*

Abstract
This study proposes a novel methodology for the derivation of a sampling plan for use in food production systems whose low microbial counts can be represented by a correlated random-effects Poisson-gamma model. The methodology presumes the establishment of a criteria defining the batch as unacceptable (i.e., if more than a tolerance percentage of the food units has microbial concentrations exceeding a critical concentration), and proposes the derivation of decision landscape plots representing collectively the producer's and consumer's risks α, β at different microbiological limits m_L (CFU/cm^2) along with their 90% confidence intervals representing the batch-to-batch variability. It was found that for a pre-defined tolerance of a maximum of 2.5% of pre-chill sheep carcasses exceeding a critical *Enterobacteriaceae* concentration of 2 log CFU/cm^2, in order to offer the producer at least a confidence of 90% of accepting reasonable quality batches, a sampling regime of n=14 carcasses and m_L=17 CFU/g should be established in Irish sheep abattoirs.

Keywords: Poisson-gamma, variables sampling plan, *Enterobacteriaceae*, sheep.

1. Introduction

In earlier work, Gonzales-Barron & Butler (2011a) demonstrated that the discrete Poisson-gamma distribution is by far more appropriate than the Poisson-lognormal or lognormal when modelling microbial counts data consisting of zero counts. They further showed (Gonzales-Barron & Butler, 2011b) that the Poisson-gamma model had the ability to overcome the assumption of constant within-batch variance by incorporating the quantification of between-batch variability as a bivariate normal distribution representing the association between within-batch mean and within-batch spread. Gonzales-Barron et al. (2012) also showed that sampling plans with a microbiological limit expressed in arithmetic mean (CFU/g) are more effective that those with the limit expressed in mean log (log CFU/g), since the former tends to yield Operating Characteristic (OC) curves with lower uncertainty about the acceptance probabilities and produces steeper OC curves which is a sign of higher discriminatory power. Thus, it is clear that to establish within-batch testing regimes that are more effective and discriminative, its mathematical derivation should be performed in arithmetic mean scale, which from a technical point of view, is more compatible with the Poisson-gamma model. Building on previous work, this study proposes a novel methodology for the derivation of a within-batch testing regime for use in food production systems whose microbial counts are known to be low and can be represented by a Poisson-gamma model.

2. Methodology

Based on a tolerance criterion, the derivation of the within-batch testing regime is addressed as a classification problem of samples having to discern between an acceptable batch and an unacceptable batch. The method uses the new notion of between-batch variability and the more effective arithmetic mean scale for expressing microbiological limits (Gonzales-Barron et al., 2012); borrows elements from classical variables sampling plans for the establishment of the tolerance criterion (Duncan, 1986), and will be explained in the following sections.

2.1 Modelling within-batch and between-batch variability in microbial counts

The mathematical representation of the within-batch and between-batch variability in *Enterobacteriaceae* counts on pre-chill Irish sheep carcasses was previously detailed in Gonzales-Barron et al. (2012). Briefly, plate count data was available in duplicate (k=2) for twenty carcasses swabbed on each of the four sampling visits to five large Irish abattoirs (n=400 carcasses, j=20 batches). The between-batch and within-batch variability in *Enterobacteriaceae* was modelled by a Poisson-gamma regression model with correlated random effects for the mean (m_j) and dispersion parameter (k_j). The model is defined by five parameters which have been fitted in Gonzales-Barron et al. (2012). By performing a Monte Carlo simulation using the Poisson-gamma model, it is possible to visually assess a 'universe of contaminated batches'. In Figure 1, each point represents a random batch of Irish pre-chill carcasses whose *Enterobacteriaceae* true concentration is characterised by a Gamma (k, m) distribution. Notice that the lower the within-batch mean, the wider the range of dispersion values that the within-batch gamma distribution can take. This is due to the variable proportion of zero counts comprised in the observed distributions of low microbial counts.

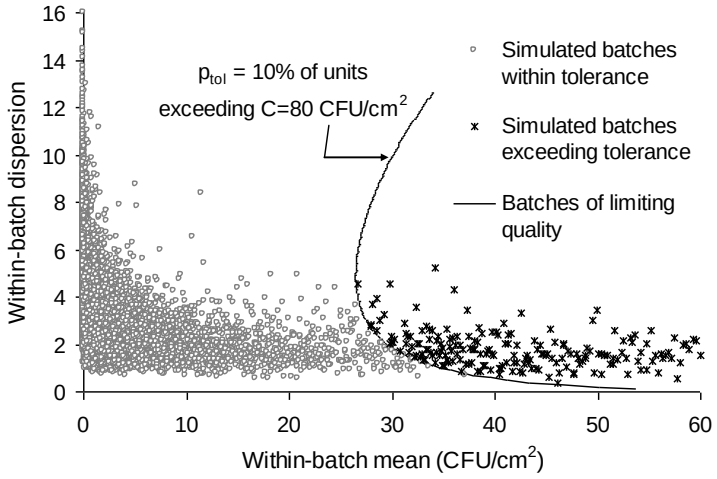


Figure 1: A representation of *Enterobacteriaceae* true concentration in batches of sheep carcasses simulated using the random-effects Poisson-gamma regression model (n=10000) showing a limiting quality contour partitioning batch acceptability and rejectability regions, as derived assuming a tolerance that a maximum of 10% carcasses within a batch can have microbial concentrations exceeding 80 CFU/cm^2

2.2 Definition of a tolerance criterion

In acceptance sampling plan theory, when the underlying distribution of the quality/safety variable (i.e., microbial concentration) is known, it is more efficient to derive sampling plans that make full use of the microbial counts information (i.e., variables sampling plans), rather than ascribing them to categories or classes (i.e., attributes sampling plans). The design of a variables sampling plan however involves a few decisions to be made. The first is to mathematically describe what makes a batch of food unacceptable. In classical variables sampling plans (Dahms, 2004), a batch is defined as unacceptable if more than a tolerance percentage (p_{tol}) of the food units in it has microbial concentrations exceeding a critical concentration (C). At the same time, the purpose of the within-batch testing regime should be known so that its derivation can be made either on the consumer's side (for safety specifications such as Performance Objectives (PO)), or on the producer's side (for control purposes such as adherence to GMP limits or statistical process control). For safety specifications, the probability of accepting a defective batch or consumer's risk β is of main concern, while for GMP specifications, the probability of taking action although the batch is of good quality or producer's risk α is of main concern. When the hygiene and the microbiological safety of the production process is known to be under control, the food producer requires some assurance that products of acceptable quality should not be rejected because of the imprecision of the sampling scheme.

In our case, as the overall concentration of *Enterobacteriaceae* on Irish sheep carcasses was generally low (Gonzales-Barron et al., 2012), the derivation of the within-batch sampling regime will be performed on the producer's side. Once suitable values for p_{tol} and C have been ascertained, the representation of the universe of contaminated batches can be partitioned into batch acceptability and batch rejectability regions by means of a *limiting quality contour* (Figure 1). The limiting quality contour is comprised by the true distributions Gamma (k', m'), whose (1-p_{tol})% of the food units produced in a batch have true microbial concentrations equal or lower than C. The paired values (m', k') defining the true within-batch gamma distributions of limiting quality for p_{tol}=0.1 and C=80 CFU/cm^2 are shown in Figure 1 as a limiting quality contour segregating acceptable batches from defective batches.

2.3 Derivation of decision landscape curves

Since the plotted parameters m and k from the universe of contaminated batches correspond to the *true* within-batch distributions of microbial concentration (Figure 1), the (nontrivial) problem consists of finding a decision criterion (the arithmetic mean of the individual analytical results m_L and sample size n) that satisfies a pre-defined minimum confidence (1-α) measured on the *samples' mean* distributions. A neat way of approaching this problem is by obtaining a decision landscape plot of the values of the producer's and consumer's risks α, β for different microbiological limits m_L at a constant sample size n. Since the decision landscape should represent collectively the misclassification errors of the universe of contaminated batches, it is constructed by Monte Carlo simulation so that the between-batch variability could be propagated to the α and β risks.

Conditional-probability and ***joint-probability*** decision landscapes (Figure 2) were constructed for the assessment and determination of the within-batch sampling regime; nevertheless, they are constructed for one sample size n. The producer's and consumer's risks α and β are conditional probabilities as by definition the probabilities of batch rejection or acceptance assume previous circumstances, these are, that the batch is in fact acceptable or defective, respectively. The conditional-probability decision landscape plots these conditional probabilities at various microbiological limits for a fixed sample

size. The joint producer's and consumer's risk *α'* and *β'* are still misclassification probabilities yet multiplied, respectively, by the marginal probabilities of having an acceptable batch and a rejectable batch from the universe of contaminated batches. Thus, *α'* can be seen as the probability of rejecting an acceptable batch in the long run, and likewise, *β'* the probability of accepting a defective batch in the long run. The joint-probability decision landscape similarly plots *α'* and *β'* against microbiological limits so that a trade-off of the risks in the long run can be assessed. Notice however that these decision landscapes presume that all possible batches will only be generated within the domain of the universe of contaminated batches (Figure 1).

3. Results and Discussion

As an illustration, the conditional- and joint-probability decision landscapes (Figure 2) were obtained using the tolerance criterion of C=80 CFU/cm^2 and p_{tol}=0.10 (Figure 1) for a sample size of 5 sheep carcasses. As expected, as the microbiological limit increases – and more number of batches will pass the criterion – it is progressively more unlikely that an acceptable batch is rejected (α decreases) while it becomes more likely that a defective batch is accepted (β increases). As α and β risks must be iteratively estimated from the universe of contaminated batches at a given m_L that subdivides the batch space into acceptable and rejectable, α and β risks are not single point estimates at every m_L (as in classical variables sampling plan theory), but instead, accounting for between-batch variability, they take the form of uncertainty distributions. For this reason, the decision landscape contains mean α and β risks (curves A and C in Figure 2) and confidence intervals. For this study, the 90% CI of α and β were used and decision landscapes are shown with the 5th and 95th percentiles for β and α (lines B and D), respectively. For instance, if an m_L of 20 CFU/cm^2 was established as the maximum average of 5 samples, the conditional α risk could be with a 90% confidence anywhere between 0-0.19, with a low expected value of 0.026, while the conditional β risk will have a high expected value of 0.12 with a 90% CI of 0-0.36. However, an m_L can be derived so as to minimise the expected values of both risks. This $\mathbf{m_L}^*$ is found at the intersection of the mean α and β curves (corresponding to the point 'e' in Figure 2 where m_L=13 CFU/cm^2 and EV(α)=EV(β)= 0.055). However, it should be born in mind that the conditional risks at this m_L=13 still can be as high as 0.45 (95th pct) for α, and as high as 0.20 (5th pct) for β. To further diminish these confidence intervals, decision landscapes for higher samples sizes should be derived and be comparatively assessed.

As discussed before, there are cases where the safety of the production system is under control or characterised by sustainable very low microbial concentration levels. In such cases (as is the present one), an m_L can be found at a desirable maximum α. Suppose that a producer establishes that a reasonable quality batch is that where at least 90% of the units are below 80 CFU/cm^2 (C=80, ptol=0.10), and furthermore wishes to derive a sampling plan for GMP purposes that ensures that these reasonable quality batches will be accepted most of the times, with a minimum confidence of 90% (*1-α*). In the decision landscape, the microbiological limit is read at the point where the upper percentile curve of α (curve B) equals α=0.10. This obeys to the point $f_{\alpha=0.10}$, with a resulting m_L=25 CFU/cm^2 (Figure 2). The decision landscape additionally suggests that a lower minimum confidence (established by the producer) results in a more conservative m_L (i.e., compare points $f_{\alpha=0.10}$ and $f_{\alpha=0.05}$ in Figure 2). In terms of the consumer's risk, at the m_L of 25 CFU/cm^2, the *conditional* β risk is high, with an expected value of 0.20 and a 95th pct of 0.46. However, in practice, the joint decision landscape (Figure 2, bottom) suggests that the *joint* β' risk has a very low expected value of 0.0059, and a 95th pct of 0.015. This large difference occurs because the actual probability of having a rejectable batch (from the modelled universe of contaminated batches as classified by the tolerance criterion) is only 0.033. In fact, notice that the upper percentile of the joint β' risk asymptotically tends to this value (Figure 2, bottom). Since the probability of having a defective batch is only 3.3%, the

(joint) probability of misclassifying a defective batch as acceptable after sampling cannot be higher than 3.3%. Finally, the additional information provided by the decision landscape is related to the trade-off between risks. The shade area of Figure 2 indicates that at m_L values between 2 and 40, both misclassification errors are present and hence a certain trade-off between both should be ascertained. By examining the degree of overlap of the percentile curves, the effect of the sample size on the reduction of the misclassification errors can be assessed.

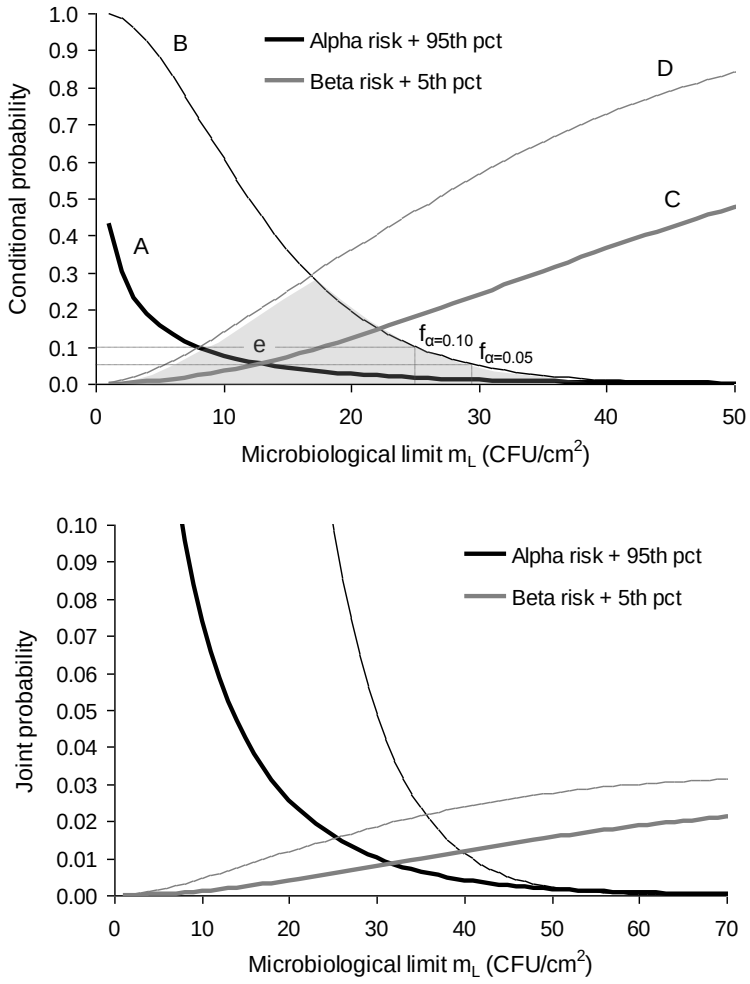


Figure 2: Decision landscapes for the conditional probabilities (top) and joint probabilities (bottom) of the producer's and consumer's risks for n=5 individual samples, derived under the tolerance criterion of C=80 CFU/cm^2 and p_{tol}=0.10

For the Irish conditions, in this work a tolerance criterion of C=100 CFU/cm^2 and ptol=0.025 is suggested, meaning that a batch of sheep carcasses should be hygienically acceptable if up to a maximum of 2.5% of carcasses produced in that batch exceeds 2 log CFU of *Enterobacteriaceae*. Under this criterion, the probability that a batch taken from the universe of contaminated batches falls

within the acceptability region is 0.9475, while the probability that a batch falls within the rejectable region is 0.0525. A series of Monte Carlo simulations were performed to produce decision landscapes for different sample sizes, and descriptors were extracted from the conditional-probability decision landscapes and compiled in Table 2.

n	m_L^* ($\overline{\alpha} = \overline{\beta}$) (CFU/cm^2)	$m_L\|(\alpha_{95pct}=0.10)$ (CFU/cm^2)	α (Mean + 90% CI)	$\beta\|m_L$ (Mean + 90% CI)
5	9.0	20.0	0.0160 [0 – 0.10]	0.246 [0 – 0.615]
6	9.8	19.5	0.0166 [0 – 0.10]	0.217 [0 – 0.595]
8	11.0	18.5	0.0172 [0 – 0.10]	0.172 [0 – 0.537]
10	11.4	17.5	0.0188 [0 – 0.10]	0.134 [0 – 0.496]
12	12.0	17.0	0.0192 [0 – 0.10]	0.115 [0 – 0.461]
14	12.5	16.8	0.0198 [0 – 0.10]	0.101 [0 – 0.447]
16	13.0	16.5	0.0212 [0 – 0.10]	0.087 [0 – 0.413]
18	13.4	16.2	0.0223 [0 – 0.10]	0.077 [0 – 0.381]
20	13.8	15.9	0.0231 [0 – 0.10]	0.066 [0 – 0.366]

Table 1: Descriptors extracted from the conditional-probability decision landscapes produced at different sample sizes (n) for the tolerance criterion of C=100 CFU/cm^2 and p_{tol}=0.025

The values of m_L^* in Table 1 represent the points at which the expected values of the producer's and consumer's risks are minimised (point 'e' in Figure 2). Under this condition, the higher the sample size, the higher the microbiological limit, and the lower the α and β risks (results not shown). Nevertheless, as this within-batch sampling regime will be derived on the producer's side (α risk), a further assumption will be presumed. Having the producer the confidence that the hygiene of the production system is under control, he may establish a sampling plan for GMP purposes (or statistical process control) ensuring that reasonable quality batches should be accepted with a minimum confidence of 90% (1-α). Thus, the third column of Table 1 represents the microbiological limit taken from the conditional-probability decision landscape at which the 95pct of α is 0.10 (point 'f' in Figure 2). The fourth column of this table shows the mean α risk at this microbiological limit, and the fifth column the mean β risk with confidence intervals read off from the decision landscape at the same microbiological limit. Results showed that for a fixed α_{95pct} of 0.10, the higher the sample size, the lower the microbiological limit. Under this condition, as the sample size increases, the mean producer's risk also increases while the consumer's risk decreases. If we wished to establish that the consumer's risk should on average be 0.10, the recommended within-batch testing regime should then be set at n=14 units and m_L=16.8 CFU/cm^2 *Enterobacteriaceae*.

4. Conclusions

Assessing the decision landscape plots, it was found that for a tolerance of a maximum of 2.5% (p_{tol}=0.025) of pre-chill sheep carcasses exceeding a critical *Enterobacteriaceae* concentration of 2 log CFU/cm^2 (C=100 CFU/cm^2), a sampling plan of n=14 carcasses and m_L=17 CFU/g offers the producer at least a confidence of 90% (1-α) of accepting reasonable quality batches. The proposed methodology based on a correlated random-effects Poisson-gamma model was proven to be statistically sound; it is the first to address the derivation of a sampling plan as a classification problem capable of propagating the batch-to-batch variability; it is adequate for microbial data consisting of many zero counts, and

uses the more effective arithmetic means (as opposed to mean logs) for expressing microbiological limits.

References

Dahms, S. (2004). Microbiological sampling plans – statistical aspects. *Mitt. Lebensm. Hyg*, 95, 32-44.

Duncan, A. J. (1986). Quality control and industrial statistics, 5th ed., Irwin IL: Homewood.

Gonzales-Barron, U., & Butler, F. (2011a). A comparison between the discrete Poisson-gamma and Poisson-Lognormal distributions to characterise microbial counts in foods. *Food Control*, 22, 1279-1286.

Gonzales-Barron, U., & Butler, F. (2011b). Characterisation of within-batch and between-batch variability in microbial counts in foods using Poisson-gamma and Poisson-lognormal regression models. *Food Control*, 22, 1265-1278.

Gonzales-Barron, U., Lenahan, M., Sheridan, J., & Butler, F. (2012). Use of a Poisson-gamma model to assess the performance of the EC process hygiene criterion for Enterobacteriaceae on Irish sheep carcasses. *Food Control*, 25, 172-183.

Construction d'un modèle dose réponse pour *Clostridium perfringens* par inférence bayésienne

Bayesian modelling of *Clostridium perfringens* dose response

S. Jaloustre[123*] and M.L. Delignette-Muller[34]

[1] *Agence Nationale de Sécurité Sanitaire (Anses), LSA, 23 av. du Gal de Gaulle, F-94706, Maisons-Alfort Cedex, France.*
E-mail :*Severine.SEVRIN-JALOUSTRE@anses.fr*
[2] *AgroSup Dijon, F- 21079 Dijon, France*
[3] *Université de Lyon, F-69000, Lyon ; Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France*
[4] *Université de Lyon, F-69000, Lyon ; VetAgro Sup Campus Vétérinaire de Lyon, F-69280 Marcy l'Etoile, France*
E-mail :*ml.delignette@vetagro-sup.fr*

Résumé
Le but de cette étude était d'estimer les paramètres d'un modèle dose réponse pour prédire le risque de diarrhée liée à l'exposition à *Clostridium perfringens* dans une population donnée. Pour explorer de potentielles sources de variabilité, deux sources de données ont été utilisées, permettant de construire deux modèles dose réponse '*single hit*' sur dose individuelle : un modèle dit modèle d'expérience à partir de données issues d'expériences sur volontaires sains, et un modèle dit modèle TIAC (Toxi Infection Alimentaire Collective) à partir de données issues de TIAC. L'incertitude sur les doses individuelles a été décrite à partir d'informations publiées sur la dispersion de la contamination et sur la variabilité de la taille des portions d'aliments. Dans chaque modèle, une variabilité inter-exposition, c'est-à-dire inter-expérience ou inter-TIAC selon le modèle, a été décrite. Les paramètres de chaque modèle ont été estimés par inférence bayésienne. La source des données apparaît avoir un impact sur l'estimation de ces paramètres et sur la prédiction du risque, avec une sous-estimation du risque si on utilise le modèle construit à partir de données d'expériences.
Mots-clés : *Clostridium perfringens*, modèle dose réponse, inférence bayésienne

Keywords : *Clostridium perfringens*, dose response, Bayesian modelling

1. Introduction

Clostridium perfringens food poisoning is a really common food borne illness particularly in institutions and restaurants (Crouch and Golden, 2005). Vegetative cells provoke often abdominal cramps and diarrhea, sometimes fever and headache and seldom vomiting. Symptoms begin from 6 to 30 hours after the ingestion of contaminated food and last for one day or less. They are usually mild but vulnerable people may die of the disease (Mead *et al.*, 1999). Outbreaks are often associated with processed meat (WHO, 2003) as involved processes may allow *Clostridium perfringens* germination and growth sometimes without

final thermal inactivation of vegetative cells (Fazil *et al.*, 2002). The aim of this study was to estimate *Clostridium perfringens* dose response model parameters for risk assessment.
One dose response model has been published to date by Golden *et al.* (2009). The authors used data collected from published human volunteer feeding studies with mean doses to fit an exponential model using maximum likelihood methods and described a between-strain variability on the *Clostridium perfringens* virulence.
Many factors related to microorganism, host and food matrix may affect the frequency and the severity of adverse effects and are likely to induce a variability on response, which has to be described in order to capture all the possible outcomes (ILSI, 2000). Assessing such variability requires the use of datasets that capture the diversity of human population, pathogen strains and food matrices (FAO, 2003). Therefore it may be interesting to use various sources of data, provided that a data selection based on objective arguments leads to a consistent dataset with the same definitions of the dose and the modeled adverse effect (FAO, 2003).
Human volunteer feeding studies provide interesting data as ingested doses, numbers of exposed and ill people are precisely known. It is thus of interest to use them, as done by Golden *et al.* (2009). But such studies may not represent variability on host susceptibility (Teunis *et al.*, 2010). On the opposite, outbreak data are more likely to represent variability on host susceptibility. Dose response models fitted on volunteer feeding study data may thus underestimate the risk of food borne illness in the whole population including susceptible groups. That is why in the present study, some outbreak data were added to data from human volunteer feeding studies and two dose response models were fitted on volunteer feeding study data for the first and on outbreak data for the second.
For both data sources, uncertainty on actual ingested dose was taken into account. Using collected data and information from a study on French dietary habits, actual doses were simulated, leading us to fit a 'single hit' model on actual doses (Haas *et al.*, 1999).
According to international organization recommendations, potential variability related to various factors (host, pathogen and food matrix,) was explored but could not be modeled due to lacking data. As variability between outbreak or experiment (for simplicity named exposure) could not be explained by one of those factors, a between-exposure variability was described in each model. To estimate the model parameters, a Bayesian approach was used making it easy to describe separately variability and uncertainty. Risk predictions by both models were then performed and compared to those by the model published by Golden *et al.* (2009).

2. Materials and methods

2.1 Collected data

The aim of the present study was to estimate *Clostridium perfringens* dose response model parameter to predict food borne diarrhea due to *Clostridium perfringens* vegetative cells. In the present study, modeled food borne symptom was defined as diarrhea within 6 to 24 hours following exposure to *Clostridium perfringens* vegetative cells. Therefore we selected published and personal data for which at least information on the ingested dose and the number of exposed and ill people were available. For each collected data, available information about strain, food product and host susceptibility according to FAO classification

(FAO, 2004) was collected. Two sources of data were taken into account : human volunteer feeding studies and outbreaks.

2.1.1 Human volunteer feeding studies

Many studies (Dack *et al.*,1954; Dische and Elek, 1957; Hauschild and Thatcher, 1967; Strong *et al.*, 1971; Skjelkvale and Uemura, 1977a, 1977b) have been published. In the selected studies (Dack *et al.*,1954; Dische and Elek, 1957; Hauschild and Thatcher, 1967; Strong *et al.*, 1971), healthy volunteers were fed with *Clostridium perfringens* vegetative cells in various products at relatively high doses ($>10^8$ ufc). In these studies, data obtained with strains known to lack the cpe gene (Strong *et al.*, 1971) were not collected.

Information on doses was reported by authors either as a single mean dose (Dack *et al.*,1954; Dische and Elek, 1957; Strong *et al.*, 1971) or a mean dose with a dose range (Dische and Elek, 1957; Hauschild and Thatcher, 1967). Observed dose dispersion was in these studies only due to the dispersion of *C. perfringens* concentration in food. When information on dose range was available for an experiment, a relative standard deviation *RSD* of this concentration was estimated by dividing the estimated standard deviation by the mean concentration. A mean relative standard deviation was then estimated and used to model uncertainty intervals of *C. perfringens* concentration when only a single mean dose had been reported by authors.

2.1.2 Outbreaks

Of all the published outbreaks, we only selected sufficiently described outbreaks. . Thus only 6 published outbreak data were used from 2 publications (Hobbs *et al.*, 1953; Sutton and Hobbs, 1968). Of the nine outbreaks reported by the central laboratory of veterinary services (CLVS), only 4 outbreaks reported by CLSV were selected as for some of them, the number of exposed people was considered too low and uncertain or *Clostridium perfringens* was not the only identified pathogen.

For these outbreaks, strains were not identified and information on host susceptibility was sometimes lacking. Information on ingested dose was not directly reported and the only available information was about *Clostridium perfringens* concentration in contaminated food. In this case, uncertainty on actual dose may derive from variability on serving size in the exposed population and from uncertainty on *C. perfringens* concentration in food. As no information was available on the dispersion of contamination in food, *Clostridium perfringens* concentration in contaminated food was described using the mean relative standard deviation estimated from study data. Variability on serving size was modeled using data obtained in a study on French dietary habits (Dubuisson *et al.*, 2010; Lioret *et al.*, 2010).

The whole dataset (volunteer feeding studies and outbreak data) is represented in Fig.1. Uncertainty on actual ingested dose was estimated using the estimated mean relative standard deviation for both source of data and variability on serving size for outbreaks. Estimated uncertainty derived from sampling.

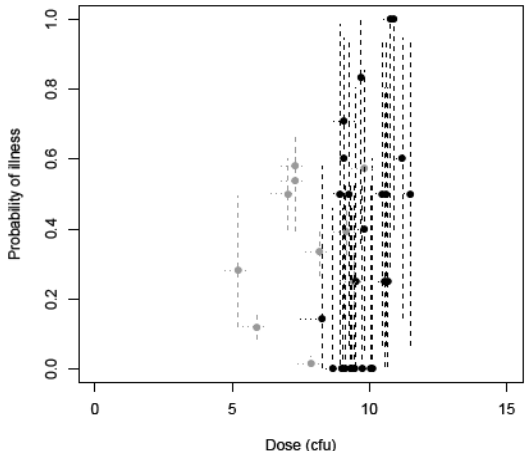


Figure 1 : Collected data. Grey levels represent the source of data (black for volunteer feeding studies, grey for outbreaks). Points represent observed data. Vertical dashed lines represent the sampling uncertainty. Horizontal dotted lines represented the uncertainty on ingested dose.

2.2 Dose response model

Infection and illness result from the ingestion of one or more *Clostridium perfringens* vegetative cells able to resist all barriers to reach intestine (infection) and to provoke diarrhea (illness). In the present study, we directly modeled probability of illness as a function of the ingested dose using a 'single hit' model on actual doses. First assumption in this model was that a single cell was capable of provoking illness. This implied that even for very low doses there was always a very small, a non-zero probability of illness, this probability increasing with the dose. Second assumption was that the mean probability for cells to provoke illness did not depend on the size of the dose (FAO, 2003). In this 'single hit' model, the probability of illness P_{ill} is expressible as :

$$P_{ill}(d,r) = 1-(1-r)^{d} \qquad Eq.1$$

with r the probability that any cell resists barriers and provokes illness and d the actual dose. r corresponds to the *Clostridium perfringens* virulence defined by Golden *et al.* (2009).

2.3 Bayesian modelling

2.3.1 Proposed models

A Bayesian approach was chosen for estimating the model parameter r. Factors related to the pathogen, the host and the food matrix were identified as potential sources of variability on response by international committees (ILSI, 2000; FAO, 2003). In our dataset, information on those factors, particularly concerning outbreaks, was too sparse to make it possible to explore a between-factor variability on r, so that only a between-experiment/outbreak variability (for simplicity named between-exposure variability) was explored. This factor was considered as a random factor as the levels under study were a random sample of their population of interest.
At last, as volunteer feeding studies (for simplicity named experiments) are usually performed on healthy people with non wild strains (Teunis *et al.*, 2010), response in these studies at a certain dose may be underestimated compared to observed response in outbreaks at the same

dose. This trend clearly appears in Fig.1. That is why we fitted two dose response models : the first on experiment data and the second on outbreak data (respectively named experiment and outbreak models).

In the proposed models, to describe variability on r, the $\log it(r) = \ln\left(\frac{r}{1-r}\right)$ was used. The random effect related to 'Exposure' was described by normal distributions : $N(\mu_{\log itr\_\exp}, \sigma_{\log itr\_\exp})$ for experiments and $N(\mu_{\log itr\_out}, \sigma_{\log itr\_out})$ for outbreaks.

2.3.2 Prior distributions

Prior distributions of $\mu_{\log itr\_out}$ and $\mu_{\log itr\_\exp}$ were defined from $\ln r$ values reported by Golden *et al.* (2009). Non informative distributions were defined for the two standard deviations $\sigma_{\log itr\_out}$ and $\sigma_{\log itr\_\exp}$ (Gelman, 2006).

2.3.3 Computations

The empirical posterior distribution of each parameter was computed from its prior one and from the corresponding data.Computations were performed using the JAGS software (Plummer, 2009) and the rjags package of R software (R Development Core Team, 2009). For each model, inferences were made on 5.10^4 iterations for each of 3 independent MCMC chains after an adaptation phase of 5.10^3 iterations. A thinning interval of 10 was used and 5000 values were thus kept for each chain. Convergence was checked by visually analyzing MCMC chain traces and examining Gelman and Rubin convergence statistics, as modified by Brooks and Gelman (1998).

2.3.4 Assessment of the goodness-of-fit

In order to check the ability of the proposed models to describe the observed data, the parameter values sampled from the four MCMC were used to simulate, for each observed data, $1.5.10^3$ values of the probability of illness. Median simulated values of the probability of illness and their 95% credibility intervals, defined from the 2.5^{th} and the 97.5^{th} percentiles, were compared to the observed values. Median simulated values were expected to be close to the observed ones while the simulated 95% credibility intervals were expected to be reasonably tight.

2.3 Model prediction

For simulations, three doses ranging from 5 to 9 $\log_{10}$cfu were fixed. In order to separate variability on r and uncertainty on parameters, second order Monte Carlo simulations were performed as follows for each fixed dose :

1. A set of parameters considered as uncertain ($\mu_{\log itr\_\exp}$ and $\sigma_{\log itr\_\exp}$ or $\mu_{\log itr\_out}$ and $\sigma_{\log itr\_out}$) were randomly selected in their joint posterior distributions for the proposed model.
2. Given this set of parameters, 1001 r values were randomly selected from their variability distributions, making it possible to predict 1001 probabilities of illness using Eq.1. The 1001 predicted probabilities lead to a variability distribution, summarized by its mean.

3. Steps 1 and 2 were performed 601-fold in order to obtain an uncertainty distribution of the mean probability of illness.

3 Results

3.1 Dose response model parameters

Statistics of MCMC replicates obtained for the proposed model and prior distributions are presented in Table 1. As shown in this table, MCMC replicates are much narrower than prior distributions. The effect of the source of data appears clearly in the non overlapping $\mu_{\text{logit}r\_out}$ and $\mu_{\text{logit}r\_\exp}$ 95% credibility intervals but not in the $\sigma_{\text{logit}r\_out}$ and $\sigma_{\text{logit}r\_\exp}$ 95% credibility intervals. The lower values of $\mu_{\text{logit}r\_out}$ are consistent with observations on collected data.

Goodness-of-fit graph of the proposed models is presented on Fig. 2. As observed in this figure, median predictions are often close to observed values. It seems that the model particularly well describes outbreak data as median predictions are close to observed values while 95% credibility intervals are really narrow. Human volunteer feeding study data are worse described by the model with broader 95% credibility intervals and median predictions often underestimating the observed proportion of illness. Collected data and predictions by the proposed model are presented in Fig. 3. As observed in this figure, the proposed model correctly describes the observed data. The effect of the source of data appears even if the two 95% credibility intervals overlap due to the huge between-exposure variability, summarized by $\sigma_{\text{logit}r\_out}$ and $\sigma_{\text{logit}r\_\exp}$.

Parameter σ	Prior distribution	MCMC replicates
$\mu_{\text{logit}r\_out}$	-25 [-15,-35]	-19.38 [17.45,-21.53]
$\mu_{\text{logit}r\_\exp}$	-25 [-15,-35]	-24.40 [-23.40,-25.61]
$\sigma_{\text{logit}r\_out}$	5 [0.025,0.975]	2.98 [1.91,4.68]
$\sigma_{\text{logit}r\_\exp}$	5 [0.025,0.975]	2.25 [1.49,3.67]

Table 1: Statistics of prior distributions and MCMC replicates for the proposed model. The first value represents the median value while values in brackets represent 95% credibility intervals.

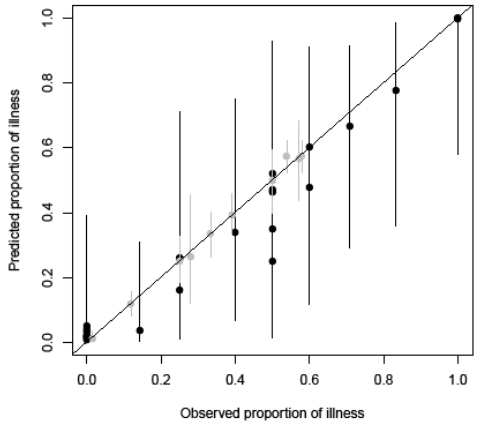


Figure 2 : Comparison of observed proportion of illness with probability of illness predicted by the proposed models. Grey levels represent the source of data (black for volunteer feeding studies, grey for outbreaks). Points represent observed data and median predictions. Vertical lines represent the 95% credibility intervals of predicted probability of illness.

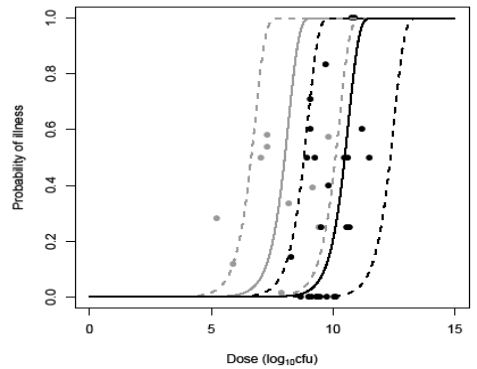


Figure 3 : Collected data and predictions by the proposed models. Grey levels represent the source of data (black for volunteer feeding studies, grey for outbreaks). Points represent observed data. Solid curves represent the median predictions by the proposed models. Dashed curves represent the 95% credibility intervals of prediction by the proposed models.

3.2 Model prediction

Predicted probabilities of illness for each fixed dose are presented in Table 2. Prediction intervals by the ‘experiment model’ are consistent with deterministic predictions by the model published by Golden *et al.* (2009). These two predictions are far lower than predictions by th outbreak model. For the same dose, the probability of illness predicted by the ‘experiment

model' is often one hundredth the one predicted by the 'outbreak model'. Thus the 'experiment model' really appears to underestimate the risk.

Dose (log_{10}cfu)	Prediction by the model		
	'Outbreak model'	'Experiment model'	Model published by Golden *et al.* (2009)
5	$5.2.10^{-2}[7.8.10^{-3}, 1.9.10^{-1}]$	$3.7.10^{-5}[6.8.10^{-6}, 1.5.10^{-3}]$	$2.0.10^{-5}$
7	$2.9.10^{-1}[1.5.10^{-1}, 5.1.10^{-1}]$	$3.3.10^{-3}[6.8.10^{-4}, 2.8.10^{-2}]$	$2.4.10^{-3}$
9	$7.3.10^{-1}[4.8.10^{-1}, 8.8.10^{-1}]$	$1.2.10^{-1}[5.2.10^{-2}, 2.5.10^{-1}]$	$9.6.10^{-2}$

Table 1: Mean probabilities of illness predicted by the 'outbreak' and 'experiment models' and by the model published by Golden *et al.* (2009). The first value represent the median value predicted by the proposed models while values in brackets represent predicted 95% credibility intervals.

3 Discussion and conclusion

In the present study, the parameters of a dose response model using published data obtained from volunteer feeding studies and outbreaks and personal outbreak data, in order to describe as well as possible variability on response related to pathogen, host and food matrix. Two single hit models on actual doses were used to fit the collected data. After checking its ability to predict collected data, the model fitted on outbreak data and describing a between-outbreak variability was used for prediction in risk assessment.

Concerning outbreak data, information on dose was reported by authors or CLVS as *C. perfringens* concentration in food so that we used data from a French dietary study to model portion size. These data were directly used to describe consumption of contaminated food in other countries and sometimes forty years ago. Insofar published outbreaks occurred in developed and often European countries, recent standardization of dietary habits in European countries (van der Wilka and Jansenb, 2005) makes it possible to assume that portion size do not differ from an European country to another. Concerning evolution of portion size since 1950 (time of first published outbreaks), it seems that portion size did not evolve (Dubuisson *et al.*, 2010; Lioret *et al.*, 2010), so that current data on portion size may be used to describe consumption of contaminated food 50 years ago, as done in this study.

If numbers of exposed and ill people in human volunteer feeding studies are precisely known, it is often the opposite for outbreaks. It is very difficult to know precisely the number of exposed people during outbreaks except in certain cases of outbreaks in institutions. Moreover, symptoms related to *Clostridium perfringens* are usually so mild that some ill people may be undetected, leading to an underestimation of the number of ill people. Thus uncertainty on numbers of exposed and ill people during outbreaks may be huge. To lower it, a selection of published and reported outbreaks was performed such as to remove outbreaks which had been insufficiently described or for which strong doubts on numbers of exposed and ill people appeared. Even after this selection, uncertainty on both numbers still remains but it was not possible to model it.

According to the predictions by the 'outbreak' and 'experiment models', it should be avoided to fit a dose response model on human volunteer feeding study data, which lead to an underestimation of the predicted proportion of illness. It would be thus necessary to enhance the outbreak detection and the quality of their reports to use more outbreak data.

References

Brooks, S.P. and Gelman, A., 1998. General Methods for Monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 434-455.

Crouch, E. and Golden, N., 2005. A Risk Assessment for *Clostridium perfringens* in Ready-To-Eat and Partially Cooked Meat and Poultry Products, USDA, Food Safety Inspection Service, September 2005.

Dack, G.M., Sugiyama, H., Owens, F.J., and Kirsner, J.B., 1954. Failure to produce illness in human volunteers fed *Bacillus cereus* and *Clostridium perfringens*. *J. Infect. Dis*. 94: 34-38.

Dische, F.E. and Elek, S.D., 1957. Experimental food poisoning by *Clostridium welchii*. *Lancet* 273, 71-74.

Dubuisson, C., Lioret, S., Touvier, M., Dufour, A., Calamassi-Tran, G., Volatier, J.L. and Lafay, L., 2010. Trends in food and nutritional intakes of French adults from 1999 to 2007: results from the INCA surveys. *Br J Nutr*. 103, 1035-1048.

Fazil, A.M., Ross, T., Paoli, G., Vanderlinde, P., Desmarchelier, P. and Lammerding, A.M., 2002. A probabilistic analysis of Clostridium perfringens growth during food service operations. International Journal of Food Microbiology 73, 315-329.

FAO/WHO, 2003. Hazard Characterization for Pathogens in Food and Water. Guidelines. FAO Food and Nutrition. FAO/WHO, Roma. Available at: http://www.fao.org/docrep/006/y4666e/y4666e00.htm.

FAO, 2004. Risk assessment of *Listeria monocytogenes* in ready-to-eat foods. WHO/FAO 2004. Available at: http://www.fao.org/ag/agn/agns/jemra_riskassessment_listeria_en.asp.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1, 515–533.

Golden, N.J., Crouch, E.A., Latimer, H., Kadry, A.R. and Kause, J., 2009. Risk assessment for *Clostridium perfringens* in ready-to-eat and partially cooked meat and poultry products. *Journal of Food Protection* 72, 1376-1384.

Haas, C.N., Rose, J.B., Gerba, C.P., 1999. *Quantitative microbial risk assessment*. John Wiley and sons, USA, New York : 449 pp.

Hauschild, A., and Thatcher, F., 1967. Experimental food poisoning with heat-susceptible *Clostridium perfringens* type A. *J. Food Sci*. 32, 467-471.

Hobbs, B.C., Smith, M., Oakley, C., Warrack, G., and Cruickshank, J., 1953. *Clostridium welchii* food poisoning. *J. Hyg*. 51, 75-101.

ILSI [International Life Sciences Institute], 2000. Revised framework for microbial risk assessment. Washington, D.C: ILSI Risk Science Institute Press.

Lioret, S., Dubuisson, C., Dufour, A., Touvier, M., Calamassi-Tran, G., Maire, B., Volatier, J.L. and Lafay, L., 2010. Trends in food intake in French children from 1999 to 2007: results from the INCA (étude Individuelle Nationale des Consommations Alimentaires) dietary surveys. *Br J Nutr*. 103, 585-601.

Mead, P.S., Slutsker, L. and Dietz, V., 1999. Food-related illness and death in the United States. Emerging Infection Diseases 5, 607-625.

Plummer, M., 2009. JAGS Version 1.0.9 Manual. Lyon: International Agency for Research on Cancer. Available at http://www-ice.iarc.fr/~martyn/software/jags/.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 3-900051-07-0, http://www.R-project.org.

Skjelkvale, R., and Uemura, T., 1977a. Experimental diarrhoea in human volunteers following oral administration of *Clostridium perfringens* enterotoxin. *J. Appl. Bacteriol.* 43, 281-286.

Skjelkvale, R., and Uemura, T., 1977b. Detection of enterotoxin in feces and anti-enterotoxin in serum after *Clostridium perfringens* food-poisoning. *J. Appl. Bacteriol*. 42, 355-363.

Strong, D., Duncan, C., and Perna, G., 1971. *Clostridium perfringens* type A food poisoning II. Response of the rabbit ileum as an indication of enteropathogenicity of strains of *Clostridium perfringens* in human beings. *Infect. Immun.* 3, 171-178.

Sutton, R.G.A. and Hobbs, B.C., 1968. Food poisoning caused by heat sensitive *Clostridium welchii*. A report of five recent outbreaks. *Journal of Hygiene* 66, 135-146.

Teunis, P.F., Kasuga, F., Fazil, A., Ogden, I.D., Rotariu, O. and Strachan, N.J., 2010. Dose-response modeling of *Salmonella* using outbreak data. *Int J Food Microbiol*. 144, 243-249.

van der Wilka, E.A. and Jansenb, J., 2005. Lifestyle-related risks: are trends in Europe converging? *Public Health* 119, 55-66.

WHO, 2003. Surveillance Programme for Control of Foodborne Infections and Intoxications in Europe 8th report 1999-2000. Available at http://www.bfr.bund.de/internet/8threport/8threp_fr.htm.

Session 3 : Sensométrie I / *Sensometrics I*

Thurstonian and Statistical models

Rune Haubo Bojesen Christensen

DTU Informatics, Section for Statistics
Technical University of Denmark, Build. 305, Room 122,
DK-2800 Kgs. Lyngby, Denmark

Thurstonian models for sensory discrimination tests link the observed answers to an underlying measure of sensory difference known as the Thurstonian delta. These models provide a more detailed understanding of the discrimination task, for instance, explaining why the expected proportion of correct answers is different for the triangle and 3-AFC protocols with the same product differences. It turns out that several Thurstonian models for sensory discrimination tests can be identified as particular versions of well-known statistical model classes. For instance, the standard binomial discrimination protocols such as duo-trio, triangle and m-AFC can all be identified as instances of so-called generalized linear models. Similarly the Thurstonian models for the A-not A with sureness and 2-AC protocols can be identified as so-called cumulative link models. This identification makes it possible to combine probabilistic inference with regression and ANOVA techniques for more insightful analyses, more powerful significance tests, reduced bias in parameter estimates and more accurate quantification of the statistical uncertainty. So-called random effects versions of these models can help us overcome one of the greatest challenges in sensory discrimination testing, namely the issue of replications. These models also combine Thurstonian inference with regression tools, facilitate subject-specific inference and produce high powered tests of product differences.

Les modèles CUB pour l'analyse sensorielle dans l'industrie agro-alimentaire

CUB models for sensory analysis in food industry

Marica Manisera[1], Domenico Piccolo[2], Paola Zuccolotto[1]

[1] *Department of Quantitative Methods, University of Brescia*
C.da S. Chiara, 50 - 25122 Brescia, Italy
E-mail: {*manisera, zuk*}*@eco.unibs.it*

[2] *Department TEOMESUS, Statistical Sciences Unit, University of Naples Federico II*
Via Leopoldo Rodinó, 22 - 80138 Napoli, Italy
E-mail: *domenico.piccolo@unina.it*

Résumé
Les préférences et les perceptions des attributs sensoriels des produits sont très importants pour les fabricants de l'industrie agro-alimentaire, afin d'éviter la déception du marché et d'améliorer la qualité des aliments. En effet, les analyses sensorielles combinées avec les méthodes statistiques appropriées permettent de segmenter le marché, d'obtenir le positionnement des produits (marques, organisations,...) et d'identifier le niveau d'acceptabilité du marché. Enfin, cela a un grand impact sur la qualité des aliments et la compétitivité industrielle. Dans cet article, nous utilisons les modèles CUB pour analyser les données sensorielles provenant d'une enquête sur le café italien (espresso).
Mots-clés : analyse sensorielle, modèles CUB , café italien

Abstract
Consumers and experts' preferences and perceptions of the sensory attributes of products are very important to manufacturers in the food industry, in order to avoid market disappointment and improve food quality. Indeed, appropriate sensory analyses, combined with appropriate statistical methods allow to segment market, to obtain positioning of products (brands, organizations, ...) and to identify market acceptability. This, finally, has a great impact upon food quality and industrial competitiveness. In this paper, we use CUB models to analyse sensory data coming from a survey on the Italian espresso.
Keywords : sensory analysis, CUB models, Italian coffee

1 Introduction

Sensory evaluation is a scientific method where experimental results are collected on a set of sampled consumers who express preferences and reactions with respect to food and drink. Since samples are generally obtained according to standard statistical designs, this field attracts many approaches for a correct analysis of the results insofar as formal conditions for inferential procedures are respected.

On the other hand, consumer preferences result from complex interactions where subjective, objective and contextual factors are present with different roles. In fact, the expressed choice is the result of a human decision and we should assume that this process is a final act conditioned by personal history, environmental variables, subjective covariates and objects' characteristics, which all surely interact with the modality of the survey. As a consequence, it may be worth studying the stochastic structure of the choice process in order to adequately model the observed preferences.

Operationally, to collect sensory data, experts or untrained subjects are often asked to rate or rank different products on the basis of some sensory descriptors (items), by expressing their perceptions on hedonic response scales (usually 9-point Likert scales). For example, consumers can be asked to evaluate quality attributes and express their preferences towards colour, smell, taste and mouth feel for a collection of coffee varieties, as we will pursue in this paper.

In this way, affective tests concern ordinal measurements. Such scales are substantially of qualitative nature although some numerical coding, as the integers $\{1, 2, \ldots, m\}$, is generally proposed. Then, a correct statistical analysis must be related to ordinal data modelling and current literature focuses on the models generated by cumulative probability in order to take the ordinal nature of sensory data into account (Agresti, 2010).

In this paper, following previous research in the area promoted by Piccolo (2003), we adopt a different structure by assuming that the response of each consumer is the combination of a *feeling* attitude towards the food being evaluated and an intrinsic *uncertainty* component surrounding the discrete choice. This class of models have been successfully applied in several fields (D'Elia and Piccolo, 2005; Iannario, 2007) and sensory analysis is a favoured context (Piccolo and D'Elia, 2008; Piccolo and Iannario, 2010). In fact, these models allow to measure how the perception process is transformed into personal evaluations which are a mixture of several components: the relevant ones are defined as feeling and uncertainty. Moreover, we will show that the added value of the proposal is mainly related to a sharp visualization of a huge amount of information by a graphical pattern of the estimated models represented in the parametric space.

This work shows how several varieties of Italian coffee (*espresso*) have been rated by a number of Italian and foreign tasters with respect to visual, olfactory and gustatory perceptions. The data set has been released without information on product and usage characteristics and the whole analysis will be concerned with the ability of the proposed models to cope with information derived by the frequency distributions of expressed preferences. The paper is organized as follows: in Section 2, we discuss the fundamentals of CUB modelling approach and in Section 3 we present the case study. Some final remarks conclude the paper (Section 4).

2 CUB models

As mentioned in Section 1, the observed preferences result from the consumers' evaluation of food and drink, that is the expression of their preferences on a hedonic response scale. Perception and evaluation result from complex psychological mechanisms determined by many interacting factors of different nature (psychological, social, biological, physiological, etc.). Especially when eating and drinking behaviour is involved, human decision making occurs at a non-conscious level and sensory and consumer research should take psychological insights into account (Köster, 2009).

The philosophy of CUB models is perfectly in line with this, since feeling and uncertainty

represent the latent components combined together in order to express the consumers' judgements (i.e., the observed discrete choices). The feeling component is the degree of agreement with a given item and results from subjective motivations. According to the latent variable approach, it can adequately be interpreted as a continuous latent random variable that is then discretized, since the consumers' ratings assigned to an item are discrete. On the other hand, the uncertainty component is the indecision intrinsically present in human choices and resulting from factors related to the evaluation process (for example, the limited knowledge of the problem, the nature of the chosen questionnaire and response scale, the subjective interest towards items). Both components are explicitely considered in the CUB models, by means of a mixture of two random variables, as explained in Subsection 2.1.

2.1 Basic issues and extended CUB models

By definition, any model is strictly arbitrary; thus, the rationale for their structure comes from a blend of logical arguments and empirical facts. Overall, parsimony of parameters is a key issue. In line with these arguments, the class of models we are going to introduce aims at parametrically defining the behaviour of respondents as generated by two main latent components.

Specifically, *uncertainty* may be modelled with regard to the extreme choice of a person who assigns the same probability to each category, with a complete indifference. As a consequence, for the distribution related to uncertainty we introduce the discrete Uniform random variable U defined over the support $\{1, 2, \ldots, m\}$, for a given m:

$$Pr\,(U = r) = \frac{1}{m} = U_r, \quad r = 1, 2, \ldots, m\,.$$

This random variable maximizes the entropy, among all the discrete distributions with finite support $\{1, 2, \ldots, m\}$, for a fixed m, and it is minimally informative about the choice (when one knows only the number m of categories).

Instead, we model the *feeling* component by means of a shifted Binomial random variable V whose probability distribution is:

$$Pr\,(V = r \mid \xi) = \binom{m-1}{r-1} \xi^{m-r}\,(1-\xi)^{r-1} = b_r(\xi)\,, \quad r = 1, 2, \ldots, m\,,$$

The rationale for such distribution stems from heuristic and pragmatic point of views: the (shifted) Binomial distribution is able to cope with different shapes of sample data and just with a single parameter. From a statistical point of view, combinatorial and selective arguments confirm the convenience to adopt such distribution, as argued by Iannario (2011).

If we weight the components assumed for uncertainty and feeling, we are introducing a (convex) ***C***ombination of a discrete ***U***niform and a shifted ***B***inomial distributions, and this justifies the CUB acronym. Then, a CUB random variable R expressing the final choice of the respondent is defined by the probability mass function:

$$Pr\,(R = r \mid \boldsymbol{\theta}) = \pi\, b_r(\xi) \,+\, (1-\pi)\, U_r\,, \quad r = 1, 2, \ldots, m\,,$$

where $\boldsymbol{\theta} = (\pi, \xi)'$, $\pi \in (0, 1]$ and $\xi \in [0, 1]$. The parametric space is then the (left open) unit square, $\Omega(\boldsymbol{\theta}) = \Omega(\pi, \xi) = \{(\pi, \xi) : \; 0 < \pi \leq 1; \; 0 \leq \xi \leq 1\}$. Iannario (2010) proved that CUB models are identifiable for any $m > 3$.

The class of CUB models turns out to be a very flexible parametric family since the shape of the distribution largely varies over $\Omega(\pi, \xi)$, as shown by Piccolo (2003). This allows to fit data with positive or negative skewness, any modal value and peaked or flat distributions.

Parameters are immediately related to the latent components of the responses. The *feeling parameter* (ξ) is mostly related to location measures and strongly determined by the skewness of responses: it increases when respondents prefer low ratings. Usually, high values of the responses imply high consideration towards the food; then, in sensory analysis, the quantity $(1-\xi)$ increases with sensory satisfaction with the product. Instead, the *uncertainty parameter* (π) modifies the heterogeneity of the distribution and it is mostly related to the comparisons among probabilities. Then, uncertainty of the choice increases with $(1-\pi)$.

Since there is one-to-one correspondence among a CUB random variable and the parameter vector $\boldsymbol{\theta} = (\pi, \xi)'$, we represent each CUB model as a point in the unit square. This visualization is a focal point of the approach since a single point summarizes any aspect of the probability distribution and allows for immediate comparison with respect to time, space and circumstances.

Since $1-\pi$ measures the *propensity* of respondents to behave in accordance to a completely random choice, and $1-\xi$ measures the *strength of feeling* of the subjects for a direct and positive evaluation of the food, hereafter we will consider the plot of CUB models with coordinates $1-\pi$ and $1-\xi$, respectively.

The expectation of R is given by: $\mathbb{E}(R) = \frac{(m+1)}{2} + \pi\ (m-1)\left(\frac{1}{2} - \xi\right)$. It confirms that the mean value moves towards the central value of the support depending on the sign of $(\frac{1}{2} - \xi)$ and this behaviour is related to the skewness of the distribution. In fact, a CUB random variable is symmetric if and only if $\xi = 1/2$.

A peculiar aspect of the last formula is that the expectation of R is constant for infinitely many values of the parameter vector $\boldsymbol{\theta} = (\pi,\ \xi)'$; as a consequence, we may obtain the same mean value for rating distributions which are quite different. In addition, expectation does not convey all the characteristics of a random phenomenon since these are explained by a sequence of higher moments.

CUB models have been extended in several directions as recently pointed out by Iannario and Piccolo (2011), and these generalizations concern the probability distribution of the components, the inclusion of subjects' and objects' covariates, the joint consideration of several objects/items in a multivariate context.

For example, if one considers that both uncertainty and feeling may be conditioned by subjects' characteristics, we can define CUB models with covariates by introducing a logistic link among parameters and covariates of the respondents. This extension is particulary noticeable since it allows for testing and measuring the effect of known characteristics on the responses and thus such models are especially valuable for marketing studies.

Another generalization stems from the circumstance that respondents may sometimes prefer a quick response instead to weigh up more demanding choices. This behaviour is frequent in sensory analysis and induces an anomalous value of the frequency of a given category. Since this component may imply both biases and inefficiencies in the statistical analysis, it can explicitely be modelled in CUB models with a *shelter effect* (Corduas *et al.*, 2009; Iannario, 2011).

2.2 Inferential issues

When sample data are available, the classical steps of the iterative cycle of specification, estimation and validation of a CUB model may be consistently pursued by maximum likelihood (ML) methods which lead to efficient asymptotic properties of the statistical procedures. Moreover, the involved mixture distribution advocates the EM procedure as an effective algorithm to reach convergence almost everywhere on $\Omega(\boldsymbol{\theta})$, as shown by Everitt and Hand (1981), McLachlan and Krishnan (2008), McLachlan and Peel (2000), among others.

For a general CUB model with covariates, the ML estimation has been derived by Piccolo (2006) and extended by Iannario (2011) to models with *shelter effect*; several suggestions have been given for improving the convergence of the procedure by means of accurate preliminary estimators. In this context, the significance of the estimated parameters, the relevance of the covariates and the validation of the model are obtained by exploiting the asymptotic properties of the ML estimators.

A critical review of fitting measures for ordinal models, and specifically for CUB models, is in Iannario (2009). When sample data are summarized by the observed frequencies $n_1, n_2, \ldots, n_m$, the log-likelihood of the *saturated* CUB model is

$$\ell_{sat} = -n \log(n) + \sum_{r=1}^{m} n_r \log(n_r) .$$

This quantity is easily computable on the basis of sample data and acts as a benchmark for comparing the effectiveness of more elaborate structures, and also for fitting purposes. If $\ell(\hat{\boldsymbol{\theta}})$ and ℓ_0 are the log-likelihoods of the estimated model and of a Uniform model, respectively, a convenient measure of fitting has been proposed as

$$\mathcal{I} = \frac{\ell(\hat{\boldsymbol{\theta}}) - \ell_0}{\ell_{sat} - \ell_0} .$$

A further normalized fitting measure has been introduced for comparing observed f_r and expected $p_r(\hat{\boldsymbol{\theta}})$ relative frequencies:

$$\mathcal{F}^2 \;=\; 1 - \frac{1}{2} \sum_{r=1}^{m} \mid f_r - p_r(\hat{\boldsymbol{\theta}}) \mid .$$

It is related to a standard dissimilarity index and has an immediate interpretation as the proportion of correct predicted responses.

A program in ***R*** –where the whole inferential procedure is effectively implemented with estimation, test results, statistical indexes and graphical displays– is freely available (Iannario and Piccolo, 2009). Work is currently in progress to release a standard ***R*** package.

3 Case study

In this Section we present the results of a case study dealing with sensory data about coffee tasting. Usually the coffee tasting method consists of three main evaluations (sensory attributes):

- the *visual analysis*, taking into account the colour (should not be either too light or too dark, but rather nutty-color with dark red streaks), the texture (should be dense, with a fine texture and without any gaps), and the persistence (quite long) of the cream;
- the *olfactory analysis*, taking into account the smell (should be pleasant and intense) and fragrances or aromas (toasted, chocolaty, floral, fruity, peanuts, spiced, ...);
- the *gustatory analysis*, taking into account flavour (sweet, acidic, bitter) and aftertaste (aroma, persistence).

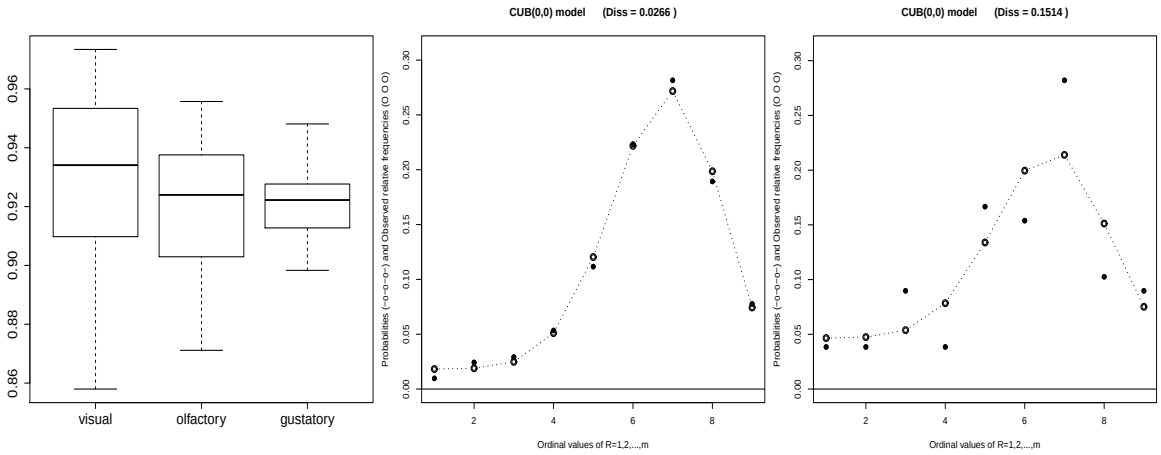


Figure 1: Boxplots of $\mathcal{F}^2$ separately for the three sensory attributes (left). Plot of estimated probabilities versus observed relative frequencies in the best ($\mathcal{F}^2 = 1 - Diss = 0.973$) and worst ($\mathcal{F}^2 = 1 - Diss = 0.849$) case (middle and right, respectively).

The survey which produced the analyzed data was carried out by Centro Studi Assaggiatori (CSA, http://www.assaggiatori.com) of Brescia, Italy, along with the International Institute of Coffee Tasters (IIAC)[1] and was concerned with the sensory analysis of 43 different coffee varieties, evaluated by a number of experienced and non-experienced judges through the above described tasting method. For each coffee variety a set of judges (from a minimum of 6 to a maximum of 421) was selected from the 1650 judges involved in the survey, who formulated visual, olfactory, gustatory evaluations of the coffee on an 9-point Likert scale. After removing the coffee varieties evaluated by less than 60 judges, the data set turns out to be composed by 36 coffee varieties for which a total number of 7604 judgments on each sensory attribute are available. On the whole, each of the 1650 judges was asked to taste from a minimum of 1 to a maximum of 11 coffees, but more than 78% of judges tasted exactly 5 coffee varieties. For each judge some personal information is also available (gender, age, experience in tasting, consumption, ...).

We fit CUB models separately to each of the 36 varieties of coffees with respect to visual, olfactory and gustatory perceptions. The estimated models are all significant and with good fitting measures ($\mathcal{F}^2$ varies in (0.849, 0.973)), as shown in Figure 1 (more detailed results are available from Authors).

We summarize results by plotting the estimated parameter vectors on the unit square. So, according to the estimated CUB models we locate the 36 coffee varieties on a map describing their relative positioning with respect to the selected sensory attributes, focusing attention on both the level of their evaluation and the degree of uncertainty of the judgements (Figure 2). It should be evident how the complex pattern of this experiment may be sharply simplified by CUB models in a unique representation. The ranking of preferences is not constant with respect to the three evaluations and this confirms that respondents react in different ways when faced to visual, olfactory and gustatory sensations. The close position of visual and olfactory

[1]The authors thank Luigi Odello (director of CSA) and prof. Eugenio Brentari (University of Brescia) for making the data available.

perceptions is a further confirmation of well known results in sensometrics: as a matter of fact, sight and smell are senses which manifest themselves with high similarity.

In addition, we notice that all evaluations (except for varieties 34 and 35) are expressed with a limited uncertainty, confirming that respondents are giving meditated preferences. However, the uncertainty generally increases when we move from visual to olfactory and then to gustatory perceptions. Thus, gustatory perceptions are more related to subjectivity than olfactory perceptions, which, in turn, are more related to subjectivity than the visual ones. Thus, we conjecture that perceptions more heavily depend on the personal history, attitude, and habits when we consider gustatory with respect to visual and olfactory sensations.

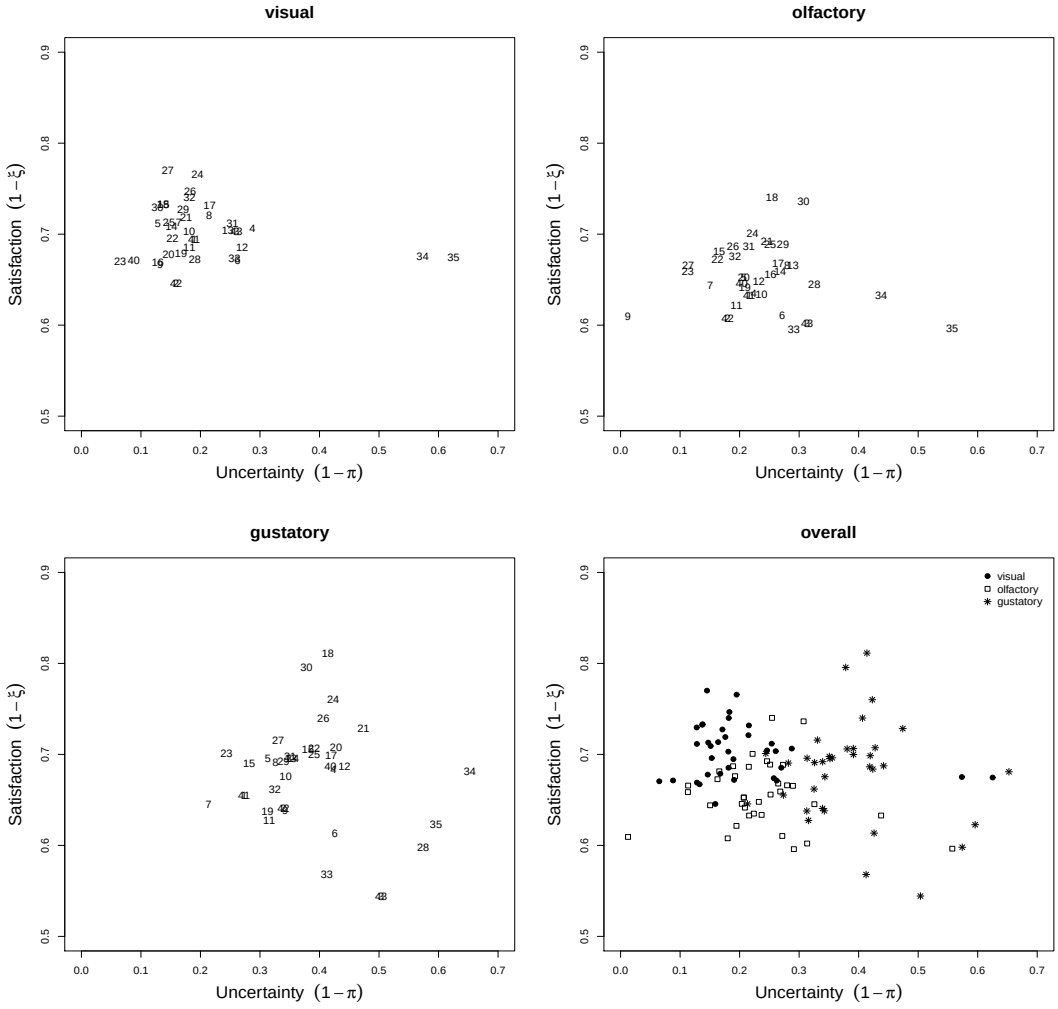


Figure 2: CUB models visualization of visual, olfactory, gustatory perceptions of the 36 coffee varieties.

Figure 3 shows the estimated models of visual, olfactory and gustatory perceptions for each coffee variety separately: gustatory perceptions are generally more uncertain and also the

atypical location of varieties 34 and 35 is confirmed.

In addition, some of the personal characteristics of the judges available from the questionnaire turned out to be a significant covariate for the estimated CUB models (see Iannario *et al.*, 2011).

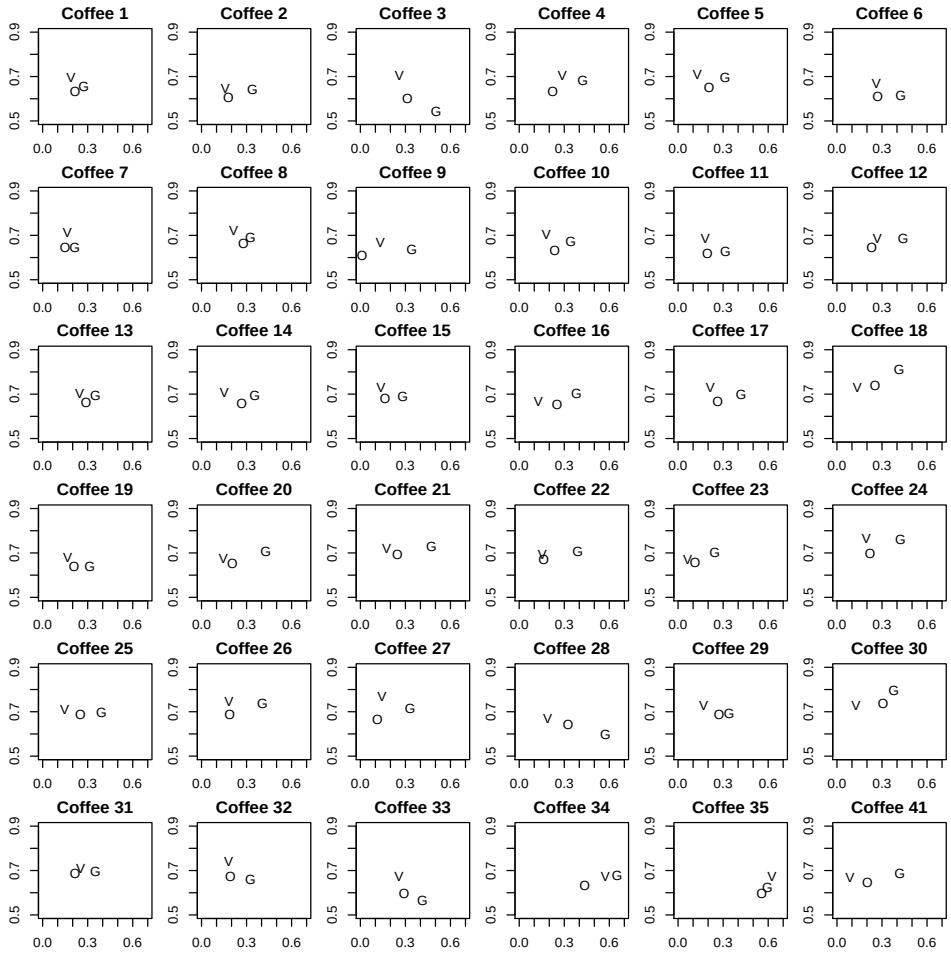


Figure 3: CUB models visualization of visual, olfactory and gustatory perceptions for each coffee variety.

At this step, it is interesting to inspect the relationships among the judges' perceptions expressed through the visual, olfactory and gustatory ratings and the satisfaction about each coffee variety. Since the gustatory satisfaction towards coffee is significantly dependent by both visual and olfactory ratings but olfactory is much more relevant, we explore the relationship between the expressed level of olfactory rating and the gustatory satisfaction for the 36 coffee varieties, separately. More specifically, using the expressed scores on the olfactory sensory attribute as covariate in the CUB model of gustatory satisfaction, we verify if gustatory satisfaction can be predicted by means of single judges' perceptions on the olfactory attribute. The logistic link $\xi_i = \dfrac{1}{1+e^{-\boldsymbol{w}_i\gamma}}$ is introduced in the CUB models, with ξ_i and $\boldsymbol{w}_i$ indicating respectively

the gustatory satisfaction and the olfactory rating of subject i, for $i = 1, 2, \ldots, n$. Results are plotted in Figure 4. The coffee varieties show different reaction rates but it is insightful to observe that the shape is regularly homogeneous for all the coffees. This result may be usefully exploited by producers of coffees with poor olfactory perceptions, since an improvement of the consumer gustatory perception towards the product seems to be highly dependent upon a positive evaluation of the coffee's smell.

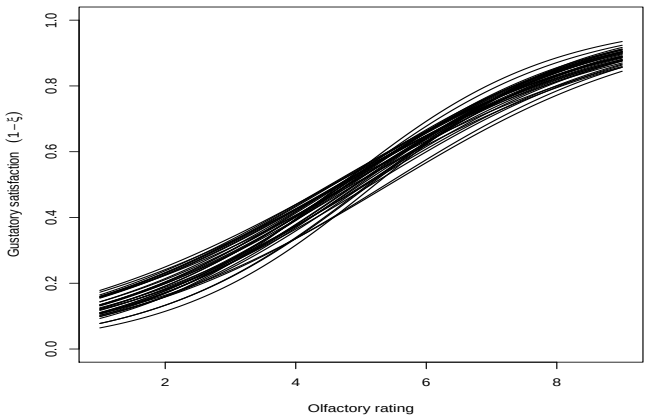


Figure 4: Prediction of gustatory satisfaction given the olfactory rating.

4 Concluding remarks

In this paper, CUB models have been studied for interpreting uncertainty and feeling of different brands of coffee but they manifest themselves as useful also for measuring the predictive ability of gustatory responses given the olfactory ones. The experimental results on a very large data set of different brand of coffees confirm that CUB models may be usefully exploited for comparing and summarizing several aspects of the data in an effective graphical display.

The analysis so far proposed may be further deepened if we could insert product characteristics in the sensory analysis. It could allow, for example, to identify which coffee varieties show a peculiar behaviour, in order to better understand relationships among variety and perceptions and to finally direct the manufacturers' efforts to improve their competitiveness.

Results from CUB models could be integrated with other advanced statistical techniques useful for sensory analysis (among others, Brentari and Zuccolotto, 2011) in order to get a more complete picture of the phenomenon (see, for example, Iannario *et al.*, 2011).

Acknowledgement. The second Author has been partly supported by MIUR projects PRIN2008: "Modelling latent variables for ordinal data: statistical methods and empirical evidence" (CUP E61J10000020001), within the research Unit at University of Naples Federico II.

References

Agresti A. (2010). *Analysis of Ordinal Categorical Data*, 2^{nd} ed., Wiley, NY.

Brentari E., Zuccolotto P. (2011). The impact of chemical and sensory characteristics on the market price of Italian red wines, *Electronic Journal of Applied Statistical Analysis*, **4**, 265–276.

Corduas M., Iannario M., Piccolo D. (2009). A class of statistical models for evaluating services and performances, in: M. Bini *et al.* (eds.), *Statistical methods for the evaluation of educational services and quality of products*, Springer, 99–117.

D'Elia A., Piccolo, D. (2005). A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, **49**, 917–934.

Everitt, B.S., Hand D.J. (1981). *Finite mixture distributions*, Chapman & Hall, London.

Iannario M. (2007). A statistical approach for modelling urban audit perception surveys, *Quaderni di Statistica*, **9**, 149–172.

Iannario M. (2009). Fitting measures for ordinal data models, *Quaderni di Statistica*, **11**, 39–72.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data, *Metron*, **LXVIII**, 87–94.

Iannario M. (2011). Modelling *shelter* choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, forthcoming.

Iannario M., Manisera M., Piccolo D., Zuccolotto P. (2011). Sensory analysis in the food industry as a tool for marketing decisions, *Submitted for publication.*

Iannario M., Piccolo D. (2009). A program in `R` for CUB models inference, Version 2.0, available at `http://www.dipstat.unina.it/CUBmodels1/`

Iannario M., Piccolo D. (2011). CUB Models: Statistical Methods and Empirical Evidence, in: Kenett, R. S. and Salini, S. (eds.), *Modern Analysis of Customer Surveys*, Wiley, NY, 231–254.

Kï¿½ster E. P. (2009). Diversity in the determinants of food choice: A psychological perspective, *Food Quality and Preference*, 20, 70–82.

McLachlan G., Krishnan T. (2008). *The EM algorithm and extensions*, 2^{nd} ed., Wiley, NY.

McLachlan G., Peel G. J. (2000). *Finite mixture models.* J. Wiley & Sons, NY.

Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, **5**, 85–104.

Piccolo D. (2006). Observed information matrix for MUB models, *Quaderni di Statistica*, **8**, 33–78.

Piccolo D., D'Elia A. (2008). A new approach for modelling consumers' preferences, *Food Quality and Preference*, **19**, 247–259.

Piccolo D., Iannario M. (2010). A new approach for modelling consumers' preferences, *Proceedings of the* 11^{th} *European Symposium on Statistical Methods for the Food Industry*, University of Sannio, Benevento, Academy School, Afragola, 139–148.

Prise en compte de l'expérience des sujets dans une épreuve de catégorisation : problématique et traitement des données

Taking into account the subjects experience in a free sorting task

Pauline Faye[1], Philippe Courcoux[1], El Mostafa Qannari[1], Agnès Giboreau[2]

[1]*ONIRIS, Unité de Sensométrie et Chimiométrie, Domaine de la Géraudière, 44 322 Nantes, France ; Université Nantes Angers Le Mans, France*. E-mail: pauline.faye@oniris-nantes.fr

[2]*Institut Paul Bocuse, Château du vivier, 8 chemin du Trouillat, 69130 Ecully, France ; Université Claude Bernard, Centre européen pour la nutrition et la santé, Lyon, France.*

Résumé

En analyse sensorielle, l'épreuve de tri libre s'avère particulièrement intéressante pour analyser les différences de perception entre les consommateurs. Il s'agit, ici, d'étudier l'impact de la connaissance dans le domaine du vin et de la dégustation sur la perception d'un ensemble de verres à vin. Les connaissances en vin de 209 consommateurs ont été évaluées par un quiz. Le modèle de Rasch a permis de construire un score de connaissance sur la base des réponses des consommateurs et d'identifier trois groupes nommés connaisseurs, intermédiaires et non connaisseurs. L'Analyse des Correspondances Multiples a permis de caractériser chacun des groupes et de corroborer la segmentation des consommateurs réalisée sur la base de leur connaissance. Le modèle INDSCAL appliqué aux données issues des tris libres a permis d'établir un espace de représentation des verres et déterminer le poids que chaque groupe accorde à chaque axe de la configuration. Les connaisseurs et les non connaisseurs perçoivent différemment les verres ; les connaisseurs donnant plus de poids aux axes liés à l'usage des produits.

Mots-clés : tri libre, modèle de Rasch, analyse des correspondances multiples, INDSCAL.

Abstract

In sensory analysis, the free sorting task is useful for the assessment of differences in perception among subjects. In this study, we investigate the impact of the experience and the knowledge in wine and wine consumption on the perception of a set of wine glasses. A panel of 209 consumers took part in a free sorting task of a set of wine glasses. They were also instructed to fill in a questionnaire composed, on the one hand, of questions regarding their socio-demographic characteristics and their usage and attitude towards wine and, on the other hand, of a quiz which aims at assessing the knowledge of the consumers regarding wine. The questions from the quiz were submitted to Rasch model. This resulted in ability scores associated with the consumers. On the basis of these scores, the subjects were segmented into three groups: non connoisseurs, intermediates and connoisseurs. These groups were further characterized by applying Multiple Correspondence Analysis on the data from the first part of the questionnaire. By applying INDSCAL model to the free sorting data from the three groups of consumers, it was possible to set up a stimuli space to represent the wine glasses and assess the importance that the three groups of consumers assign to the dimensions which underlie the space of representation of the wine glasses. It turned out that the consumers which were identified as connoisseurs put a heavier weight than the non connoisseurs on the dimensions which are related to the actual usage of the glasses.

Keywords : free sorting, Rasch model, Multiple Correspondence Analysis, INDSCAL

1. Introduction

Dans le cadre de l'évaluation sensorielle et des études des préférences, l'épreuve de catégorisation dite 'tri libre' connaît un intérêt grandissant car elle permet de cerner de manière simple et fiable les perceptions d'un ensemble de produits par un panel d'experts ou de consommateurs.
Basée sur le processus holistique de catégorisation (évaluation globale), l'épreuve de tri libre s'avère plus naturelle et permet d'évaluer un plus grand nombre de produits que les approches sensorielles analytiques. Du fait de son caractère intuitif, global, non verbal et comparatif, elle est particulièrement adaptée pour appréhender la perception de produits par des consommateurs (Lawless *et al*, 1995 ; Faye *et al*, 2004 ; Cartier *et al*, 2006), étudier l'organisation des connaissances et des représentations (Picard *et al*, 2003 ; Dacremont *et al*, 2006) ou révéler des différences de perceptions entre consommateurs en termes culturel (Chrea *et al*, 2004 ; Blancher *et al*, 2007) ou en termes d'expertise (Giboreau *et al*, 2001 ; Soufflet *et al*, 2004 ; Lelièvre *et al*, 2008 ; Ballester *et al*, 2008). Concernant l'utilisation du tri pour étudier l'impact de l'expertise sur la perception, les conclusions sont variables selon les auteurs, en fonction des produits mais aussi du type d'expertise considérée.

Nous nous intéressons particulièrement aux différences de perceptions entre les consommateurs en fonction de leur connaissance préalable du produit. Cette connaissance peut être théorique (basée sur des connaissances scientifiques et empiriques) ou pratique (basée sur l'expérience et l'usage). Considérant que la connaissance préalable d'un sujet et sa familiarité au produit peut influencer sa perception, l'objectif de ce travail est de proposer une approche méthodologique permettant de classer, sur la base des résultats d'un quiz, les sujets en fonction de leurs connaissances vis-à-vis des produits concernés par l'étude. Par la suite, cette classification est étayée et validée par les réponses des sujets à un questionnaire sur les usages et pratiques de consommation. Les résultats d'une épreuve de tri libre sont analysés en tenant compte des groupes déterminés dans la première phase. L'approche générale est illustrée grâce à une étude de cas portant sur 209 sujets et visant à évaluer la perception de verres à vin par les consommateurs. Au-delà de la problématique scientifique consistant à évaluer l'impact des connaissances a priori sur la perception, l'intérêt de notre étude porte également sur le traitement statistique des données. Pour cela, nous faisons appel à des techniques différentes telles que le modèle de Rasch, l'Analyse des Correspondances Multiples et le modèle INDSCAL.

2. Matériel et méthodes

2.1 Procédure

2.1.1 Echantillons

Trente verres à vin, présentant des différences en termes de taille, volume, forme et usage (verres à champagne, verres à vin rouge ou vin blanc) ont été sélectionnés dans les collections ARC international (Table 1). Les verres sont présentés sur photos mates (format 10*15) à 60% de leur taille réelle. Les photos codées aléatoirement sont disposées sur une nappe grise selon un carré latin de Williams pour balancer les effets d'ordre de présentation et de report d'ordre 1.Toutes les photos sont présentées simultanément au sujet.

1 (35)	**4**(26)	**7** (16.5)	**10** (47)	**13** (38)	**16** (47)	**19** (47)	**22** (62)	**25** (19)	**28**(19)
2 (41)	**5** (55)	**8** (32)	**11**(20)	**14** (30)	**17**(75)	**20** (24)	**23** (47)	**26** (45)	**29**(31)
3 (47)	**6** (73)	**9** (40)	**12**(47)	**15** (16)	**18** (35)	**21** (47)	**24** (18)	**27** (16)	**30**(38)

Table 1 : Codes, volumes (cl) et illustration des 30 verres à vin

2.1.2 Sujets

Deux cents neuf consommateurs de vin ont participé à l'expérimentation. Afin d'assurer la plus grande variabilité possible en termes de connaissances en vin et en dégustation mais aussi en termes de pratiques, les consommateurs ont été recrutés a priori selon des critères de fréquence de consommation (de consommation quotidienne à exceptionnelle). Le panel de consommateurs sélectionné est, par ailleurs, équilibré en termes d'âge, sexe et activité professionnelle. Afin de pouvoir caractériser leurs pratiques et d'évaluer leurs connaissances, les consommateurs sont invités à remplir un questionnaire en deux parties en fin d'expérimentation.

La première partie comporte 20 questions déclaratives sur les usages et pratiques en termes d'achat, de consommation et de formation en vin, les caractéristiques sociodémographiques mais également sur l'évaluation de leur propre expertise en vin.

La deuxième partie (quiz) porte sur l'évaluation des connaissances en vin et en dégustation et est constituée de 26 questions objectives sur le vin et les verres à vin, validées par des professionnels lors d'un pré-test. A titre d'exemples, les questions sont du type « Quelle est la signification du sigle A.O.C ? » ou « Quel artiste est à l'honneur sur l'étiquette du Mouton-Rothschild 1973 ? (A) Pablo Picasso, (B) Marc Chagall ou (C) Andy Warhol. La première question s'avère a priori assez facile alors que la deuxième est sans doute plus difficile pour les consommateurs. Les réponses à ces questions sont codées de manière binaire, 0 pour une mauvaise réponse et 1 pour une bonne réponse. Au total, 72 variables binaires permettent de coder l'ensemble des données issues de cette deuxième partie du questionnaire.

2.1.2 Expérimentation et codage des données

L'expérimentation est menée dans différents endroits, en fonction du lieu de recrutement des sujets (cabine d'évaluation sensorielle à l'université, à domicile, dans des bars et des restaurants, chez des cavistes).
La procédure se décompose en deux étapes consécutives :

Étape 1 : la tâche de tri libre consiste à réaliser des groupes des photos de verres selon les ressemblances et les différences perçues. Pour chaque sujet, les données issues d'une épreuve de tri libre se présentent sous forme d'une partition d'un ensemble de produits (verres à vin).

Étape 2 : une fois les groupes formés, chaque sujet est invité à décrire librement chacun des groupes de verres réalisés avec ses propres termes (mots ou expressions).

2.2 Analyse des données

2.2.1 Questionnaire sur les usages, pratiques et connaissances en vin

2.2.1.1 Score de connaissance en vin et groupes de consommateurs

Afin de comparer le niveau de connaissance en vin des consommateurs, un score de connaissance est déterminé pour chaque sujet. Ce score est obtenu grâce au modèle de Rasch appliqué aux 72 variables binaires issues des réponses au quiz sur le vin.
Le modèle de Rasch est très populaire dans de nombreux domaines d'applications notamment en psychométrie (Boomsma *et al*., 2000). Il permet de déterminer des scores de performances des sujets en tenant compte des difficultés des questions constituant le quiz. Ainsi, un sujet aura un score de performance d'autant plus élevé qu'il répondra correctement à un grand nombre de questions, notamment des questions difficiles (i.e. les questions pour lesquelles le nombre total de bonnes réponses est relativement faible). De manière plus précise, le modèle de Rasch est basé sur une régression logistique qui intègre deux effets : l'effet sujet et l'effet question. Le modèle stipule que pour un sujet h et une question m, la probabilité P_{mh} que le sujet h donne est une bonne réponse pour la question m est donnée par : $P_{hm} = \frac{e^{\beta_h - \delta_m}}{1 + e^{\beta_h - \delta_m}}$ où β_h reflète la performance du sujet h et δ_m le niveau de difficulté de la question m.

Dans notre étude, nous nous intéressons exclusivement à la performance des sujets qui correspond au score de connaissances en vin. Ainsi, le modèle de Rasch permet de déterminer un score pour chaque sujet : plus le score est élevé, plus le consommateur est connaisseur en vin. Par la suite, les 209 consommateurs sont segmentés en fonction de leur score de connaissance.

2.2.1.2 Caractérisation des consommateurs.

Dans l'objectif de caractériser les groupes de consommateurs déterminés par le niveau de connaissance, la première partie du questionnaire est utilisée (i.e. critères sociodémographiques, pratiques, usages, consommation de vin, formation en œnologie, expertise auto-évaluée). Un test de Chi2 permet de déterminer les critères qui différencient les groupes de consommateurs de manière significative au seuil de signification de 1%. Les questions qui se sont révélées comme ayant un effet non significatif sont exclues de l'analyse.
Une analyse des correspondances multiples (ACM) est ensuite réalisée sur les données catégorielles du tableau (sujets x questions). Cette analyse descriptive permet de visualiser les proximités entre les sujets, les modalités des questions et de caractériser les sujets et les groupes de sujets.

Afin d'étudier le lien entre usages, pratiques et connaissances en vin et en dégustation, le score de connaissance issu du modèle de Rasch est intégré en variable illustrative à l'ACM.

2.2.2 L'épreuve de tri libre

Chaque partition individuelle est associée à une matrice binaire de dissimilarités entre produits, où 0 signifie que deux produits sont dans le même groupe et 1 qu'ils sont dans des groupes différents. Les matrices de dissimilarités sont agrégées par groupe de sujets pour obtenir autant de matrices de dissimililarités que de groupes de consommateurs.
Un usage classique est de réaliser, pour chaque groupe de consommateur, une MDS non métrique sur les matrices de dissimilarités agrégés afin de représenter les proximités entre les verres et d'identifier les dimensions sous-jacentes qui structurent ces proximités (Kruskal, 1964; Schiffman *et al.*, 1981; Borg and Groenen, 1997). Une des limites de cette approche est de ne pas tenir compte directement des différences entre les groupes de sujets dans l'analyse des proximités des produits. La méthode MDS sur dissimilarités individuelles basée sur le modèle INDSCAL (Carrol et Chang, 1970) permet de pallier cette limite. Cette méthode permet de positionner les produits dans un espace multidimensionnel de faible dimension et de déterminer le poids que chaque groupe de sujets accorde à chacun des axes. Les axes factoriels reflètent des variables latentes qui structurent l'espace perceptif et les poids sont directement liés à la dispersion des produits sur chaque axe. Ainsi, l'introduction de poids permet de tenir compte du fait que les groupes de sujets ne perçoivent pas nécessairement ces variables latentes avec la même intensité.

De manière relativement sommaire, supposons que nous disposions de H tableaux de dissimilarités. Soit X l'espace de représentation de dimension p, fixée par l'utilisateur, nous désignons par $d_{ij}^{h}(X)$ la distance entre les produits i et j pour le tableau h (h=1,…,H) dans cet espace. Soit X_{ia} la coordonnée du produit i sur l'axe a de X et $w_a^{(h)}$ le poids du tableau h sur cet axe. L'espace de représentation optimal est obtenu par minimisation du Stress défini par:

$$Stress = \left[\sum_h \sum_{ij} \left(\delta_{ij}^{(h)} - d_{ij}^{(h)}(X) \right)^2 / \sum_h \sum_{ij} d_{ij}^{(h)}(X)^2 \right]^{1/2} \quad avec \quad d_{ij}^{(h)}(X) = \sqrt{\sum_a w_a^{(h)} (x_{ia} - x_{ja})^2}$$

Ce problème de minimisation est résolu de manière itérative (Carrol et Chang, 1970).

Il est clair que le modèle INDSCAL donne des informations utiles sur le positionnement des produits les uns par rapport aux autres et permet également de mettre en évidence des différences entre les tableaux de dissimilarités, ce qui peut être un objectif en soi pour une épreuve de catégorisation. Cependant, il faut souligner que son aspect métrique nous semble peu adapté aux données issues du tri libre. C'est pourquoi nous préconisons d'adopter une version non métrique de l'approche INDSCAL. A notre connaissance, cette approche tout à fait adaptée aux données issues d'une épreuve de tri libre n'a jamais été utilisée en évaluation sensorielle. La méthode INDSCAL avec transformation non métrique développée dans le contexte de la psychométrie par Takane *et al* (1977) est basée sur la minimisation du critère suivant :

$$Stress = \left[\sum_h \sum_{ij} \left(f^{(h)}(\delta_{ij}^{(h)}) - d_{ij}^{(h)}(X) \right)^2 / \sum_h \sum_{ij} d_{ij}^{(h)}(X)^2 \right]^{1/2} \quad avec \quad d_{ij}^{(h)}(X) = \sqrt{\sum_h w_a^{(h)} (X_{ia} - X_{ja})^2}$$

où $f^{(h)}$ *(h=1,..., H)* sont des fonctions (à déterminer) croissantes.

3. Résultats

3.1 Groupes de consommateurs

3.1.1 Identification des groupes

Le modèle de Rasch a été appliqué sur les réponses des 209 consommateurs aux questions du quiz sur le vin. Les scores individuels sont rangés par ordre croissant et représentés sur la figure 1. La répartition des scores met nettement en évidence deux points d'inflexion et permet ainsi de définir trois groupes de sujets : deux groupes extrêmes correspondant chacun à 20% de l'effectif total et un groupe intermédiaire regroupant 60% des consommateurs. Ainsi, avec en moyenne 17% de bonnes réponses, les 20% des consommateurs (41 sujets) dont les scores sont les plus faibles sont définis comme les non connaisseurs. A l'opposé, avec en moyenne 75% de bonnes réponses, les 20% des consommateurs (41 sujets) dont les scores sont les plus élevés sont définis comme connaisseurs. Les 60% des consommateurs (127 sujets) dont les scores sont intermédiaires présentent en moyenne 42% de bonnes réponses au quiz sur le vin et sur la dégustation.

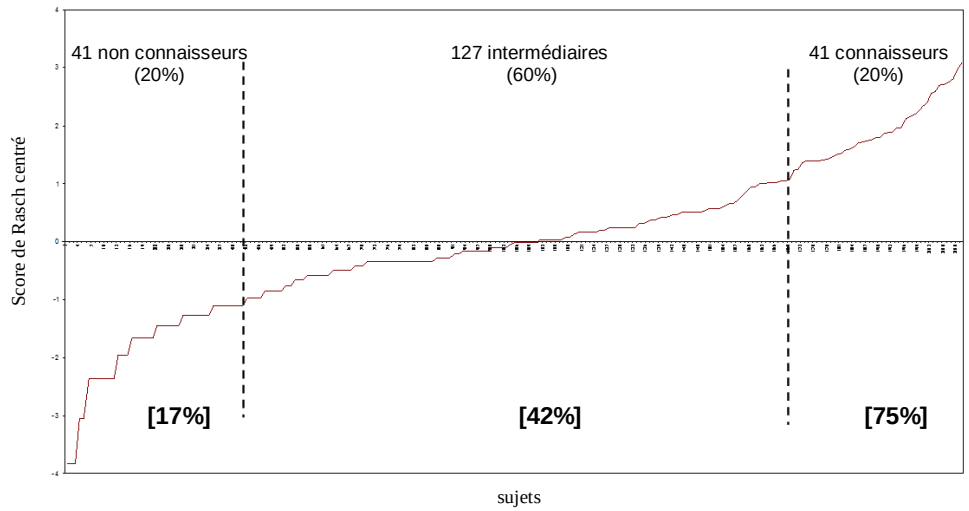


Figure 1 : Scores individuels de connaissance en vin issus du modèle de Rasch, centrés et ordonnés **[moyenne du pourcentage de bonnes réponses par groupe]**

3.1.2 Caractérisation des groupes

Hormis les connaissances préalables en vin et en dégustation qui ont permis de définir les trois groupes de consommateurs, les questions concernant les critères sociodémographiques, les usages, les pratiques et une auto-évaluation du niveau expertise sont utilisées pour caractériser les différences entre les groupes. A cet effet, une ACM est réalisée en considérant les variables descriptives sur les usages et les pratiques comme variables actives et les variables sociodémographiques, les variables concernant l'expertise auto-évaluée (variables qualitatives) et le score Rasch (variable quantitative) comme illustratives. Le premier plan factoriel (non représenté ici par manque de place) de l'ACM restitue 20.7% de l'inertie totale. Le premier axe factoriel permet de différencier, en particulier, les

sujets qui ont répondu « jamais » aux questions concernant l'achat et qui sont donc assez peu investis dans la consommation. Le deuxième axe factoriel s'avère particulièrement intéressant puisqu'il permet de bien séparer les trois groupes de consommateurs. La corrélation relativement forte du score de Rasch avec le deuxième axe factoriel (R=-0,65) confirme cette observation. L'analyse de l'ACM et des tests de chi2 permet de décrire les trois groupes de consommateurs de la manière suivante :

- *Connaisseurs* : Agés de 30 à 50 ans, ce sont des professionnels du vin (cavistes, viticulteurs) ou des consommateurs avisés, formés à l'œnologie. Ils consomment du vin très fréquemment et possèdent une cave. Ils préfèrent réaliser leurs achats dans des magasins spécialisés et basent leurs choix d'experts sur le cépage, le millésime et le nom du domaine. Ils se jugent experts en vin.
- *Non connaisseurs* : Principalement étudiants, ouvriers ou employés, ils consomment du vin très occasionnellement. Non formés à l'œnologie, ils ne possèdent que quelques bouteilles de vin, qu'ils choisissent au supermarché selon des critères liés principalement au prix. Ils ne se sentent pas compétents dans le domaine du vin.
- *Intermédiaires*: Comparés aux deux groupes précédents, ils ne présentent pas de caractéristiques particulières.

Ce résultat confirme que la connaissance préalable en vin et en dégustation est liée aux usages et pratiques des consommateurs, recrutés dans le cadre de cette expérimentation. Le score de connaissance s'avère être un bon indicateur d'une connaissance générale sur le vin qui peut être théorique ou liée à l'expérience des consommateurs. Cette conclusion tend à valider la pertinence de la segmentation des consommateurs sur la base de leur connaissance préalable sur le vin et la dégustation et d'étudier les différences de perception entre les groupes ainsi définis.
Le groupe des intermédiaires ne présentant pas de caractéristiques particulières, seuls les deux groupes extrêmes seront comparés dans la suite de l'analyse.

3.2 Les différences de perceptions des verres entre les consommateurs

Le modèle INDSCAL non métrique est appliqué aux matrices agrégées des connaisseurs et des non connaisseurs. Avec un stress de 0,013, la configuration de dimension 3 est conservée. Les figures 2 et 3 représentent respectivement les proximités entre les verres sur les axes 1 et 2 et 1 et 3 de la configuration. Pour interpréter les dimensions sous-jacentes qui structurent cet espace perceptif, les corrélations entre les axes issus d'INDSCAL et les termes générés lors de l'épreuve de tri par les connaisseurs d'une part et les non connaisseurs d'autre part sont calculées. Les descriptions recueillies nous renseignent sur la nature des propriétés prises en compte par les sujets dans la catégorisation des verres et donc structurant leurs perceptions. L'analyse de ces corrélations nous indique que :

- l'axe 1 oppose les verres à champagnes aux verres à vin pour les deux groupes de consommateurs
- l'axe 2 oppose
    - pour les non connaisseurs, les verres ronds aux verres avec un angle, qualifiés de contemporains.
    - pour les connaisseurs, les verres classiques pour grands vins qui permettent d'aérer le vin aux verres modernes qui permettent un retour des arômes.
- l'axe 3 oppose
    - pour les non connaisseurs, les verres qualifiés de petits et beaux aux verres qualifiés de normaux et banaux.
    - pour les connaisseurs, les verres standards, démodés et peu appréciés, pour le bistrot ou pour le domicile, aux verres élégants adaptés à la dégustation ou pour recevoir des amis.

A la différence des connaisseurs, l'axe 3 est expliqué par un faible nombre termes pour le groupe des non connaisseurs (corrélations faibles des termes avec l'axe 3).

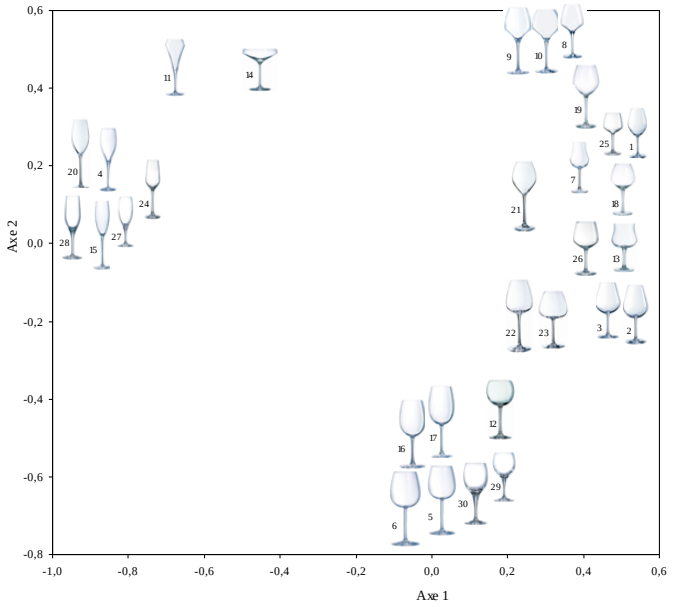


Figure 2 : Axes 1 et 2 de la configuration produits issue du modèle INDSCAL

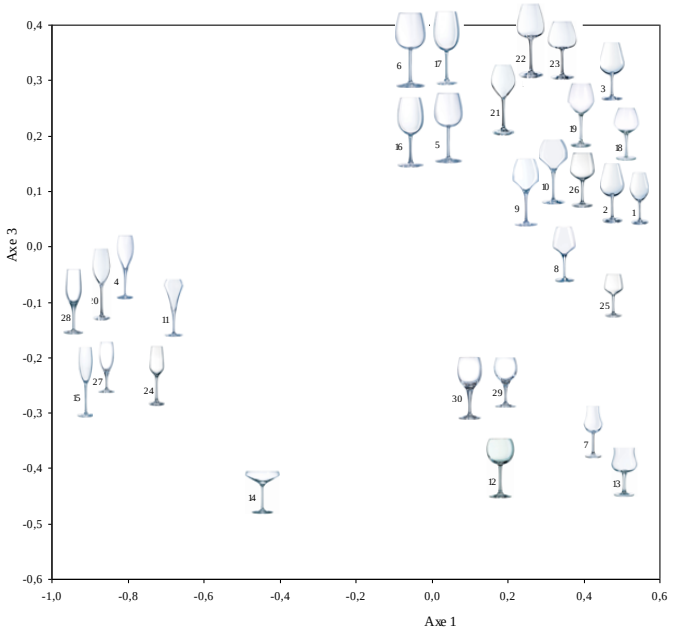


Figure 3 : Axes 1 et 3 de la configuration produits issue du modèle INDSCAL

Ces résultats indiquent que les consommateurs n'emploient pas le même type de vocabulaire pour décrire les verres. Ainsi, les perceptions des connaisseurs se structurent principalement selon des propriétés de forme, des qualificatifs (modernes, classiques…) mais surtout sur des propriétés d'usage (retour d'arômes, aération du vin, verre à dégustation, verre pour usage quotidien). Quant aux non connaisseurs, ils semblent se focaliser sur des propriétés physiques des verres et sur quelques qualificatifs et propriétés d'usages qui s'avèrent consensuelles et partagées par l'ensemble des consommateurs interrogés (verres à vin, verres à champagne).

La figure 4 représente les poids que chaque groupe de consommateurs associe aux différents axes de la configuration. Cette figure indique que les deux groupes de consommateurs accordent des poids pratiquement identiques au premier axe factoriel, opposant verres à champagnes et verres à vin. En revanche, les connaisseurs accordent des poids similaires aux axes 2 et 3 alors que les non connaisseurs accordent relativement plus d'importance à l'axe 2 qu'à l'axe 3 qui semble moins structurant dans leur perception.

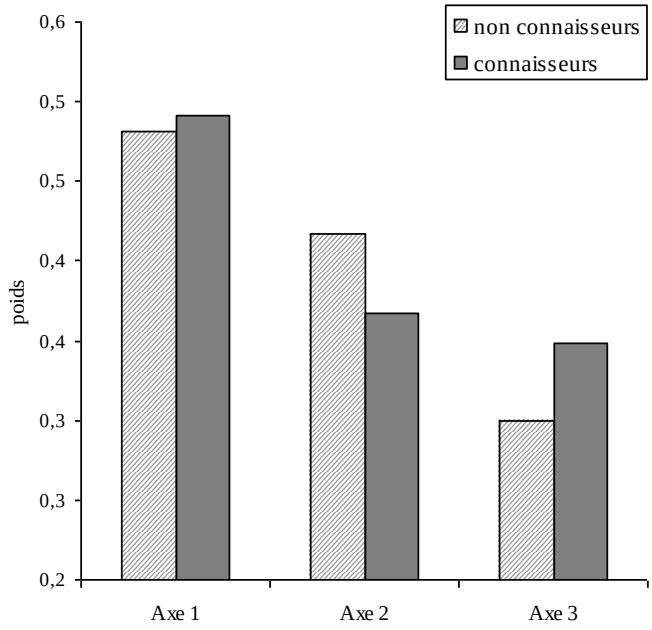


Figure 4 : Répartition des poids par axe de la configuration INDSCAL et par groupe de consommateurs

L'interprétation des axes indique que l'axe 3 est lié principalement à des propriétés d'usages. Ces propriétés structurant peu la perception des non connaisseurs, ils semblent accorder un poids relativement moins important à l'axe 3 que les connaisseurs.

4. Discussion et conclusions

Cette recherche avait pour objectif de proposer une démarche méthodologique pour identifier des groupes de consommateurs selon leur connaissance préalable et d'examiner l'impact de cette connaissance sur leur perception.

Les résultats indiquent que des consommateurs de vin peuvent présenter des niveaux de connaissance différents. Certains d'entre eux peuvent être considérés comme très avisés ou « experts » au même titre que des professionnels du vin car ils partagent les mêmes connaissances qui peuvent être théoriques, acquises par l'expérience ou induites par les pratiques de consommations. Dans cette étude, le niveau de connaissance en vin et en dégustation est très lié aux pratiques et usages des consommateurs. Les résultats confirment notre hypothèse concernant l'impact de la connaissance préalable sur la perception : les non connaisseurs structurent majoritairement leur perception sur des propriétés physiques liées à la forme des verres et les connaisseurs sur des propriétés plus globales liées à l'usage et au jugement.

Sur le plan des méthodes statistiques, le modèle de Rasch s'est révélé tout à fait pertinent pour synthétiser les niveaux de connaissance en vin des sujets. Sur la base des scores obtenus à l'aide de ce modèle, il a été relativement facile de classer les sujets en trois groupes. L'ACM a permis de mieux expliquer et caractériser les groupes et de corroborer la segmentation des consommateurs en trois groupes, sur la base de leur connaissance préalable dans le domaine du vin. Le modèle INDSCAL présente l'intérêt de pouvoir directement comparer les deux groupes de consommateurs sur la base du même espace perceptif.

Outre l'intérêt de la méthodologie générale qui permet d'appréhender les niveaux de connaissance des sujets dans un domaine particulier et d'en tenir compte dans l'analyse des résultats d'une épreuve de catégorisation, nous soulignons dans cette article l'apport d'outils statistiques tels que les modèles de Rasch et INDSCAL et l'analyse des correspondances multiples.

Bibliographie

Ballester J., Patris B., Symoneaux R., Valentin D. (2008). Conceptual vs perceptual wine spaces : Does expertise matter? *Food Quality and Preference*, 19, 267-276

Blancher G., Chollet S., Kesteloot R., Nguyen Hoang D., Cuvelier G., Siefferman J.-M. (2007). French and Vietnamese: How do they describe texture characteristics of the same food ? A case study with jellies. *Food Quality and Preference***,** 18, 560 -575.

Borg I., Groenen P. (1997). Modern Multidimensional scaling. Theory and applications. Springer series in statistics, Springer Verlag, New York.

Boomsma A., Van Duijn M.A.J., Snijders T.A.B. (2000). Essays on Item Response Theory. Springer-Verlag, New York.

Carroll J.D., Chang J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalisation of 'Eckart-Young' decomposition. *Psychometrika*, 35, 283-319

Cartier R., Rytz A., Lecomte A., Poblete F., Krystlik J., Belin E., Martin N.(2006). Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map. *Food Quality and Preference*, 17, 562-571

Chrea C., Valentin D., Sulmont-Rossé C., Ly Mai H., Hoang Nguyen D., Abdi H. (2004). Culture and odor categorization : agreement between cultures depends upon the odors. *Food Quality and Preference*, 15, 669-679.

Dacremont C., Soufflet I. (2006). Impact of fabric end-use knowledge on handle perception. *Revue européenne de psychologie appliquée*, 56, 273-277.

Faye P., Brémaud D., Durand Daubin M., Courcoux P., Giboreau A., Nicod H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mapping. *Food Quality and Preference, 15,* 781-791.

Giboreau A., Navarro S., Faye P., Dumortier J. (2001). Sensory evaluation of automotive fabrics :the contribution of categorization tasks and non verbal information to set-up a descriptive method of tactile properties. *Food Quality and Preference*, 12, 311-322.

Kruskal J.B. (1964). Nonmetric multidimensionnal scaling : a numerical method. *Psychometrika*,29, 115-129.

Lawless H., Sheng N., Knoops Stan S.C.P. (1995). Multidimensional scaling of sorting data applied to cheese perception. *Food Quality and Preference*, 6, 91-98.

Lelièvre M., Chollet S., Abdi H., Valentin D. (2008). What is the validity of the sorting task for describing beers ? A study using trained and untrained assessors. *Food Quality and Preference*, 19, 697-703.

Picard D., Dacremont C., Valentin D., Giboreau A. (2003). Perceptual dimensions of tactile textures. *Acta Psychologica*, 114, 165-184.

Schiffman S.S., Reynolds M.L., Young F.W. (1981). Introduction to multidimensional scaling. Theory, methods and applications. Academic Press, Orlando

Soufflet I., Calonnier M., Dacremont C. (2004). A comparison between industrial experts' and novices' haptic perceptual organization: a tool to identify descriptors of the handle fabrics. *Food Quality and Preference*, 15, 689-699.

Takane Y., Young F.W, De Leeuw J.(1977) Nonmetric Individual differences multidimensional scaling : An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 8-67

Ellipses de confiance pour les approches holistiques

Confidence ellipses in holistic approaches

Marine Cadoret & François Husson

Applied mathematics department - Agrocampus Rennes
65 rue de Saint-Brieuc - 35042 Rennes cedex (France)
E-mail : *cadoret@agrocampus-ouest.fr*
E-mail : *husson@agrocampus-ouest.fr*

Résumé
L'évaluation sensorielle de produits par un jury de consommateurs ou d'experts par des méthodes holistiques met en jeu des données de nature qualitative (en catégorisation), quantitatives structurées en groupes (en napping ou en profil libre), quantitatives et qualitatives (en napping catégorisé). Quelle que soit la nature des données, une configuration moyenne des produits est construite et il est crucial d'évaluer sa stabilité surtout quand les juges ne sont pas entraînés. Malheureusement, la plupart des ellipses de confiance construites autour de la position des produits ne constituent pas une zone de confiance dans le sens où deux produits sont significativement différents lorsque leurs ellipses ne se chevauchent pas. En effet, la plupart des ellipses proposées dans la littérature sont trop petites. Elles conduisent l'utilisateur à des interprétations fausses car trop optimistes. Nous proposons ici d'utiliser le bootstrap total pour construire des ellipses de confiance qui s'interprètent réellement comme des zones de confiance.
Mots-clés : Bootstrap total, ellipses de confiance, méthode holistique, tri hiérarchique, napping, napping catégorisé

Abstract
The sensory evaluation of products by a panel of experts or consumers with holistic methods involves qualitative data (in sorting task description), quantitative variables structured by groups (in napping or in free choice profiling), quantitative and qualitative (in sorted napping). Whatever the nature of data, a mean configuration of products is built and it is crucial to assess its stability, especially when the judges are untrained. Unfortunately, most of the confidence ellipses constructed around the position of the products do not give a confidence area in the sense that two products are significantly different when their ellipses do not overlap. Indeed, most ellipses proposed in the literature are too small. They lead the user to misinterpretations because too optimistic. Here we propose to use total bootstrap to build confidence ellipses that can be actually interpreted as confidence areas.
Keywords: Total bootstrap, confidence ellipses, holistic method, sorting task, napping, sorted napping, hierarchical sorting task

1. Introduction

La construction d'ellipses de confiance pour évaluer la stabilité d'une carte des produits obtenue à l'issue d'une dégustation sensorielle est cruciale. En effet, il n'est pas possible de savoir si deux produits sont perçus comme différents d'un point de vue sensoriel si leur position sur la carte n'est

pas agrémentée d'une zone de confiance. En profil classique (QDA), les ellipses de confiance proposées par Husson *et al.* (2005) sont utilisées pour construire les cartes des produits obtenues à l'issue d'analyse en composantes principales (ACP) ou d'analyse factorielle multiple (AFM) quand une structure doit être prise en compte sur les descripteurs.

Le principe de construction de ces ellipses est d'utiliser l'information apportée par chaque juge pour construire des zones de confiance autour de la position moyenne des produits. Plus précisément, des jurys virtuels de même taille que le vrai jury sont construits grâce à un tirage aléatoire et avec remise des juges. Les moyennes des jurys virtuels sont alors calculées par produit et par descripteur. L'analyse factorielle (ACP ou AFM) est ensuite lancée avec comme individus actifs les produits vus par le vrai jury et comme individus supplémentaires les produits vus par chacun des jurys virtuels (voir tableau de droite figure 1). Ces projections des produits vus par les jurys virtuels représentent alors l'incertitude autour de la position moyenne d'un produit fournie par le jury. Cette incertitude peut alors être matérialisée par des enveloppes convexes ou des ellipses de confiance.

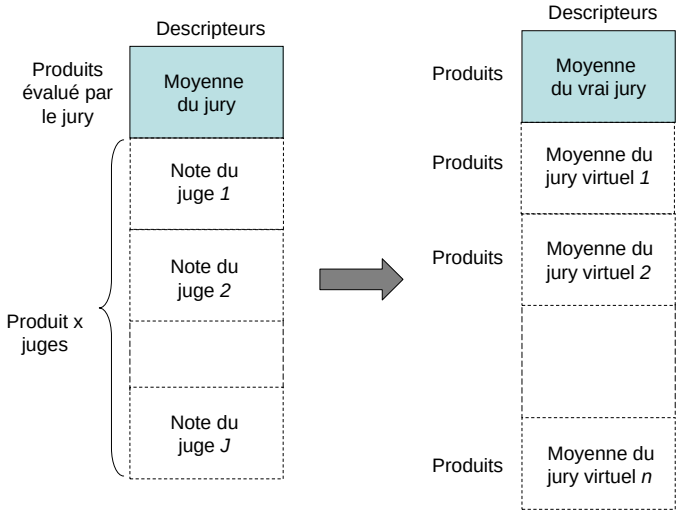


Figure 1 : Structure du jeu de données (à gauche) et des jurys virtuels (à droite) pour le profil classique

Pour les approches holistiques, les évaluations individuelles sont directement utilisées pour construire l'évaluation du jury. En catégorisation par exemple, l'évaluation d'un juge correspond à une colonne du tableau qui résume l'ensemble des classes à laquelle le juge affecte chacun des produits. En napping, les données d'un juge sont deux variables quantitatives correspondant aux abscisses et ordonnées de la position des produits. En napping catégorisé, les données d'un juge correspondent à une variable qualitative correspondant à la classe à laquelle le juge affecte le produit et à deux variables quantitatives correspondant aux coordonnées des produits. En profil libre, les données d'un juge correspondent à autant de variables quantitatives que de descripteurs choisis et évalués par le juge. Pour toutes ces méthodes, des jurys virtuels peuvent être constitués par rééchantillonnage des juges. Cependant, les jurys virtuels peuvent être considérés comme supplémentaires uniquement en colonnes (voir figure 2) car d'un jury à l'autre seules les lignes sont communes.

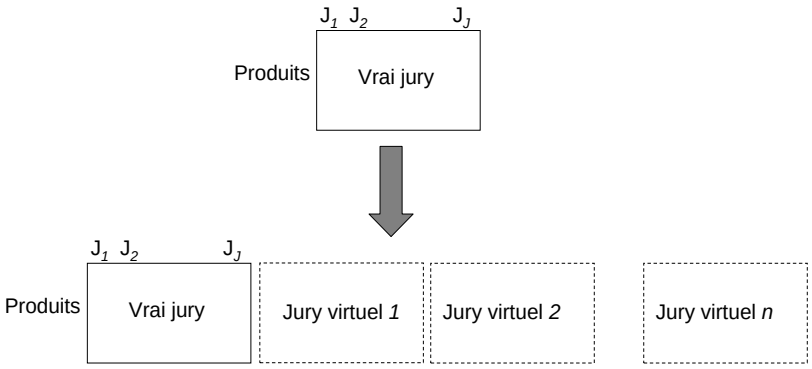


Figure 2 : Structure du jeu de données (en haut) et des jurys virtuels (en bas) pour les approches holistiques

2. Construction d'ellipses de confiance par la méthode de Cadoret *et al*. (2009) en catégorisation

2.1 Principe de la méthode de Cadoret *et al*. (2009)

En catégorisation, Abdi *et al.* (2007) proposent de construire des enveloppes convexes autour de la position de chaque produit obtenue par la méthode DISTATIS. Cependant, ils s'intéressent à la position relative des produits vus par chacun des juges et non à la stabilité de la position des produits vus par l'ensemble du jury. Or c'est la variabilité autour de la position moyenne du jury qui permet de déterminer si le jury, dans son ensemble, a réussi à différencier ou non les produits d'un point de vue sensoriel.

Cadoret *et al*. (2009 et 2011) ont proposé de construire les régions de confiance autour des produits vus par l'ensemble du jury. Pour ce faire, ils construisent un jury virtuel en choisissant des juges au hasard avec remise. Pour chacun de ces juges, ils projettent la position de chaque classe de produits sur la configuration obtenue par le vrai jury. Puis, pour chaque produit, ils calculent le barycentre des positions des classes (de chaque juge du jury virtuel) à laquelle ce produit appartient. Ils obtiennent alors, pour chaque produit, une nouvelle position du produit pour un jury virtuel. Cette procédure est répétée de nombreuses fois pour obtenir des positions de chaque produit vu par de nombreux jurys virtuels. Pour chaque produit, une ellipse de confiance ou une enveloppe convexe contenant 95% de ces positions est construite. La figure 3 donne la représentation des produits avec les ellipses de confiance autour de la position de chaque produit pour le jeu de données de catégorisation où 12 parfums sont décrits par 98 juges (Cadoret *et al*., 2009).

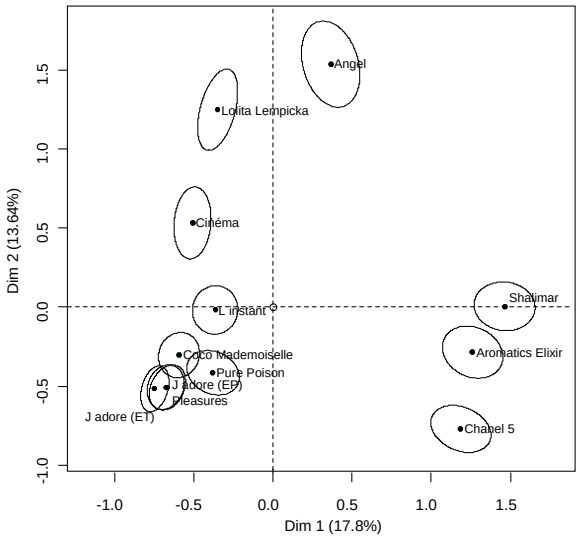


Figure 3 : Représentation des produits avec ellipses de confiance obtenue par la méthode de Cadoret *et al*. (2009).

2.2 Evaluation de la méthode de construction des ellipses de confiance

Pour évaluer la pertinence des ellipses de confiance, il est possible de perturber le jeu de données pour supprimer la structure présente dans les données. Pour ce faire, nous conservons les données d'un juge mais permutons, par juge, les numéros de produit. En catégorisation, le nombre de classes choisi par juge est conservé mais la structure sur les produits est complètement cassée. Ainsi, dans l'analyse globale des données, aucune structure ne devrait se dégager d'un tel jeu de données.

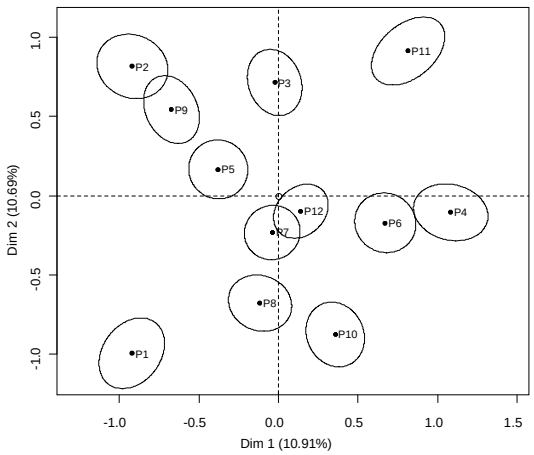


Figure 4 : Ellipses de confiance construites sur un jeu de données de catégorisation fictif et non structuré de 12 produits décrits par 98 juges.

La figure 4 correspond aux ellipses de confiance obtenue avec la procédure proposée par Cadoret et al. (2009) pour un jeu de données non structuré. Cette figure montre des ellipses de confiance de petites tailles et souvent disjointes. Il semble donc que la variabilité autour de la position des produits soit

largement sous-estimée. La notion de zone de confiance sous-entendue par les ellipses conduit alors à interpréter cette représentation de façon erronée. Ainsi, les produits 1 et 10, par exemple, seraient interprétés comme différents de l'ensemble des autres produits d'un point de vue sensoriel alors que la construction même du jeu de données nous assure que le jeu de données n'est pas structuré (nous avons réalisé plusieurs perturbations du jeu de données et à chaque fois nous obtenons le même type de graphique avec des ellipses petites et disjointes).

On peut se demander pourquoi la configuration des produits semble aussi stable quand les données ne sont pas structurées. Le tableau de données sur lequel l'ACM est construite contient 12 lignes (les 12 produits) et 98 colonnes (les 98 juges). Ces 98 juges ont utilisé au total 453 catégories et donc l'ACM revient à faire une AFC sur un tableau disjonctif de 12 lignes et 453 colonnes. Les produits sont donc dans un espace de très grande dimension (ici de dimension 11 car il y a 12 lignes) ; l'objectif de l'analyse factorielle est de fournir un sous-espace de dimension restreinte (souvent en deux dimensions) qui met en évidence une structure sur les produits, c'est-à-dire en ACM un plan de projection qui maximise la variabilité des points (i.e. des produits) projetés, et donc un plan qui permet de bien différencier les produits. Cette configuration maximise aussi la variabilité des modalités de chaque variable qualitative, à savoir ici les classes d'appartenance des produits proposées par chaque juge. Donc, même s'il n'y a pas de structure sur les données, l'ACM fournit une représentation des produits qui les sépare au mieux et qui sépare au mieux les classes auxquelles ces produits appartiennent. D'après les propriétés barycentriques de l'ACM, un produit est au barycentre des classes auxquelles il appartient (une classe par juge) et toutes les classes auxquelles il appartient sont dans la même région du sous-espace. La figure 5 illustre cette situation car toutes les positions du produit P1 vu par chacun des juges sont dans la partie inférieure gauche du graphique. Les ellipses de confiance sont alors construites en calculant le barycentre de nombreux points qui sont dans la même région du sous-espace que le produit vu par le vrai jury. Comme le nombre de juges est souvent important en catégorisation, le barycentre est calculé à partir de beaucoup de points et la bonne propriété de la moyenne (théorème central limite) conduit à une très bonne stabilité du barycentre.

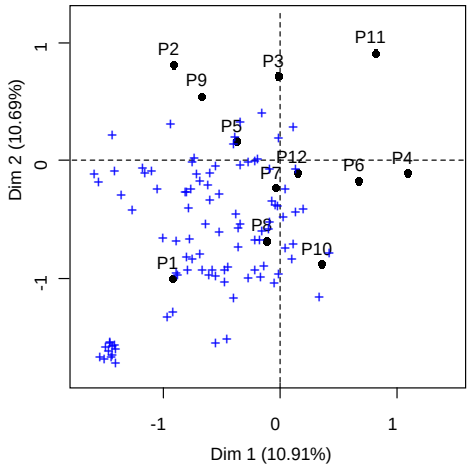


Figure 5 : Représentation de la position du produit P1 vu par chacun des juges pour des données non structurées.

Le rééchantillonnage des juges permet de perturber la position du barycentre des classes mais en s'appuyant sur l'analyse initiale qui a positionné les classes d'appartenance d'un même produit dans une même région du plan. Le bootstrap utilisé ici est donc un bootstrap partiel selon la terminologie de

Lebart (2007). Ce bootstrap ne perturbe pas le sous-espace obtenu par ACM car les axes principaux calculés sur les données originales non perturbées jouent un rôle privilégié. C'est pour cette raison que nous proposons d'utiliser un bootstrap total de type 3 qui va permettre de valider globalement le sous-espace engendré par les axes principaux de l'analyse factorielle.

3. Méthodologie pour construire des ellipses de confiance dans les approches holistiques

3.1 Utilisation du bootstrap total pour construire des ellipses de confiance en catégorisation

Le principe de construction de jurys virtuels reste parfaitement adapté aux méthodes holistiques. Cependant, il ne faut pas projeter l'ensemble de la configuration obtenue par le jury virtuel sur la configuration obtenue par le vrai jury car sinon on retrouve la même méthode proposée par Cadoret *et al*. (2009). Il faut déterminer la configuration obtenue par le jury virtuel et comparer cette configuration à celle du vrai jury. C'est ce que propose le bootstrap total (Château et Lebart, 1996, Lebart, 2007) qui consiste à bootstrapper les individus statistiques puis à refaire une analyse complète pour chaque réplication et enfin à concaténer les résultats des échantillons bootstraps. Dans le cas de la catégorisation, cela revient à construire un échantillon bootstrap (un jury virtuel), à faire l'ACM, à récupérer la configuration (en 2 dimensions) des produits obtenue par ACM et à positionner cette configuration sur la configuration obtenue par le vrai jury. Pour positionner la configuration d'un jury virtuel sur la configuration du vrai jury, l'analyse procrustéenne (voir, par exemple, Krzanowski, 2000) est tout à fait adaptée. Cette procédure nécessite de déterminer la dimensionnalité des configurations du vraie jury et des jurys virtuels avant de faire l'analyse procrustéenne. L'estimation du nombre de dimensions de la configuration des produits est difficile. Josse et Husson (2012) ont proposé une approximation d'un critère de validation croisée dans le cadre de l'ACP. Ce critère peut être utilisé sur une matrice pondérée par les poids d'une analyse factorielle telle que l'ACM ou l'AFM. Si le nombre de dimensions est égal à 2, ce qui est assez fréquent, alors la configuration des produits du jury virtuel (le plan principal de l'ACM) subit une rotation procrustéenne sur la configuration du vrai jury. Pour que la rotation procrustéenne puisse être effectuée, au minimum deux dimensions sont utilisées même si l'estimation du nombre de composantes principales est égal à 1.

Le graphique de gauche de la figure 6 donne les ellipses de confiance obtenues sur le jeu de données non structuré précédent. Les ellipses de confiance se chevauchent très largement ce qui est en accord avec l'interprétation suivante des résultats : les produits ne peuvent pas être considérés comme différents d'un point de vue sensoriel. On peut noter que si on utilise cette méthodologie avec toutes les dimensions de l'ACM plutôt que 2 dimensions, alors les ellipses de confiance sont très petites et le graphique ressemble au graphique obtenu figure 4. Ceci montre que le choix de la dimensionnalité est cruciale et que c'est bien le bootstrap total qui permet d'obtenir des ellipses de taille raisonnable (en effet, si toutes les dimensions de l'ACM sont utilisées, cela revient à ne pas faire d'ACM sur le jury virtuel).

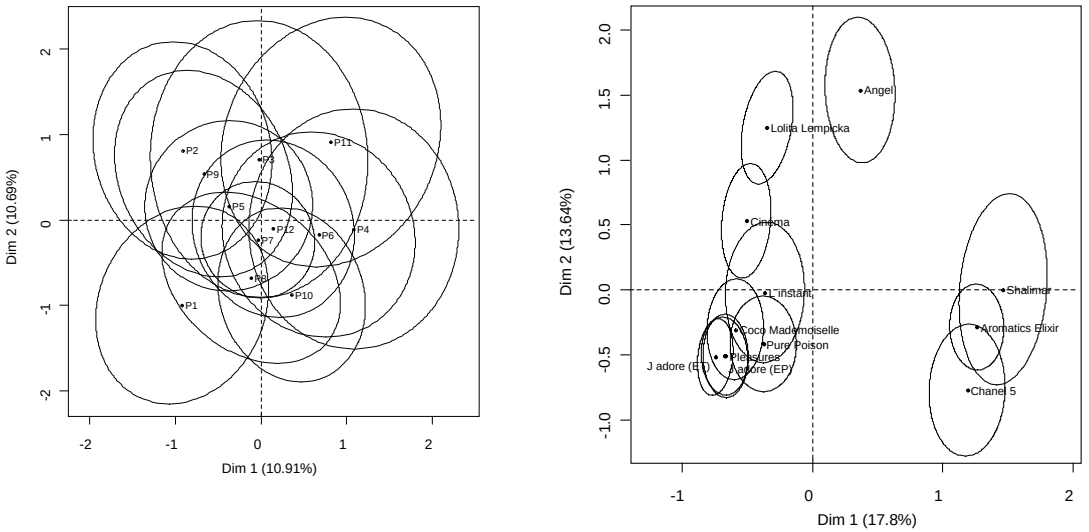


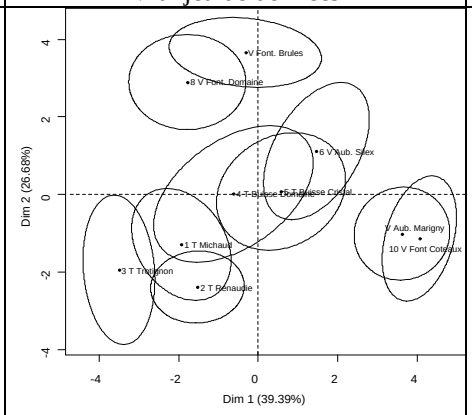
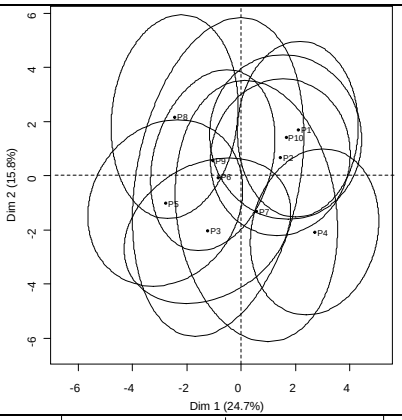
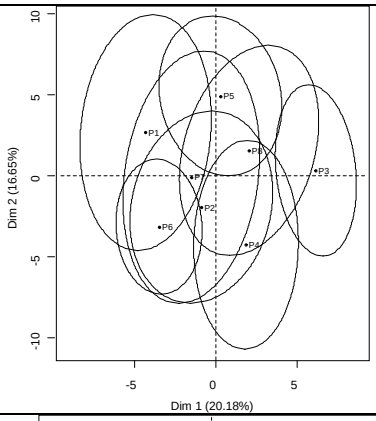
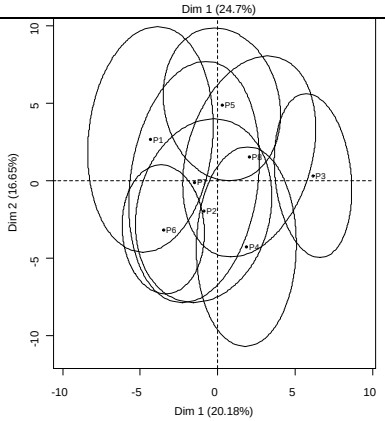
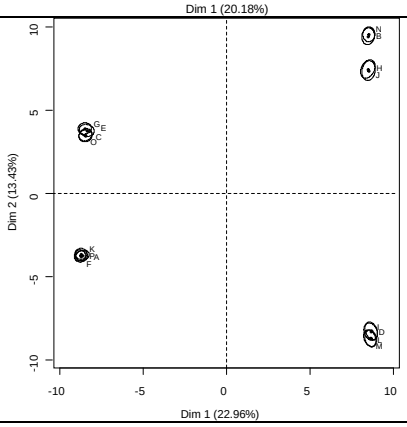
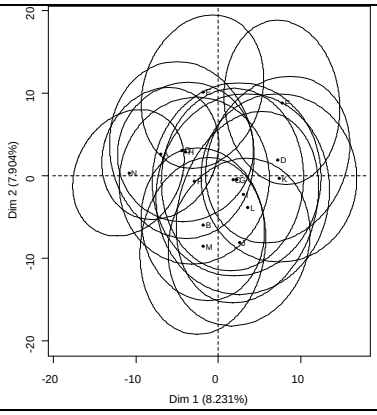
Figure 6 : Ellipses de confiance construites avec le bootstrap total sur un jeu de données de catégorisation de 12 produits décrits par 98 juges ; le graphique de gauche concerne un jeu de données fictif et non structuré, le graphique de droite concerne le jeu de données parfum de Cadoret *et al*. (2009).

Si on applique maintenant la méthodologie du bootstrap total sur un jeu de données structuré, alors certaines ellipses de confiance sont bien disjointes ce qui montre que certains produits sont perçus comme différents d'un point de vue sensoriel tandis que d'autres sont très proches (voir figure 6, graphique de droite).

3.2 Construction d'ellipses pour les autres approches holistiques

Cette stratégie du bootstrap total peut se décliner facilement aux diverses méthodes holistiques telles que le napping, le napping catégorisé, le tri hiérarchique, etc. Le principe est donc de construire des jurys virtuels en choisissant au hasard et avec remise des juges, puis de lancer l'analyse sur le jeu de données virtuel, et ensuite d'effectuer une rotation procrustéenne du sous-espace obtenu à partir du jeu de données virtuel sur le sous-espace obtenu à partir du jeu de données original. La figure 8 montre, pour différentes méthodes de recueil de données, des ellipses de confiance pour un vrai jeu de données et pour un jeu de données fictif non structuré obtenu par permutation des lignes de chaque juge.

L'exemple de napping correspond à un jeu de données dans lequel 10 vins sont positionnés par 11 juges sur une nappe de 60 x 40cm (Pagès, 2005). L'exemple de napping catégorisé correspond au jeu de données smoothies de Pagès et al. (2010) dans lequel 8 smoothies sont décrits par 24 juges à la fois grâce à un positionnement des produits sur des nappes 60x40cm et à un regroupement des produits à l'issue de l'épreuve de napping. Le jeu de données de tri hiérarchique correspond à un jeu de données sur des cartes pour enfants (Cadoret *et al.*, 2011) dans lequel 16 cartes ont été triées par 89 enfants. Enfin, le jeu de données de profil libre correspond à une description de 12 parfums par 6 juges (Gazano *et al.*, 2005).

	Vrai jeu de données	Jeu de données fictif non structuré
Napping		
Napping catégorisé		
Tri hiérarchique		

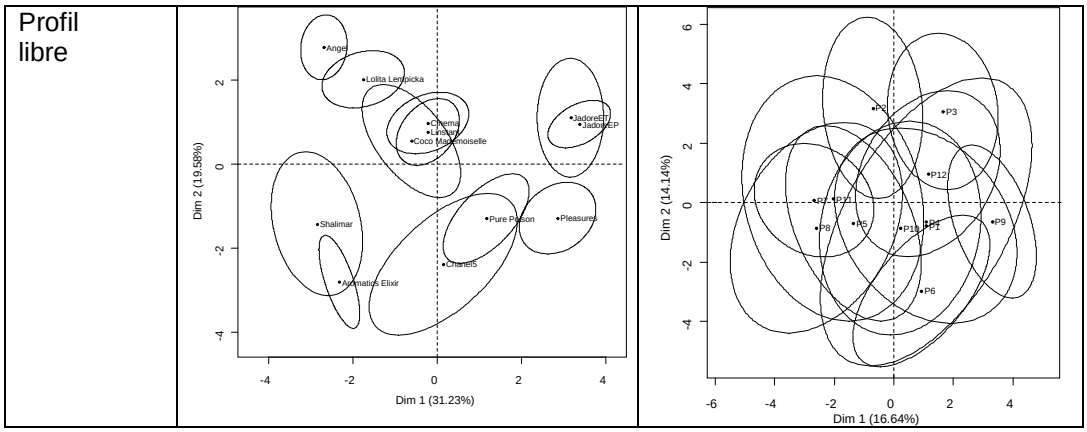


Figure 8 : Ellipses de confiance construites sur des jeux de données réels (colonne de gauche) et des jeux de données non structurés (colonne de droite) pour différentes méthodes d'évaluation sensorielle (napping, napping catégorisé, tri hiérarchique, profil libre).

Pour les jeux de données non structurés, les ellipses de confiance ne permettent pas, comme attendu, de mettre en évidence des différences sensorielles entre produits tandis que les ellipses de confiance obtenues sur les jeux de données réels permettent de mettre en évidence des différences sensorielles entre certains produits.

3. Conclusion

La méthodologie proposée ici est prometteuse car elle offre la possibilité de construire des ellipses de confiance à partir de données obtenues par de nombreuses méthodes de description sensorielles et notamment les méthodes holistiques. Les configurations moyennes des produits obtenues par des approches holistiques peuvent donc être interprétées avec beaucoup plus de sérénité puisque la stabilité de la configuration peut être visualisée directement sur le graphique. Nous avons décrit l'utilisation de cette procédure pour des méthodes factorielles mais cette procédure est également adaptée à d'autres méthodes de positionnement multidimensionnel comme DISTATIS par exemple.

Le choix du nombre de dimensions pour effectuer la rotation procrustéenne reste un choix délicat car aucune méthode éprouvée ne permet d'obtenir avec certitude le bon nombre de dimensions sous-jacentes dans le jeu de données. En pratique, il semble que le nombre de dimensions des configurations moyennes soit souvent de 2 pour ce type d'évaluation sensorielle.

La méthodologie proposée ici est disponible dans la librairie de fonction SensoMineR (Husson et Lê, 2006, Lê et Husson, 2008) du logiciel R grâce à la fonction *boot*. Cette fonction permet de construire des ellipses de confiance pour des données de catégorisation, des données de napping, de napping catégorisé, pour du tri hiérarchique ou encore pour des données de profil libre.

Bibliographie

Abdi, H., Valentin, D., Chollet, S. & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, 18, 627–640.

Cadoret M., Lê S., Pagès J. (2009). A Factorial Approach for Sorting Task data (FAST). *Food Quality and Preference.* 20, 410-417.

Cadoret M., Lê S., Pagès J. (2011). Statistical analysis of hierarchical sorting data. *Journal of Sensory Studies*. 26, 96-105.

Chateau, F. & Lebart, L. (1996). Assessing Sample Variability in the Visualization Techniques related to Principal Component Analysis : Bootstrap and Alternative Simulation Methods. Proceedings de COMPSTAT, Physica Verlag, Heidelberg, Alberto Prats, Editor.

Efron, B. (1979). Boostrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Gazano, G., Ballay, S., Eladan, N., Sieffermann, J.M. (2005). Flash Profile and fragrance research:using the words of the naive consumers to better grasp the perfume's universe In: ESOMAR Fragrance Research Conference, 15-17 May 2005, New York.

Husson, F., Le Dien, S. & Pagès, J. (2005). Confidence ellipse for the sensory profiles obtained by Principal Component Analysis. *Food Quality and Preference*, 16 (3), 245-250.

Husson, F. & Lê, S. (2006). SensoMineR : un package pour le traitement de données sensorielles avec R. *Sciences des aliments.* 26, 355-356.

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis; application to the study of ten white from the Loire Valley. *Food quality and preference*, 16, 642-649.

Pagès, J., Cadoret, M., Lê, S. (2010) The Sorted Napping: a new holistic approach in sensory evaluation. *Journal of Sensory Studies*, 25, 637-658.

Pagès, J. & Husson, F. (2005). Multiple Factor Analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data. *Journal of chemometrics*, 19, 138-144.

Josse, J. & Husson, F. (submitted). Selecting the number of components in PCA using cross-validation approximations

Krzanowski, W.J. (2000). *Principles of multivariate analysis; a user's perspective*, Clarendon Press, Oxford.

Lê, S. & Husson, F. (2008). SensoMineR: a package for sensory data analysis. *Journal of Sensory Studies*. 23 (1), 14-25.

Lebart, L. (2007). Which Bootstrap for Principal Axes Methods? *Selected Contributions in Data Analysis and Classification.* 581-588. Springer Berlin Heidelberg,

Saporta, G & Hatabian, G. (1986) Régions de confiance en analyse factorielle. *Data analysis and informatics* (499-508). Elsevier. Amsterdam.

Session 4 : Chimiométrie I / *Chemometrics I*

Multi-block regression based on combinations of orthogonalisation and PLS.

Tormod Næs*+, Oliver Tomic*, Nils-Christian Afseth*, Vegard Segtnan*, Ingrid Måge*

* Nofima, Oslovegen 1, 1430 Ås, Norway
+ Dept. of Food Science, University of Copenhagen, Faculty of Life Sciences.

Abstract

Exploring relationships between several large data blocks is important in many areas in modern science. Typical applications can be found in industrial process modelling, in investigations of various -omics data sets and in consumer science. In some cases one is interested in understanding the relation between the data sets without any particular ordering of them while in other cases one is interested in studying blocks of data with a predictive direction. The latter will be in focus here.

Using regular least squares (LS) regression analysis for this type of data is often impossible because of strong collinearities. In addition, one is also typically interested in exploring relations among the variables within each of the blocks and how the different blocks contribute to the regression model. In order to handle these problems, one needs regression methodology based on data compression that both solves collinearity problems and that can be used for graphical visualisation for improved interpretation of the results. The multi-block PLS regression method is an important method which has these properties. It is essentially a concatenated PLS approach, but provides additional plotting tools for understanding the different blocks as well as their joint contribution to prediction.

The SO-PLS (sequential and orthogonalised partial least squares, Jørgensen et al(2007)) and PO-PLS (parallel and othogonalised PLS, Måge et al(2008)) methods have recently been proposed as alternatives to regular multi-block PLS regression. These methods are, as opposed to standard MB-PLS regression, invariant with respect to the relative weighting of the blocks. This can be an important aspect if the data blocks represent different sources of information with very different measurement units. Secondly, these methods handle explicitly situations with very different dimensionality of the blocks. It can for instance easily be used for situations where one of the blocks is a design matrix and the other ones are highly collinear. Both these methods are based on sequential use of PLS regression and orthogonalisation. The methods are closer to classical statistical methods such as regression and ANOVA than standard MB-PLS regression and can therefore be considered as bridges between classical statistical approaches and more chemometric strategies.

This talk will give an overview of the PO-PLS and SO-PLS methods with focus on ideas and the relationships between them and with other related methods. The incorporation of interactions as well as interpretation of the result will be given attention. A data set from NIR and Raman spectroscopy will be used for illustration.

References

Jørgensen, K. Mevik, B-H. and Næs, T. (2007). Combining designed experiments with several blocks of spectroscopic data. *Chemometrics and Intelligent laboratory Systems*, *88*(2), 143-212.
Måge,I. Mevik, B-H. and Næs, T. (2008). Regression models with process variables and parallel blocks of raw material measurements. *J. Chemometrics, 22, 443-456.*

Sélection conjointe de régions de spectres MIRS et RAMAN et de variables en régression PLS à l'aide d'Algorithmes Génétiques

Joint selection of wavelength regions for MIRS and RAMAN spectra and variables in PLS regression using Genetic Algorithms

Lidwine Grosmaire[1], Christelle Reynès[2] & Robert Sabatier[2]

[1] *Laboratoire de Physique Moléculaire et Structurale - UMR Qualisud - Université Montpellier 1 - France*
[2] *Laboratoire de Physique Industrielle et Traitement de l'Information - EA 2415 - Université Montpellier 1 - France*

E-mail : *lidwine.grosmaire@univ-montp1.fr*

Résumé
De nombreuses méthodes adaptées pour la regression PLS, s'intéressent aux choix de variables explicatives, quand celles-ci sont en nombre trop important. Quand il s'agit de sélectionner des intervalles pour des spectres, la panoplie des techniques est plus réduite. L'origine de ce travail est une problématique de régression pour des données sur la transformation de manioc. Ces données sont constituées de trois tableaux : des spectres RAMAN, MIR et des variables physico-chimiques. Il s'agit d'adapter au contexte de régression une stratégie précédemment mise au point pour la sélections d'intervalles uniquement pour des spectres NIR en discrimination. Nous avons développé un algorithme génétique spécialement adapté à ce type de données (multitableau), pour le cas de la régression PLS1.

Mots-clés : Méthode PLS, Algorithme Génétique, Spectres MIR et RAMAN, choix de variables, sélection d'intervalles.

Abstract
Many methods exist for feature selection in PLS regression when there are too many variables. Less methods are available for selecting wavelength regions for MIR or RAMAN spectra. In this work, PLS has been coupled with genetic algorithms to allow the selection of intervals in spectra. Our application goal is to be able to perform feature selection among variables from different origins : MIRS spectra, RAMAN spectra and physico-chemical variables. A new algorithm is proposed to adapt to such multiway data in PLS1 regression context. An illustration on real data is given.
Keywords : PLS regression, Genetic Algorithm, MIR and RAMAN spectra, Choosing variables, selection of wavelength regions.

1 Introduction

En chimiométrie, le choix de variables explicatives est un problème souvent abordé dans le cas particulier de la régression PLS. Lorsqu'il s'agit de sélectionner des intervalles de longueur d'onde choisis dans des spectres MIR (Moyen Infrarouge), les méthodologies sont plus rares, plus complexes (statistiquement parlant) et plus *coûteuses* en temps calcul. Les Algorithmes Génétiques (AG) ont quelque fois été utilisés avec la méthode PLS pour sélectionner des intervalles, voir Leardi (2000) et Leardi et Norgard (2004). Quelques autres méthodes spécifiques (sans AG), ont été mises au point comme la iPLS de Norgard *et al.* (2000). L'article de Hoskuldsson (2001) fait un état des lieux assez général pour le choix d'intervalles dans ce cadre.
Les données à l'origine du travail présenté s'inscrivent dans le contexte de la production et transformation de manioc. L'augmentation régulière de la production et de la consommation de ce produit témoigne de son importance économique grandissante dans le monde et plus particulièrement dans les régions tropicales (Tonukari, 2004). Néanmoins, la production et la transformation du manioc sont à l'heure actuelle le fait de petites exploitations peu rentables. Par conséquent, les études sur ce tubercule, en particulier dans un but industriel, semblent à ce jour essentielles. L'objectif de ce travail est d'essayer d'expliquer la capacité de panification à partir des différents paramètres étudiés.
Dans cet article, nous allons utiliser et adapter un AG, qui a été mis au point dans un contexte de discrimination (LDA usuelle) dans le cas où l'on désire choisir, pour variables explicatives, des intervalles de longueurs d'onde (voir Reynès *et al.* (2006)). Mais, pour le cas pratique qui nous a été soumis, les données consistent en un multitableau composé de trois tableaux (mesuré sur les mêmes $n = 52$ observations) pour lequel il faut choisir des intervalles de longueurs d'onde pour les deux premiers (composés de spectres MIR et RAMAN) et des variables physico-chimiques pour le dernier, dans le but de prédire une seule variable d'intérêt.

2 Les données

Les variétés de manioc qui ont retenu notre attention proviennent de Colombie. L'amidon, extrait des farines, est utilisé dans ce pays après un procédé empirique de fermentation naturelle et de séchage au soleil qui confère au produit fini des propriétés de panification très intéressantes. Afin de déterminer l'impact variétal et l'impact procédé sur les propriétés de panification, nous avons sélectionné 13 variétés : 10 cultivées en altitude (1800 m) et 3 cultivées en plaine (1000 m) et 4 procédés de traitement contrôlé : non fermenté séché au four (NFO), fermenté séché au four (FO), non fermenté séché au soleil (NFS) et fermenté séché au soleil (FSR).
Les grains d'amidon sont constitués de deux macromolécules : l'amylose (structure linéaire) et l'amylopectine (structure ramifiée). En suspension dans l'eau, les grains d'amidon chauffés développent une certaine viscosité dans le milieu (Thomas *et al.*, 1998). Le Rapid Visco Analyser (RVA) permet de suivre l'évolution de la viscosité de l'amidon au cours d'un protocole de température établi (chauffage, maintien en température, refroidissement). Lorsque la suspension d'amidon est chauffée, une température caractéristique est atteinte (température de gélatinisation) : l'eau pénètre dans les grains d'amidon qui gonflent. Ce phénomène entraine l'augmentation de la viscosité jusqu'à un maximum suivi d'une diminution s'expliquant par la perte de structure granulaire : les macromolécules sortent des grains pour se solubiliser à l'extérieur. Lorsque la solution est refroidie, sa viscosité augmente à nouveau consécutivement à la réassociation des macromolécules : la rétrogradation. Les profils RVA obtenus donnent accès à 12 paramètres de temps, de température et de viscosité au cours de la gélatinisation et

de la rétrogradation.
Des spectres IR et Raman ont été respectivement enregistrés dans les domaines 650-4000 cm-1 (3351 variables) et 230-3800 cm-1 (4562 variables). Afin d'homogénéiser le signal et d'homogénéiser les données, les spectres bruts ont nécessité un prétraitement de correction de ligne de base et de normalisation vectorielle (méthode SNV) à l'aide du logiciel LabSpec 5.
Au final, on obtient 52 échantillons (13 variétés x 4 traitements) qui ont été analysés afin de déterminer ces caractéristiques chimiques et physico-chimiques comme la capacité de panification (représentant l'expansion de la pâte au cours de la cuisson), le pourcentage d'amylose (composé de l'amidon avec l'amylopectine), la viscosité de l'amidon au cours de la gélatinisation (RVA) ainsi que les spectres MIR et Raman.

3 Rappels à propos des Algorithmes Génétiques

Un AG est une méthode d'optimisation numérique, une heuristique, introduite par Holland en 1975¡. Son processus est inspiré de la sélection naturelle. L'AG part d'un ensemble T_{pop} de solutions initiales possibles (appelées *individus*) et fait évoluer cette population de façon à optimiser un critère appelé *fitness*. L'évolution de la population se fait, à taille constante, en utilisant trois *opérateurs* qui vont être appliqués successivement à chaque individu, de façon probabiliste. Cette méthododologie permet de trouver des solutions à des problèmes d'optimisations complexes (nombre de paramètres à optimiser important et/ou échec des procédures usuelles). La convergence théorique a été montrée, mais en pratique, elle peut être plus délicate à obtenir, et en général on impose un nombre maximal d'itérations N_{gene}.
Un individu d'une population est une solution potentielle au problème posé, codé sous la forme d'un vecteur numérique (de longueur finie, fixée *a priori* pour chaque problème) dont chaque coordonnée (ou plusieurs selon le codage du problème) est la valeur d'un paramètre à optimiser. Les trois opérateurs à appliquer sont le *croisement*, la *mutation* et la *sélection*. Le croisement (ou *cross-over*) consiste en un échange aléatoire des caractéristiques de deux individus, réalisé avec une probabilité π_c donnée *a priori* par l'utilisateur. La mutation est une modification aléatoire de quelques-unes des caractéristiques d'un individu, réalisée avec une probabilité π_m donnée *a priori* par l'utilisateur. La sélection favorise la survie des individus intéressants du point de vue de la fitness. La sélection est le seul opérateur qui utilise la fitness (c'est-à-dire dépendant du problème à optimiser). Le règle générale de sélection est que plus un individu est adapté au sens de la fitness plus sa probabilité d'apparaître dans la génération suivante augmente, mais tout individu (quelle que soit la valeur de sa fitness) a une probabilité non nulle d'apparaître dans la population suivante. Ces opérateurs ont comme but de maintenir au maximun l'hétérogénéité de la population et d'assurer une évolution vers une meilleure population de solutions (au sens de la fitness).

4 La nouvelle méthodologie pour un multitableau

La nouvelle méthode proposée, consiste à utiliser conjointement d'une part la méthode PLS1 pour un multitableau (ou multibloc) $X = [X_1, X_2, X_3]$ (où X_1 représente le tableau des variables MIR, X_2 celui des variables RAMAN et enfin X_3 les variables physico-chimiques) et d'autre part un AG spécifique pour chaque tableau (ou bloc). La variable à prédire, y, ainsi que le multitableau X sont mesurés sur les mêmes $n = 52$ observations.
Par un unique souci de simplification des notations, nous n'utiliserons que $K = 3$ blocs dans

le multitableau et la méthode PLS1, les organisations plus générales ($K > 3$ et PLS2) des structures de données ne modifient en rien la présentation de cette méthode.
Cette procédure sera appelée dans la suite **AGvPLSm** (Algoritme Génétique pour la sélection de variables pour PLS multitableau).

4.1 La méthode AGvPLSm

L'analyse proprement dite, consiste donc, une fois choisis les intervalles de longueur d'onde, pour les MIR et le RAMAN, ainsi que les variables physico-chimiques (par utilisation de l'AG), à réaliser la PLS1 en validation croisée de y par $\tilde{X} = \left[\tilde{X_1}, \tilde{X_2}, \tilde{X_3}\right]$, où $\tilde{X_1}$ représente le sous-tableau des variables MIR retenues, $\tilde{X_2}$ celui des variables RAMAN retenues et $\tilde{X_3}$ les variables physico-chimiques retenues.
Enfin, le choix du nombre de composantes PLS, A, se détermine automatiquement par validation croisée, à l'intérieur de l'agorithme.

4.2 La fitness pour la sélection d'intervalles de longueur d'onde et la sélection de variables

La fitness, pour le multitableau X, va se définir par

$$fitness = cor^2(y, \hat{y}) + c \times (\alpha(N_{varsel} + \beta))$$

Dans la première partie de cette équation *cor* est la corrélation entre la variable y et sa modélisation par $\tilde{X}$ à l'aide de PLS1 en validation croisée. La deuxième partie de la fitness fait intervenir le nombre de variables sélectionnées N_{varsel} (dans X_3). Le coefficient c intervient pour pondérer les deux parties de la fitness et les coefficients α et β permettent de ramener le deuxième terme dans l'intervalle $[0, 1]$.
Pour choisir des intervalles de longueur d'onde qui aient un sens, au cours de l'évolution dans l'AG, nous imposons une longueur minimale de celui-ci *lmininter*, ainsi que la distance minimale *dmininter* entre deux intervalles (sinon il y a fusion).

4.3 Utilisation de AGvPLSm

Pour utiliser AGvPLSm, il y a un certain nombre de paramètres à choisir : T_{pop}, N_{gene}, π_c, π_m, sont les paramètres proprements dits de l'AG et sont ajustés dans des essais préalables. Les autres, N_{varsel}, *lmaxinter* et *dmininter* sont moins primordiaux, mais sont plus liés aux données traitées. L'algorithme a été programmé sous le logiciel R.

5 Application

Les données décrites au paragraphe 2 ont été traitées par la méthode AGvPLSm avec les paramètres suivants : $T_{pop} = 200$, $N_{gene} = 100$, $\pi_m = 0.9$, $\pi_c = 0.5$ en sélectionnant des variables individuelles dans $\mathbf{X}_1$, la matrice des variables physico-chimiques et en sélectionnant des intervalles dans $\mathbf{X}_2$, la matrice des spectres RAMAN et $\mathbf{X}_3$, la matrice des spectres MIR. Les résultats ont été obtenus à partir de 10 *runs* de l'algorithme.
Sur la meilleure solution de chacune des 10 populations finales, on obtient un R^2 moyen en validation croisée de 0.885 avec un écart-type de 0.017. En moyenne, 2.7 variables physico-chimiques ont été sélectionnées, ainsi que 2.8 intervalles de MIR (pour un nombre total moyen

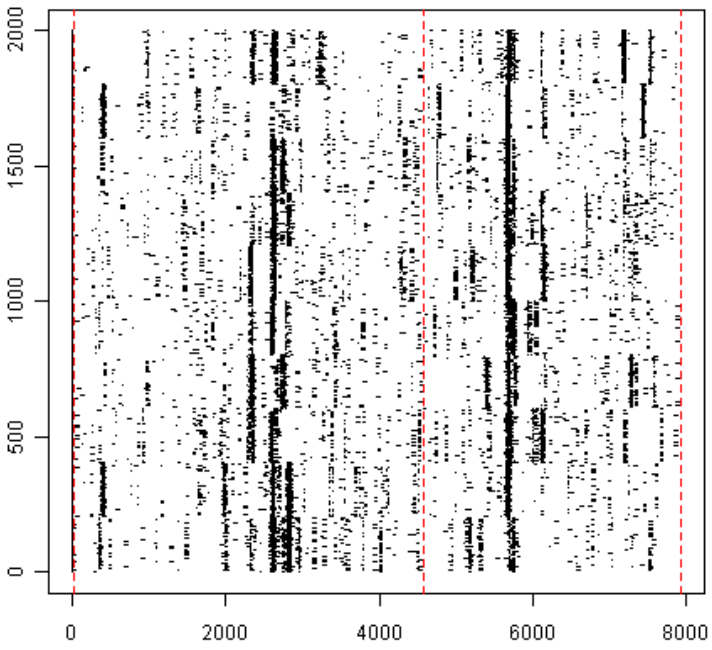


Figure 1: Représentation des variables sélectionnées (représentées par un point) dans les populations finales des 10 runs (c'est-à-dire pour 2000 solutions en ordonnée). En abscisse, on a noté le numéro des variables, pour les trois tableaux, c'est-à-dire de 1 à 7926. Les tableaux sont séparés par des traits pointillés.

de 46.3 longueurs d'onde) et 1.8 intervalles de RAMAN (pour un nombre total moyen de 37.6 longueurs d'onde).
Si on étudie l'ensemble des variables sélectionnées dans les populations finales, on obtient le graphique de la Fig.1. D'après ce graphique, on constate une sélection préférentielle dans quelques zones qui confirme une convergence des algorithmes malgré un certain nombre de solutions équivalentes. Finalement, on retient 4 variables physico-chimiques, 4 intervalles de RAMAN et 2 intervalles de MIR, ce qui correspond à 311 variables, c'est-à-dire environ 4% des variables de départ. A des fins de comparaison, un modèle a également été construit sur l'ensemble des 7926 variables ainsi que sur les variables sélectionnées par la méthode VIP (Wold *et al.*, 1993), cela correspond à 4 variables physico-chimiques. L'ensemble des résultats est donné dans la Tab.1.
On constate que le modèle obtenu sur l'ensemble des variables est pénalisé par le *bruit* apporté par un trop grand nombre de variables inutiles et/ou corrélées. La sélection VIP ne fait ressortir que des variables physico-chimiques, en petit nombre. Le modèle obtenu est assez performant. Il semble ainsi que les variables physico-chimiques jouent un rôle très important dans la modélisation (modèle quasiment équivalent à l'utilisation de l'ensemble des variables). Cependant, elles ne sont pas suffisantes et les variables de spectrométrie ont un effet non négligeable puisque le modèle proposé par AGvPLSm explique 16% de variabilité en plus.

Méthode	nb var	A	R^2	R^2_{CV}
AGvPLSm	311	12	0.9936	0.8273
PLS + VIP	4	3	0.7210	0.6650
PLS	7926	7	0.7836	0.6605

Table 1: Résultats obtenus par trois méthodes pour les données de manioc (nb var : nombre de variables utilisées par le modèle, A : nombre de composantes optimisé par le PRESS, R^2 : R^2 obtenu sur les données d'apprentissage, R^2_{CV} : R^2 moyen obtenu en 10-FCV).

En terme d'interprétation de ces résultats, les variables physico-chimiques sélectionnées sont issues des données RVA : (i) Breakdown et Relative Breakdown qui rendent compte de la taille des macromolécules ramifiées (amylopectine), (ii) Peak Viscosity qui est relié à la taille des grains d'amidon et (iii) Holding Strength qui représente la viscosité minimale de la pate. Ces paramètres sont liés à la capacité d'absorption d'eau de l'amidon et sont, par conséquent, des paramètres texturaux essentiels pour décrire la capacité de panification. Pour ce qui est des variables spectrométriques, les échantillons analysés ayant des compositions chimiques très proches (15.7 à 21.7 % en amylose), les spectres de vibration enregistrés ne montrent pas de différences notables. Cependant, les domaines spectraux sélectionnés par AGvPLSm, que ce soit à haute ou à basse fréquence, ne correspondent pas à des bandes de vibration spécifiques ou à des combinaisons de bandes, mais contiennent des informations pertinentes puisqu'elles permettent d'améliorer significativement le résultat du modèle.

6 Conclusion

La sélection de variables est un problème récurrent en analyse de données. Elle permet notamment de fournir des modèles plus performants et plus interprétables aboutissant à une meilleure compréhension des phénomènes étudiés. Quand on dispose de milliers de variables, une exploration exhaustive des combinaisons de variables est irréalisable de par l'importante combinatoire et une approche pas à pas n'est pas adaptée pour fournir un modèle optimal. De plus, les approches classiques (y compris VIP) ne permettent pas de sélectionner des intervalles qui sont les seules entités ayant un sens en spectrométrie. Une méthode heuristique est donc nécessaire. Les algorithmes génétiques fournissent une solution très adaptable et efficace pour ce type de problème où l'on doit combiner plusieurs types de sélection (individuelle vs intervalles) dans un contexte de multi-tableaux. On peut noter que la généralisation à PLS2 est aisée en utilisant, par exemple, le RV d'Escoufier à la place du R^2.
Pour l'application qui a motivé ce travail, le résultat obtenu est très intéressant pour une utilisation prédictive. En terme d'interprétation, la méthode a permis de mettre en évidence l'importance prépondérante de certaines variables physico-chimiques et de choisir un faible nombre d'intervalles assez courts qui complètent significativement le modèle obtenu. Ces intervalles sont difficiles à interpréter pour les chimistes et ne participent donc pas à la compréhension du phénomène mais sont indispensables à la qualité du modèle final.

Bibliographie

Depczinski, U., Frost, V. J., Molt, K., (2000) Genetic algorithms applied to the selection of factors in principal components regression. *Analytica Chimica Acta*, 420, 217-227.

Goicoechea, H. C., Olivieri, A. C., (2003) A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. *Journal of Chemometrics*, 17, 338-345.

Holland, J. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor, MI.

Höskuldsson, A. (2001) Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55, 23-38.

Leardi, R. (2001) Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics*, 15, 559-569.

Leardi R, Norgaard L. (2004) Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Chemometrics and Intelligent laboratory Systems*, 18, 486-497.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B. (2000) Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54, 3, 413-419.

Reynes, C., De Souza, S., Sabatier, R. (2006) Selection of discriminant wavelength intervals in NIR spectrometry with genetic algorithms. *Journal of Chemometrics*, 20, 136-145.

Thomas, D.J. & Atwell, W.A. (1998) Starches. *American Association of Cereal Chemists.*

Tonukari, N.J. (2004) Cassava and the future of starch. *Journal of Biotechnology*, 7(1).

Wold, S., Johansson, E., Cocchi, M. (1993) *3D QSAR in Drug Design: Theory, Methods, and Applications*, ESCOM, Leiden, Holland, pp. 523-550.

Détection de l'addition d'orge au café en utilisant la spectroscopie dans le proche infrarouge et techniques de chimiométrie

Detection of addition of barley to coffee using near infrared spectroscopy and chemometric techniques

Heshmatollah Ebrahimi-Najafabadi[1,2], Riccardo Leardi[2], Paolo Oliveri[2], Chiara Casolino[2], Mehdi Jalali-Heravi[1], & Silvia Lanteri[2]

[1] *Department of Chemistry, Sharif University of Technology, P.O. Box 11155-9516, Tehran, Iran*
E-mail : *ebrahimi.heshmat@gmail.com*

2 *Department of Pharmaceutical and Food Chemistry and Technology, University of Genova, Via Brigata Salerno 13, I-16147 Genova, Italy*
E-mail : *riclea@dictfa.unige.it*

Résumé
L'étude en question présente une application de spectroscopie dans le proche infrarouge pour l'identification et la quantification de l'adddition frauduleuse d'orge à échantillons de café. Neuf différents types de café ont été mélangés avec quatre types d'orge, dans le domaine 2-20% en poids d'orge. Les 100 expériments du training set et le 30 expériments du test set ont été sélectionnés en utilisant un D-optimal design. Partial Least Squares regression (PLS) a été utilisée pour prédire la quantité d'orge dans le café. Pour obténir des modèles simplifiés, ne contenant que les régions spéctrales informatives, on a utilisé un Algorithme Génétique (GA). Les performances des modèles ont été vérifiées même en utilisant un jeu de données complètement indépendant. Les modèles ont montré une très bonne predictivité, avec une erreur quadratique moyenne (RMSE) de 1.4% et 0.8% en poids pour le test set et pour le jeu indépendant.

Mots-clés: café, orge, spectroscopie dan le proche infrarouge (NIR), D-optimal design, partial least squares (PLS) regression.

Abstract
The current study presents an application of near infrared spectroscopy for identification and quantification of the fraudulent addition of barley in roasted and ground coffee samples. Nine different types of coffee including pure Arabica, Robusta and mixtures of them at different roasting degrees were blended with four types of barley. The blending degrees were between 2 and 20 weight percent of barley. D-optimal design was applied to select 100 and 30 experiments to be used as calibration and test set, respectively. Partial lest squares regression (PLS) was employed to build the models aimed at predicting the amounts of barley in coffee samples. In order to obtain simplified models, taking into account only informative regions of the spectral profiles, a genetic algorithm (GA) was applied. A completely independent external set was also used to test the model performances. The models showed excellent predictive

ability with root mean square errors (RMSE) for the test and external set equal to 1.4% w/w and 0.8% w/w, respectively.
Keywords: coffee, barley, near infrared (NIR) spectroscopy, D-optimal design, partial least squares (PLS) regression.

1. Introduction

Coffee is one of the three most widely traded foodstuffs and the second largest commodity industry worldwide [1]. There are two varieties of coffee with economic importance: *Arabica* and *Robusta* [2]. Coffee *Arabica* is generally more appreciated for its organoleptic features and, thus, it is the most expensive [3]. Assurance of quality of roasted coffees has attracted widespread attention for controlling and preventing coffee adulteration, also given the great difference in the final sale price [4]. The principal adulterants of coffee include roasted and unroasted coffee husks, twigs, barley, chicory, malt, starch, corn, maltodextrins, glucose syrups, and caramelized sugar [5]. As the simple visual inspection is not an appropriate method for differentiating between the genuine coffee samples and the fraudulent ones, a number of analytical strategies have been developed. Materny *et al.* [6] applied micro Raman spectroscopy combined with chemometric methods to discriminate between green *Arabica* and *Robusta* coffees based on chlorogenic acid and lipid contents. Komes *et al.* [7] employed UV-Vis spectroscopy and HPLC analysis to determine the polyphenolic compounds and caffeine content of four different types of coffees. Martin and coworkers [8] used the tocopherol and triglyceride content of roasted and green coffees as features for discriminating between *Arabica* and *Robusta* varieties; principal component analysis (PCA) and linear discriminant analysis (LDA) were employed as pattern recognition tools. Gonzalez *et al.* [2] applied a relatively similar approach, based on fatty acid profiles as discriminant parameters for coffee variety differentiation. Digital image processing was carried out by Sano *et al.* [9] to quantify the amounts of brown sugar, coffee husk, maize, and soybean added to coffee *Arabica*. Near infrared spectroscopy combined with multivariate calibration methods was used by Pizarro and coworkers to quantify the content of *Robusta* variety in roasted coffee mixtures [4]. Jham *et al.* [10] investigated the potential of tocopherols determined by HPLC analysis as markers to detect coffee adulteration by corn. The feasibility of detection of coffee adulteration with roasted barley, based on volatile compound profiles was studied by Oliveira *et al.* [11]; solid phase microextraction (SPME) coupled with GC-MS analysis was carried out as analytical tool and chemometric methods were used for data processing. Lago and Nogueira [12] proposed a method based on acid hydrolysis of xylan and starch and consequent electrophoretic separation for identification of adulteration in processed coffee with cereals and coffee husks.

Despite of the relative success achieved by many of these approaches for determining coffee authenticity [13-18], it is important to consider that they are, in many cases, expensive, complex and/or time consuming. For this reason, a fast, reliable and low-cost technique with easy implementation for routine analysis represents a very attractive alternative for adulteration and varietal identification purposes [19-23].

Over the last decades, the application of near infrared spectroscopy (NIRS) as a fast and non-destructive technique for the authentication of food samples has become widespread thanks to the advances in chemometrics. Furthermore, it allows to directly analyze solid samples without any complex physical/chemical pre-treatment. Thus, several studies concerning NIR applications in food quality and authentication assessment have been reported [24-28].

In many studies on coffee adulterations, usually, one or two types of coffee and adulterants have been involved. So, the models obtained are poorly representative and are just applicable for these specific samples. In order to obtain a more representative and, thus, widely applicable model, it is advisable to collect and use a wider variety of coffee and adulterants. Taking into account that the exploration of all the possible combinations of different varieties, blends and roasting degrees would be impractical, it is

worth finding the minimum number of samples that is maximally representative of all the variability factors that characterize the whole statistical population of possible combinations. The optimal design techniques select highly representative subsets according to particular criteria. The usual approach is to specify a model, to determine the region of interest, to select the number of runs to be made, to specify the optimality criterion and, finally, to find the subset of designed points from the whole set of candidate points [34]. D-optimality is the criterion most widely applied for such a purpose [35, 36].
The objective of the present study was to determine the amount of barley added in coffee samples based on NIR spectral information. In order to obtain a widely applicable model, nine types of commercial coffee samples – chosen as to extensively explore the variability of coffee present on the market – and four types of barley samples, with different roasting degrees, were used for investigation. The concentrations of barley were changed from 2 to 20 % (w/w) at 10 levels. The lowest limit is surely lower than the sensorial limit of detection. Also the resolution step (2%) is lower than human sensorial capability for distinguishing between close quantities. By taking into account all the combinations, 360 mixtures (9 coffees × 4 barleys ×10 concentrations) should have been prepared. A D-optimal design was therefore applied to reduce the number of experiments maintaining the representativeness. The prediction ability of PLS models, either on the full spectra or after variable selection by means of genetic algorithms (GA) [37-40], was evaluated on a test sample set. All the models obtained were additionally tested for their prediction ability with ten independent mixtures, prepared with one coffee and one barley which were not used for preparing the training mixtures.

2. Materials and methods

2.1 Coffee and barley samples

Nine coffee bean varieties and four barley samples were obtained from specialized markets. The coffee samples were selected as to represent the most common types of coffees available on the Italian market, including both Arabica and Robusta as well as their mixtures at different roasting degrees.

2.2 Apparatus and procedure

Spectral profiles of powder samples were recorded in the reflection mode in the range 4,000-10,000 cm^{-1} with a resolution of 4 cm^{-1}, by an FT-near infrared spectrophotometer based on a polarization interferometer (Buchi NIRFlex N-500). Before analysis, coffee and barley toasted beans were ground with an electric grinder for about 60 s and, afterward, passed through a 0.3 mm sieve. Mixtures at different concentrations were prepared by separately weighting and accurately mixing the finely ground pure powders. Spectra of two grams of samples were recorded at a temperature of 20±1 ˚C, in a cylindrical quartz holder. Each spectrum recorded was the average of 32 successive scans. In addition, three acquisitions were performed, for each experiment, by manual rotation of the cell. In order to minimize the effect of uncontrollable factors, all the experiments were carried out in a random order.

2.3 Experimental design for calibration and validation

The selection of a subset representative of all the possible combinations of coffee and barley samples at ten different concentrations (from 2% to 20% w/w) represented a crucial step prior to carrying out a suitable and significant study, since the total number of combinations was considerably large (360 experiments). A total number of 100 experiments with the maximum representativeness was chosen, based on D-optimal design, to be used in the study as the calibration set. The spectra of all the pure

coffee samples as well were recorded and included in the calibration set, which was therefore formed by a total of 109 samples. Thirty experiments were also selected as the test set, among the remaining candidate experiments, by applying a subsequent D-optimal design. The experimental matrix is reported in Table 1. An additional evaluation, with a completely external set, was also performed using a new type of coffee and a new type of barley, at ten different concentrations.

2.4 Multivariate calibration

PLS regression analysis was performed in order to obtain a quantitative model for the prediction of barley amount based on spectral information. PLS is a latent variable-based method, particularly useful when dealing with noisy and collinear data [41].
Column autoscaling was applied on the spectral data as the pre-processing.
Data processing has been performed by programs developed by the authors under MATLAB environment (The MathWorks, Inc., Natick, Massachusetts).

N.	Coffee	Barley	Conc[a].	N.	Coffee	Barley	Conc.	N.	Coffee	Barley	Conc.
1	*E*	*d*	4.09	48	C	d	10.02	95	*E*	*c*	8
2	*G*	*c*	13.95	49	F	a	18.02	96	*B*	*a*	12.01
3	*B*	*c*	8.02	50	G	c	2.01	97	*F*	*d*	6.04
4	*H*	*a*	11.89	51	H	c	16	98	*G*	*a*	3.98
5	*A*	*b*	14.02	52	A	a	6.02	99	*A*	-	0
6	*D*	*a*	3.99	53	I	b	10	100	*F*	*b*	9.99
7	*E*	*b*	5.98	54	I	c	2	101	*A*	*d*	20
8	*C*	*a*	2.06	55	F	a	20.04	102	*G*	*d*	20
9	*B*	-	0	56	G	d	16.01	103	*H*	*a*	17.98
10	*G*	*c*	10.13	57	B	d	10	104	*A*	*d*	18.01
11	*E*	*b*	17.94	58	A	c	4	105	*I*	-	0
12	*I*	*d*	17.94	59	F	-	0	105	*B*	*c*	4
13	*E*	*c*	14.01	60	G	b	11.99	107	*D*	*c*	17.99
14	*F*	*d*	2	61	G	-	0	108	*E*	*a*	20
15	*H*	*d*	6.02	62	D	c	12.03	109	*B*	*d*	13.99
16	*I*	*c*	6.04	63	H	b	4.05	1^t	*B*	*D*	3.99
17	*B*	*b*	8.03	64	D	c	20.02	2^t	*F*	*B*	2.02
18	*B*	*c*	5.98	65	C	-	0	3^t	*G*	*C*	17.98
19	*A*	*c*	17.99	66	I	a	8.01	4^t	*A*	*d*	12.03
20	*H*	*d*	2.01	67	B	b	16	5^t	*C*	*c*	2.03
21	*I*	*b*	2.03	68	I	b	14.04	6^t	*C*	*d*	14
22	*G*	*b*	18.03	69	E	-	0	7^t	*C*	*d*	20
23	*C*	*a*	8	70	D	a	6	8^t	*E*	*d*	12
24	*G*	*b*	8.22	71	G	a	14.03	9^t	*F*	*a*	8
25	*C*	*c*	12.03	72	C	b	19.99	10^t	*H*	*d*	18
26	*A*	*c*	19.98	73	F	a	14.03	11^t	*E*	*c*	20
27	*D*	*d*	8	74	H	c	16.01	12^t	*F*	*c*	3.99
28	*C*	*b*	16.01	75	B	b	9.99	13^t	*I*	*c*	10
29	*A*	*a*	16.01	76	I	a	1.99	14^t	*A*	*b*	6.04
30	*I*	*a*	19.98	77	F	b	4	15^t	*I*	*a*	15.99
31	*I*	*d*	16.02	78	E	d	7.98	16^t	*B*	*a*	9.98
32	*B*	*d*	19.97	79	A	b	9.99	17^t	*E*	*d*	13.98
33	*E*	*b*	3.98	80	E	a	1.99	18^t	*H*	*a*	8
34	*F*	*c*	14.01	81	D	a	12.02	19^t	*H*	*b*	12
35	*H*	*d*	8	82	C	b	17.99	20^t	*D*	*b*	20
36	*H*	*b*	20.01	83	D	-	0	21^t	*B*	*a*	14.05

37	*H*	-	0	84	C	c	6	22^t	*E*	*b*	16.03
38	*C*	*a*	14.01	85	A	b	11.99	23^t	*G*	*d*	2.02
39	*E*	*c*	16.02	86	I	d	12.04	24^t	*G*	*b*	3.99
40	*F*	*c*	11.99	87	D	d	16.05	25^t	*D*	*a*	9.99
41	*A*	*d*	7.99	88	G	d	6.02	26^t	*A*	*c*	16
42	*F*	*d*	4	89	B	a	18.05	27^t	*D*	*c*	8.02
43	*H*	*b*	14.01	90	C	c	4.02	28^t	*I*	*b*	18.01
44	*A*	*a*	9.99	91	D	d	14.05	29^t	*H*	*c*	5.99
45	*C*	*d*	12.01	92	D	b	9.99	30^t	*G*	*a*	5.98
46	*H*	*a*	10	93	F	c	15.98				
47	*E*	*a*	2	94	D	b	6.02				

Table 1: Design matrix of calibration and test sets.
[a]: weight percents of Barley in Coffee (w/w%) (Real values of concentration were used for modeling)
[t]: refers to test set.

3. Results and discussion

3.1 Evaluation of different sources of variability

In an ideal system, the main variability is related to the variation of the factor of interest – that is, in the present study, the concentration of barley. However, various external factors – *e.g.*, related to sampling – can affect the global variance of the system.
As a first step, it has been verified that the variability related to sample preparation (mainly grinding of the raw materials, weighting of the coffee and barley powders, mixing) is smaller than the measurement variability (results not reported).

3.2 Multivariate calibration

Table 2 shows the root mean square error, the bias and the R^2 values obtained for the PLS models. For the full-spectrum approach, the optimal complexity (estimated by cross-validation) was 12 latent variables.
PLS has a high capability to extract relevant information and to produce a reliable prediction. However, in the last two decades, it has been recognized that an efficient feature selection can be highly beneficial, both to improve the predictive ability of the model and/or to reduce its complexity [42]. For the sake of selecting the proper regions of the spectra, a genetic algorithm (GA) was chosen as the feature selection technique. For the purpose of reducing the search domain, the 1501 original spectral variables were reduced to 188 by sequentially averaging the values of eight contiguous data-points. The GA-PLS algorithm was applied five times on the calibration set. The number of evaluations of each run was set to 200. According to the GA results, 16 variables were selected. Four spectral regions have been selected: the first one is between 6,032 and 5,748 cm^{-1}, the second one includes the 4,880-4,788 cm^{-1}range, the third one the 4,688-4,628 cm^{-1} range, and the fourth includes the narrow range between 4,336 and 4,276 cm^{-1}. These regions contain 128 wavenumbers in total. Although for the nature of NIRS it is not possible to univocally assign the vibrational transition related to the selected spectral bands, the majority of them are ascribable to the first overtone of N-H, C=O, O-H, C-H, and S-H functional groups of ArOH, H_2O, ROH, CONHR, and RNH_2. The results of PLS applied on the variables selected by GA are presented in Table 2.

	Coffee types	Total	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
PLS^{c}	**RMSE**	1.23	1.33	1.11	1.07	1.32	0.92	1.08	1.45	1.64	1.06
	Bias	0.04	0.29	-0.24	0.38	0.14	-0.08	-0.76	1.00	-0.74	0.39
	R^2	0.96	0.96	0.97	0.97	0.95	0.97	0.97	0.94	0.93	0.97
PLS^{t}	**RMSE**	1.46	1.43	1.26	1.01	1.13	0.94	0.95	1.93	2.00	0.36
	Bias	0.53	0.14	0.08	0.40	-0.31	0.06	0.01	0.67	-0.23	-0.34
	R^2	0.94	0.88	0.91	0.98	0.95	0.90	0.85	0.85	0.81	0.99
PLS^{e}	**RMSE**	0.85	------	------	------	------	------	------	------	------	-------
	Bias	-0.58	------	------	------	------	------	------	------	------	-------
	R^2	0.98	------	------	------	------	------	------	------	------	-------
GA-PLS^{c}	**RMSE**	1.18	1.35	1.49	0.91	1.07	1.20	0.81	1.31	1.27	1.06
	Bias	-0.01	0.39	-0.27	0.52	-0.05	-0.36	-0.48	0.46	-0.26	0.07
	R^2	0.96	0.96	0.94	0.98	0.97	0.96	0.98	0.96	0.96	0.97
GA-PLS^{t}	**RMSE**	1.42	0.83	1.46	0.50	0.65	1.06	1.10	1.8	2.5	1.13
	Bias	0.21	0.53	1.18	-0.08	0.44	0.21	-0.12	1.43	-1.13	-0.44
	R^2	0.94	0.96	0.88	0.99	0.98	0.87	0.80	0.92	0.70	0.89
GA-PLS^{e}	**RMSE**	1.10	------	------	------	------	------	------	------	------	-------
	Bias	-0.95	------	------	------	------	------	------	------	------	-------
	R^2	0.97	------	------	------	------	------	------	------	------	-------

Table 2: Statistical performance of PLS and GA-PLS models on the calibration, test and external sets. [c:] Calibration set, [t:]Test set, [e:] External set.

Eight latent variables were selected for modeling. As it can be seen, the predictive ability of the model is of the same order of that of PLS applied on the whole spectral range. However, the complexity of this model is considerably reduced in comparison to the full-spectrum one. The parity and residual plots of the two models are shown in Figs. 1 and 2.

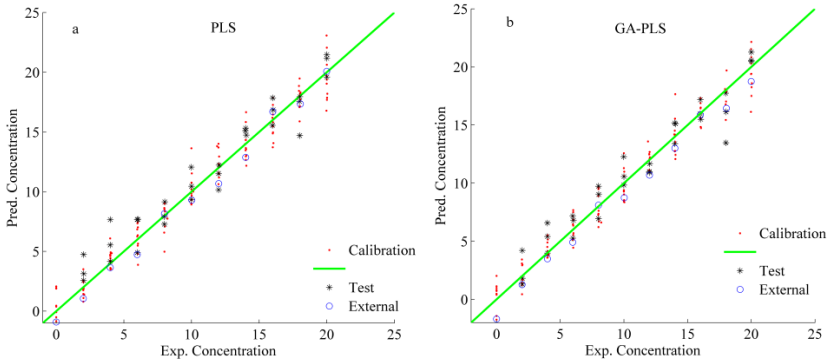


Figure 1: Experimental vs. predicted values of concentration (% w/w) of barley in coffee samples for (a) PLS model on the whole spectral range and (b) PLS model on spectral bands selected by GA.

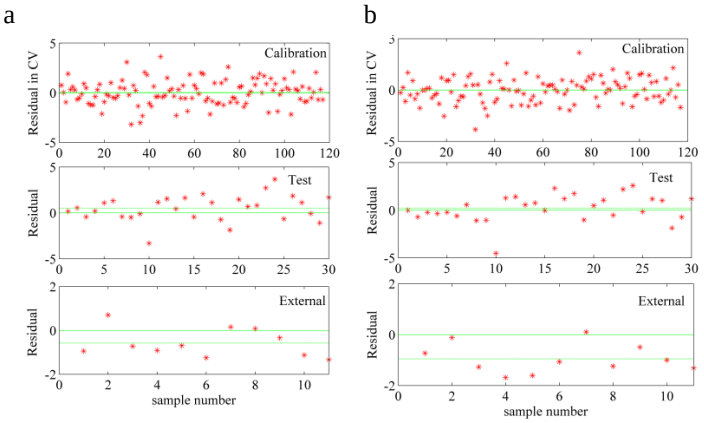


Figure 2: Residuals vs. sample number of (a) PLS model on the whole spectral range and (b) PLS model on spectral bands selected by GA. The solid line represents null residuals while the dashed line indicates the model bias value.

4. Conclusions

The excellent prediction ability obtained by multivariate calibration confirmed that non-destructive NIR measurements can be successfully employed for the detection and quantification of fraudulent addition of roasted barley to roasted coffee. Variable selection by using genetic algorithms helped to determine the spectral regions most useful to identify the adulteration of coffee with barley. The methodology allowed to quantify the amount of adulterant up to a level of 2% (w/w) of barley.
This paper clearly shows that the representatitivity of the training set is a key point in the success of a calibration model. The achievement of very low prediction errors on a totally external test set (*i.e.*, on mixtures composed by qualities of coffee and barley unknown to the model) has been possible only as a consequence of the fact that the training set was made by taking into account a relatively large number of varieties of coffee and barley. Another key point is the application of D-optimal design for the selection of a subset of adequate size from the very high set of candidate experiments.
The results reported in the present study indicate NIRS to be a promising procedure to be considered in future applications to quantify different adulterants in coffee powder.

Bibliography

[1] Pizarro, C., Esteban-Diez, I., Gonzalez-Saiz, J.M., Forina, M. (2007). Use of Near-Infrared Spectroscopy and Feature Selection Techniques for Predicting the Caffeine Content and Roasting Color in Roasted Coffees. *J. Agric. Food Chem.*, 55, 7477-7488.

[2] Martin, M.J., Pablos, F., Gonzalez, A.G., Valdenebro, M.S., Leon-Camacho, M. (2001). Fatty acid profiles as discriminant parameters for coffee varieties differentiation. *Talanta* 54, 291-297.

[3] Briandet, R., Kemsley, E.K., Wilson, R.H. (1996). Discrimination of Arabica and Robusta in Instant Coffee by Fourier Transform Infrared Spectroscopy and Chemometrics. *J. Agric. Food Chem.* 44, 170-174.

[4] Pizarro, C., Esteban-Diez, I., Gonzalez-Saiz, J.M. (2007). Mixture resolution according to the percentage of robusta variety in order to detect adulteration in roasted coffee by near infrared spectroscopy. *Anal. Chim. Acta*, 585, 266-276.

[5] Prodolliet, J., Bruelhart, M., Blanc, M.B., Leloup, V., Cherix, G., Donnelly, C.M., Viani, R. (1995). Adulteration of Soluble Coffee with Coffee Husks and Parchments. *J. AOAC Int.*, 78, 761-767.

[6] El-Abassy, R.M., Donfack, P., Materny, A. (2011). Discrimination between Arabica and Robusta green coffee using visible micro Raman spectroscopy and chemometric analysis. *Food Chem.*, 126, 1443-1448.

[7] Hecimovic, I., Belscak-Cvitanovic, A., Horzic, D., Komes, D. (2011). Comparative study of polyphenols and caffeine in different coffee varieties affected by the degree of roasting. *Food Chem.*, 129, 991-1000.

[8] Gonzalez, A.G., Pablos, F., Martin, M.J., Leon-Camacho, M., Valdenebro, M.S. (2001). HPLC analysis of Tocopherols and Triglycerids in coffee and their use as authentication parameters, *Food Chem.*, 73, 93-101.

[9] Sano, E.E., Assad, E.D., Cunha, S.A.R. (2003). Quantifying adulteration in roasted coffee powders by digital image processing. *J. Food Qual.*, 26, 123-134.

[10] Jham, G.N., Winkler, J.K., Berhow, M.A., Vaughn, S.F. (2007). γ-Tocopherol as a Marker of Brazilian Coffee (Coffea Arabica L.) Adulteration by Corn. *J. Agric. Food Chem.*, 55, 5995-5999.

[11] Oliveira, R.C.S., Oliveira, L.S., Franca, A.S., Augusti, R. (2009). Evaluation of the potential of SPME-GC-MS and chemometrics to detect adulteration of ground roasted coffee with roasted barley. *J. Food Comp. Anal.*, 22, 257-261.

[12] Nogueira, T., do Lago, C.L. (2009). Detection of adulterations in processed coffee with cereals and coffee husks using capillary zone electrophoresis. *J. Sep. Sci.*, 32, 3507-3511.

[13] Valdenebro, M.S., León-Camacho, M., Pablos, F., González, A.G., Martín, M.J. (1999). Determination of the arabica/robusta composition of roasted coffee according to their sterolic content. *Analyst*, 124, 999-1002.

[14] Alves, M.R., Casal, S., Oliveira, M.B.P.P., Ferreira, M.A. (2003). Contribution of FA Profile Obtained by High-Resolution GC/Chemometric Techniques to the Authenticity of Green and Roasted Coffee Varieties. *J. Am. Oil Chem. Soc.*, 80, 511-517.

[15] Martin, M.J., Pablos, F., Gonzalez, A.G. (1999). Characterization of arabica and robusta roasted coffee varieties and mixture resolution according to their metal content. *Food Chem.*, 66, 365-370.

[16] Martin, M.J., Pablos, F., Gonzalez, A.G. (1998). Discrimination between arabica and robusta green coffee varieties according to their chemical composition. *Talanta*, 46, 1259-1264.

[17] Schreyer, S.K., Mikkelsen, S.R. (2000). Chemometric analysis of square wave voltammograms for classification and quantitation of untreated beverage samples. *Sens. Actuators B*, 71, 147-153.

[18] Casal, S., Alves, M.R., Mendes, E., Oliveira, M.B.P.P., Ferreira, M.A. (2003). Discrimination between Arabica and Robusta Coffee Species on the Basis of Their Amino Acid Enantiomers. *J. Agric. Food Chem.*, 51, 6495-6501.

[19] Ribeiro, J.S., Ferreira, M.M.C., Salva, T.J.G. (2011). Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy. *Talanta*, 83, 1352-1358.

[20] Keidel, A., Stetten, D.V., Rodrigues, C., Maguas, C., Hildebrandt, P. (2010). Discrimination of Green Arabica and Robusta Coffee Beans by Raman Spectroscopy. *J. Agric. Food Chem.*, 58, 11187-11192.

[21] Wang, J., Jun, S., Bittenbender, H.C., Gautz, L., Li, Q.X. (2009). Fourier Transform Infrared Spectroscopy for Kona Coffee Authentication. *J. Food Sci.*, 74, 385-391.

[22] Esteban-Diez, I., González-Sáiz, J.M., Saenz-Gonzalez, C., Pizarro, C. (2007). Coffee varietal differentiation based on near infrared spectroscopy. *Talanta*, 71, 221-229.

[23] Esteban-Diez, I., González-Sáiz, J.M., Pizarro, C. (2004). An evaluation of orthogonal signal correction methods for the characterization of arabica and robusta coffee varieties by NIRS. *Anal. Chim. Acta*, 514, 57-67.

[24] Ferrari, E., Foca, G., Vognali, M., Tassi, L., Ulrici, A. (2011). Adulteration of the anthocyanin content of red wines: Perspectives for authentication by Fourier Transform-Near InfraRed and 1H NMR spectroscopies. *Anal. Chim. Acta*, 701, 139- 151.

[25] Oliveri, P., Di Egidio, V., Woodcock, T., Downey, G. (2011). Application of class-modelling techniques to near infrared data for food authentication purposes. *Food Chem.*, 125, 1450-1456.

[26] Wu, D., Nie, P., Cuello, J., He, Y., Wang, Z., Wu, H. (2011). Application of visible and near infrared spectroscopy for rapid and non-invasive quantification of common adulterants in Spirulina powder. *J. Food Eng.*, 102, 278-286.

[27] Toher, D., Downey, G., T.B. Murphy (2007). A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemom. Intell. Lab. Syst.*, 89, 102-115.

[28] Ruoff, K., Luginbuhl, W., Bogdanov, S., Bosset, J.O., B. Estermann, T. Ziolko, R. Amado (2006). Authentication of the Botanical Origin of Honey by Near-Infrared Spectroscopy. *J. Agric. Food Chem.*, 54, 6867-6872.

[29] Barnes, R.J., Dhanoa, M.S., Lister, S.J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.*, 43, 772-777.

[30] Isaksson, T., Næs, T. (1988). The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Appl. Spectrosc.*, 42, 1273-1284.

[31] Dehghani, H., Leblond, F., Pogue, B.W., Chauchard, F. (2010). Application of spectral derivative data in visible and near-infrared spectroscopy. *Phys. Med. Biol.*, 55, 3381-3399.

[32] Wold, S., Antti, H., Lindgren, F., Ohman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.*, 44, 175-185.

[33] Westerhuis, J. A., de Jong, S., Smilde, A.K. (2001). Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst.*, 56, 13-25.

[34] Montgomery, D.C. (2001). Design and analysis of experiments, 5th ed., Wiley & Sons Inc, pp.468.

[35] Mitchell, T.J. (1974). An algorithm for the construction of D-optimal experimental designs, *Technometrics*, 16, 203-210.

[36] Zunin, P., Fusella, G.C., Leardi, R., Boggia, R., Bottino, A., Capannelli, G. (2011). Effect of the Addition of Membrane Processed Olive Mill Waste Water (OMWW) to Extra Virgin Olive Oil. *J. Am. Oil Chem. Soc.*, DOI 10.1007/s11746-011-1856-2.

[37] Leardi, R., Gonzalez, A.L. (1998). Genetic algorithm applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.*, 41, 195-207.

[38] Leardi, R., Boggia, R., Terrile, M. (1992). Genetic algorithm as a strategy for feature selection. *J. Chemom.*, 6, 267-281.

[39] Jouan-Rimbaud, D., Massart, D.L., Leardi, R., de Noord, O.E. (1995). Genetic algorithm as a tool for variable selection in multivariate calibration. *Anal. Chem.*, 67, 4295-4301.

[40] Ghasemi, J., Niazi, A., Leardi, R. (2003). Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture. *Talanta*, 59, 311-317.

[41] Rannar, S., Lindgren, F., Geladi, P., Wold, S. (1994). A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: theory and algorithm. *J. Chemom.*, 8, 111-125.

[42] Leardi, R. (2003). Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data handling in science and technology*, 23, 169-196.

Détermination de la composition protéique du lait à partir de données spectrométriques moyen-infrarouge : comparaison de méthodes

Determination of protein composition in milk by mid-infrared spectrometry : comparison of methods

Sophie Guisnel[1], Marion Ferrand[1], Guy Miranda[2],
Félicie Faucon-Lahalle[1&3], Hélène Larroque[4], Patrice Martin[2], Mickaël Brochard[1]

[1] Institut de l'Elevage, 149 rue de Bercy, 75595 Paris cedex 12, France
E-mail: *Sophie.Guisnel@idele.fr*, *Marion.Ferrand@idele.fr*, *Felicie.Lahalle@idele.fr*, *Mickael.Brochard@idele.fr*
[2] INRA, UMR1313, Animal Genetics and Integrative Biology, Domaine de Vilvert 78352 Jouy-en-Josas cedex, France
E-mail : *Patrice.Martin@jouy.inra.fr*, *Guy.Miranda@jouy.inra.fr*
[3] CNIEL, 42 rue de Châteaudun, 75314 Paris cedex 09, France
[4] INRA, UR0631, Station d'Amélioration Génétique des Animaux, F-31326 Castanet-Tolosan cedex, France.
E-mail : *Helene.Larroque@toulouse.inra.fr*

Résumé

L'estimation des teneurs en protéines majeures du lait par analyse en spectrométrie moyen-infrarouge (MIR) est actuellement possible. La précision de cette méthode est cependant moyenne en particulier pour la caséine κ, l'α-lactalbulmine et la β-lactoglobuline. Dans cette étude nous avons testé et comparé différentes méthodes afin de réduire l'erreur d'estimation : prétraitement du spectre, méthodes de sélection de variables tels que les algorithmes génétiques ou Elastic Net. Notre étude montre que la dérivée première du spectre améliore notablement la précision des équations : l'erreur d'estimation pour la caséine β est réduite de 16% et son R^2 est amélioré de 13%. La sélection de longueurs d'ondes par algorithmes génétiques apporte également des améliorations pour certaines protéines (caséine α_{s1}, β-lactoglobuline). En revanche, les méthodes de régularisation ont peu d'intérêt sur ce type de données.

Mots-clés : spectrométrie moyen-infrarouge (MIR), lait bovin, protéines, caséines, lactosérum, régression PLS.

Abstract

Predicting milk protein composition by mid-infrared spectrometry is currently possible. However, the estimations are not accurate enough, particularly those of the whey proteins and the κ-casein. In this study, we compared different methods in order to reduce the estimation errors. Our study shows that the first derivative markedly improves equation accuracy. For instance, the estimation error for β-casein is reduced by 16% and its R^2 is increased by 13%. Wavelength selection by genetic algorithms also improves the results. On the other hand, regularization methods are of little relevance for this kind of data.

Keywords: mid-infrared (MIR) spectrometry, bovine milk, protein, Partial Least Squares (PLS) regression.

1. Introduction

Cette étude a été réalisée dans le cadre du programme PhénoFinlait, un projet réunissant les différents acteurs de la filière laitière et de la recherche agronomique dont l'objectif principal est de mieux connaître l'influence des effets génétiques et environnementaux sur la composition fine du lait (acides gras et protéines) afin de développer des outils de sélection génétique et d'appui technique en élevage.

La composition protéique est une information importante en transformation fromagère, un lait riche en caséine kappa (κ) aura un caillage plus rapide et plus ferme (Grosclaud, 1988). En nutrition humaine, certaines caséines et certaines protéines sériques pourraient également avoir des propriétés intéressantes (Debry, 2001). Mieux connaitre la composition protéique des laits en routine est donc pour la filière laitière un enjeu important afin de donner à chaque lait la destination la plus adéquate.

Une partie des travaux PhénoFinlait consiste au développement de deux méthodes de détermination de la composition fine en protéines : l'une est une détermination qualitative et quantitative par chromatographie en phase liquide couplée à la spectrométrie de masse (LC/MS), la seconde est l'estimation de la teneur de quelques protéines majeures à partir des données spectrales obtenues en spectroscopie moyen infrarouge (MIR). La LC/MS permet de caractériser de manière qualitative et quantitative les six protéines majeures du lait (caséines κ, caséine α_{S1}, caséine α_{S2}, caséine β, α-lactalbumine et β-lactoglobuline), toutefois cette technique est plus longue et plus onéreuse en investissement matériel que la spectrométrie moyen-infrarouge qui est utilisée en routine dans les laboratoires d'analyse du lait (interprofessionnel, contrôle laitier). Des travaux antérieurs (De Marchi, 2009 – Rutten, 2011 – Bonfatti, 2011) montrent qu'il est possible d'estimer la composition protéique des laits à partir de spectres MIR mais que ces estimations sont entachées d'erreurs élevées notamment pour la caséine κ et les protéines solubles du lactosérum. Au sein du programme PhénoFinlait, des travaux préliminaires en 2009, non publiés arrivaient aux mêmes conclusions. Pour réduire l'erreur d'estimation, différentes méthodes ont été testées en parallèle de la régression PLS, qui est la méthode utilisée dans les articles précédemment cités. Pour réduire le bruit lié aux variations spectrales, nous avons retenu la dérivée première comme méthode de prétraitement et afin de limiter le bruit apporté par les longueurs d'ondes non informatives (Leardi, 1992), nous nous sommes orientés vers des méthodes de sélection de variables, telles que les algorithmes génétiques ou les méthodes de régularisation.

2. Matériel et méthodes

2.1 Données spectrales en moyen-infrarouge (MIR)

Une banque de données spectrales MIR **(**Milkoscan FT6000, FOSS et Bentley FTS) a été constituée entre 2009 et 2011 à partir de 86 480 laits de bovins provenant de 1 136 élevages répartis dans différentes régions de France. Les travaux présentés par la suite ont été établis à partir de spectres FOSS exclusivement. Les mesures d'absorbances MIR concernent 1060 longueurs d'ondes (de 5012 à 926 cm^{-1}), toutefois seulement 436 ont été sélectionnées, correspondant aux bandes d'absorptions comprises entre 965 et 1544 cm^{-1}, entre 1716 et 2272 cm^{-1} et entre 2434 et 2970 cm^{-1}. En effet, selon la documentation FOSS (Reference Manual du Milkoscan FT 120 type 71200), les zones d'absorption de l'eau comprises entre 1715 et 1545 cm^{-1} et entre 3627 et 2971 cm^{-1} sont à proscrire car elles détériorent

la précision des estimations. De plus, la bande comprise entre 5012 et 3628 cm^{-1}, considérée comme peu informative, a également été éliminée.

2.2 Analyses quantitatives et qualitatives des laits en LC/MS

En parallèle, les laits de 271 bovins de race montbéliarde, et croisés holstein normand, ont été sélectionnés pour leurs variabilités et analysés également par une méthode récemment mise au point à l'INRA (Miranda & Martin, à paraître) faisant appel à la chromatographie liquide en phase inverse (RP-HPLC) couplée à la spectrométrie de masse (LC/MS). Cette méthode d'analyse permet, entre autre, de mesurer les quantités relatives des protéines majeures du lait (caséines κ glycosylée et non glycosylée, caséine α_{S1}, caséine α_{S2}, caséine β, α-lactalbumine et β-lactoglobuline) exprimées en % du total des pics présents dans le chromatogramme et converties par la suite en g/100g de lait après estimation du taux protéique des laits. Cette méthode permet également d'identifier les principales isoformes de ces protéines : variants génétiques, variants d'épissage et isoformes résultant de modifications post-traductionnelles (phosphorylation et glycosylation). L'analyse de ces laits a montré pour certains d'entre eux des taux de protéolyse importants. Afin de construire les équations fiables, nous avons décidé de ne tenir compte que des données spectrales des laits présentant un taux de protéolyse inférieur à 20%, soit 193 échantillons sur les 271 analysés.

2.3 Mise au point des équations

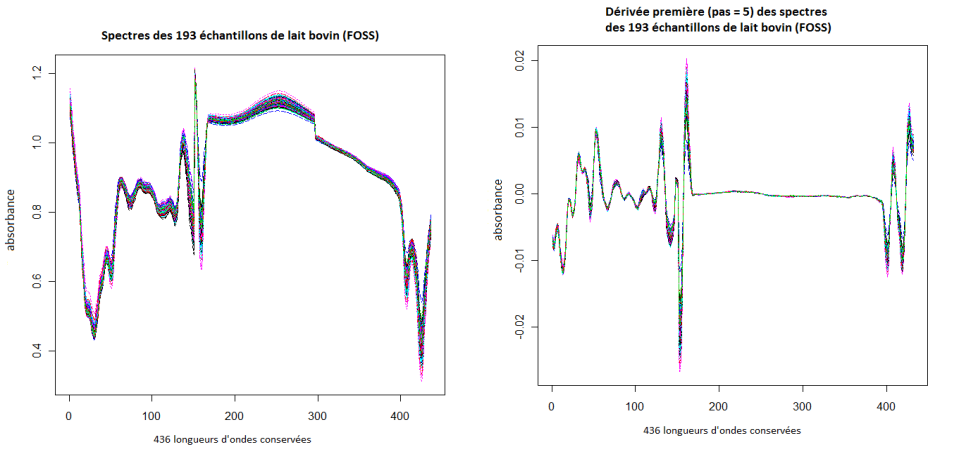
2.3.1 La régression PLS1

La régression PLS (Partial Least Squares regression) est la méthode la plus connue et la plus utilisée dans le domaine de la spectrométrie infrarouge. Elle présente deux intérêts majeurs : elle permet de traiter les cas où les individus sont moins nombreux que les variables explicatives et peut être utilisée lorsque ces dernières sont fortement corrélées entre elles. Par rapport à la régression sur composantes principales, la régression PLS permet de maximiser la corrélation entre les variables explicatives et les variables à expliquer (Tenenhaus, 1998).

2.3.2 Le prétraitement des données spectrales

Les données spectrales brutes, telles qu'elles sont acquises par un spectromètre, ne sont pas toujours de la forme la plus adaptée pour les traitements ultérieurs. L'objectif du prétraitement des données spectrales est de réduire l'effet des déformations incontrôlées des spectres ou celui des perturbations aléatoires, afin de construire par la suite des équations les plus robustes possibles (Bertrand & Dufour, 2006). Les deux principaux prétraitements appliqués généralement aux spectres sont le lissage et la dérivation : le lissage pour réduire le bruit aléatoire et la dérivation pour éliminer les variations incontrôlées du signal.

La régression PLS est une méthode peu sensible aux perturbations aléatoires, le lissage n'est donc pas nécessaire dans notre cas. A l'inverse, la dérivation du spectre, très souvent utilisée en spectrométrie du proche infra-rouge (SPIR), peut s'avérer utile comme l'ont montré Soyeurt et al. (2011). Les figures 1a et 1b présentent les spectres MIR des 193 laits analysés, respectivement avant et après dérivation.



Figures 1a et 1b. Spectres MIR des 193 laits analysés, respectivement avant et après dérivation

2.3.3 Sélection de longueurs d'ondes par algorithme génétique

La sélection de variables, associée à la PLS1, permet de limiter le bruit apporté par les longueurs d'ondes non informatives. Les algorithmes génétiques introduits par J. Holland (1992) et spécifiques à la sélection de longueurs d'ondes, imitent l'évolution des organismes vivants selon Darwin. C'est une méthode itérative mi-stochastique, mi-déterministe dont la première étape consiste à initialiser de manière aléatoire un pool de solutions, c'est-à-dire une sélection de variables. Chaque itération va construire une génération contenant de meilleures solutions que la génération précédente. L'algorithme doit converger vers une solution optimale.

2.3.4 Sélection de longueurs d'ondes par des méthodes de régularisation

En spectrométrie infrarouge, du fait de la colinéarité des données et du nombre important de variables, la régression linéaire multiple est inutilisable : la variance des estimateurs des coefficients de régression est importante et rend les estimations trop instables. Les méthodes de régularisation telles que la régression *Ridge*, LASSO (Least Absolute Shrinkage and Selection Operator) ou Elastic Net ont pour objectif de réduire la variance des estimateurs des moindres carrés ordinaires et ainsi garantir la stabilité des estimations. Elles consistent à introduire une fonction de pénalisation : les coefficients de régression β sont estimés avec un léger biais permettant de contrôler la variance. Grâce à un meilleur compromis entre biais et variance, on obtient par ces méthodes de meilleures estimations que le modèle de régression classique.

La régression *Ridge* (Hoerl and Kennard, 1988) consiste à ajouter une pénalité de norme l2 :

$$\widehat{\beta^{ridge}} = \arg\min_{\beta \in R^p} \{RSS(\beta) + \alpha \|\beta\|_2^2\}$$

Le paramètre de pénalisation α doit être judicieusement choisi afin de minimiser l'erreur d'estimation. Cette méthode conserve tous les prédicteurs dans le modèle.

La régression LASSO (Tibshirani, 1996) consiste à ajouter une pénalité en norme l1 :

$$\widehat{\beta^{lasso}} = \arg\min_{\beta \in R^p} \left\{ \frac{1}{2} RSS(\beta) + \lambda \|\beta\|_1 \right\}$$

Cette méthode a l'avantage d'annuler exactement certaines valeurs des coefficients estimés pour des valeurs suffisamment fortes du paramètre de pénalisation λ. Le modèle sélectionne ainsi les variables les plus informatives et gagne en interprétation. Toutefois, la régression LASSO présente deux principales limites dans notre cas. Elle sélectionne au plus *n* variables (*n* représentant la taille de l'échantillon) avant de saturer et dans un groupe de variables très corrélées, elle sélectionne arbitrairement une seule variable.

Une solution proposée plus récemment consiste à utiliser un compromis des deux méthodes. Le critère Elastic Net proposé par (Zou and Hastie, 2005) est de la forme :

$$\widehat{\beta^{e.net}} = \arg\min_{\beta \in R^p} \left\{ \frac{1}{2} RSS(\beta) + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2\right) \right\}$$

Les deux paramètres de pénalisation λ et α sont choisis par validation croisée, tels que $0 < \alpha < 1$. On note que lorsque le paramètre α tend vers 0, la méthode se rapproche de la régression *Ridge* et toutes les variables du modèle sont retenues. A l'inverse, lorsque le paramètre α tend vers 1, la méthode revient à la régression LASSO, et la sélection de variables est plus réduite.

La méthode Elastic Net permet d'une part de sélectionner plus de *n* prédicteurs et d'autre part de retenir toutes les variables au sein d'un groupe présentant de fortes corrélations. Elle semble ainsi être adaptée aux données spectrales.

Qing Li et Nan Lin (2010) propose une analyse bayésienne de la méthode Elastic Net où les deux paramètres de pénalités λ et α sont choisis simultanément par l'algorithme de Monte Carlo. En général, la méthode bayésienne donne des résultats comparables quant à la qualité de régression avec un avantage de la méthode usuelle pour des modèles simples et un avantage de la méthode bayésienne pour des modèles plus compliqués.

2.4 Mise en œuvre des méthodes d'analyse spectrale et comparaison

Les équations sont établies sous le logiciel R 2.13.1 en prenant 70% des données comme jeu de calibration et 30% comme jeu de validation. Pour évaluer la précision des estimations, différents paramètres statistiques sont calculés : moyenne, écart-type, écart résiduel, écart résiduel relatif, R^2.

La régression PLS1 est réalisée avec le package '*pls*'. Pour chaque équation, le nombre optimal de variables latentes est choisi selon l'erreur quadratique moyenne obtenue par validation croisée ($RMSEP_{cv}$).

Les régressions *Ridge*, LASSO et Elastic Net sont réalisées à l'aide du package '*glmnet*'. Les estimations sont calculées soit directement à partir des coefficients de régression, soit en réalisant une régression PLS1 sur la sélection de variables proposée.

L'algorithme génétique utilisé dans cette étude est celui développé par Leardi (1998), spécifique à la sélection de longueurs d'ondes. Il est programmé sous le logiciel Matlab. Les variables étant nombreuses, deux tours d'algorithme sont nécessaires. Lors du premier tour, l'algorithme est appliqué sur la moyenne des absorbances à 3 longueurs d'onde consécutives. Au second tour, il est appliqué sur les longueurs d'ondes sélectionnées au premier tour. Les paramètres de l'algorithme génétique que nous avons utilisés sont ceux proposés initialement par Leardi (1998), et repris lors de la construction des équations de prédiction des acides gras (Ferrand *et al*, 2010) : 5 variables sélectionnées en moyenne par solution dans le pool de solutions initiales, une probabilité de mutation de 1% et une probabilité de recombinaison de 50%. Pour garantir une convergence optimale, l'algorithme est lancé cinq fois de manière indépendante.

3. Résultats et discussion

Par régression PLS1 (tableau 1), les caséines α_{S1} et β sont les protéines les mieux estimées atteignant respectivement des R^2 de 0,71 et 0,68, et des erreurs relatives de 6,86% et de 7,22%. Les estimations sont de qualité moyenne pour les caséines κ (R^2= 0,54) et α_{S2} (R^2= 0,58). La fraction glycosylée de la caséine κ est très mal estimée, tout comme les protéines sériques.

Prétraitement des spectres :

Le tableau 1 permet d'observer l'effet de la dérivation première du spectre. Différents possibilités de pas ont été testées, c'est la dérivée prenant en compte un pas (intervalle) de cinq points consécutifs du spectre qui a été retenue comme meilleur compromis, de la même façon que Soyeurt H. et al. (2011). Cette dérivée est obtenue par la formule suivante :

$$\frac{dx}{d\lambda_i} = \frac{x_{i-3} - x_{i+3}}{\lambda_{i-3} - \lambda_{i+3}}$$ où $\frac{dx}{d\lambda_i}$ est la valeur de la dérivée au point *i*.

Associée à la régression PLS1, celle-ci améliore les précisions des estimations à l'exception de la caséine α_{S1} et de la β-lactoglobuline. L'erreur d'estimation pour la caséine β est réduite de 16,3 1% et son R^2 est amélioré de 12,72%. Pour l'α-lactalbulmine, les améliorations apportées par la dérivée sont également importantes (le R^2 augmentant de 23%) mais les estimations finales restent peu précises.

Tableau 1. Paramètres statistiques des équations estimant la concentration en g/100g de lait des protéines du lait bovin : PLS1 ou dérivée + PLS1.

Composition en protéines (g/100g de lait)	**N**	**Moyenne**	**Ecart-type**	**Ecart-type résiduel relatif**[1] (en %)		**R^2**	
				PLS1	**dérivée + PLS1**	**PLS1**	**dérivée + PLS1**
Caséines (CN)	57	2,458	0,271	3,93	3,72	0,88	0,89
κ-CN glycosylée	58	0,112	0,034	26,40	24,12	0,26	0,38
κCN	57	0,317	0,054	11,61	10,89	0,54	0,60
α_{S2}-CN	57	0,237	0,041	11,25	10,43	0,59	0,64
α_{S1}-CN	58	0,861	0,099	6,32	6,86	0,71	0,65
β-CN	57	1,038	0,131	7,22	6,04	0,68	0,77

Protéines sériques	57	0,385	0,058	9,96	9,64	0,58	0,61
α-LA	57	0,123	0,018	11,80	10,90	0,39	0,48
β-LG	58	0,263	0,054	15,79	15,86	0,42	0,41

[1] Ecart-type résiduel relatif= écart-type résiduel / moyenne.

Sélection des longueurs d'ondes :

- par algorithmes génétiques

Au premier tour, les algorithmes appliqués sur les 436 variables de la présélection FOSS sélectionnent une centaine de variables pour chaque protéine individuelle et environ 200 pour les quantités totales de caséines ou protéines sériques. Pour le deuxième tour, on a gardé entre 8 et 83 variables selon la protéine considérée (tableau 2). L'utilisation des algorithmes génétiques (AG) avant d'effectuer la PLS1 permettent d'améliorer la qualité des équations pour toutes les protéines, exceptée pour la caséine α_{S1}.

Cependant, les résultats sont équivalents ou moins bons par rapport à l'utilisation de la dérivée. L'estimation de la β-lactoglobuline est ainsi légèrement meilleure : le R^2 est augmenté de 7,14% et l'erreur relative est réduite de 3,2%. A l'inverse, l'estimation de la caséine κ est bien moins précise : le R^2 est diminué de 32% et l'erreur relative augmente de 20,75%.

Tableau 2 Paramètres statistiques des équations estimant la concentration en g/100g de lait des protéines du lait bovins : algorithme génétique (un ou deux tours) + PLS1.

Composition en protéines (g/100g de lait)	**N**	**Moy.**	**Ecart-type**	**Nb var AG 1 tour**	**Nb var AG 2 tours**	**Ecart-type résiduel relatif** (en %)		**R^2**	
						AG 1tour + PLS1	**AG 2tours + PLS1**	**AG 1tour + PLS1**	**AG 2tours + PLS1**
Caséines	58	2,457	0,269	205	76	3,88	3,85	0,88	0,88
κ-CN glycosylée	57	0,110	0,032	92	8	26,87	25,99	0,15	0,20
κ-CN	57	0,316	0,052	133	28	13,42	13,15	0,33	0,41
α_{S2}-CN	58	0,237	0,041	79	15	10,29	10,59	0,65	0,63
α_{S1}-CN	58	0,861	0,099	113	9	5,47	6,31	0,69	0,70
β-CN	58	1,041	0,132	110	34	5,91	6,70	0,78	0,72
Protéines sériques	58	0,387	0,060	223	39	13,67	9,35	0,24	0,63
α-LA	57	0,123	0,018	109	69	10,90	10,91	0,42	0,48
β-LG	58	0,263	0,054	110	83	15,29	15,39	0,45	0,44

- par méthodes de régularisation

Sur les 436 variables initiales, la régression LASSO en sélectionne entre 20 et 30. L'Elastic Net (EN) avec un paramètre de pénalité α égal à 0,5 en sélectionne environ deux fois plus. En réalisant une régression PLS1 sur ces sélections de variables, seules les estimations des caséines β et α_{S1} sont améliorées (tableau 3) et ces améliorations restent plus faibles que celles apportées par l'utilisation de la dérivée ou des algorithmes génétiques. En calculant les estimations directement à partir des

coefficients de régression Elastic Net, on obtient les mêmes conclusions. Ces méthodes semblent peu adaptées à notre jeu de données.

Tableau 3 Paramètres statistiques des équations estimant la concentration en g/100g de lait des protéines du lait bovin : Elastic Net (α=0,5) ou Elastic Net (α=0,5) + PLS1 ou LASSO + PLS1.

Composition en protéines (g/100g de lait)	**N**	**Moy.**	**Ecart-type**	**Nb var EN (α=0,5)**	**Nb var LASSO**	**Ecart-type résiduel relatif**			**R^2**		
						EN (α=0,5)	**EN (α=0,5) + PLS1**	**LASSO + PLS1**	**EN (α=0,5)**	**EN (α=0,5) + PLS1**	**LASSO + PLS1**
κ-CN glycosylée	57	0,110	0,032	22	4	28,47	28,49	26,97	0,13	0,04	0,14
κ-CN	57	0,316	0,052	43	19	14,05	14,56	14,59	0,33	0,28	0,27
α_{S2}-CN	58	0,237	0,041	56	24	11,43	11,76	11,64	0,57	0,55	0,56
α_{S1}-CN	58	0,861	0,099	46	21	6,37	6,97	6,65	0,70	0,64	0,63
β-CN	58	1,041	0,132	47	25	7,09	6,38	6,99	0,69	0,75	0,70
α-LA	57	0,123	0,018	68	29	12,92	13,60	13,50	0,32	0,19	0,20
β-LG	58	0,263	0,054	67	27	16,63	16,28	15,89	0,35	0,36	0,39

Au final, les caséines β et α_{S1} restent les mieux estimées, avec des R^2 respectifs de 0,77 et 0,71 et des erreurs relatives de 6,04% et 6,32% (tableau 4). Les caséines α_{S2} et κ sont correctement estimées, avec des R^2 respectifs de 0,65 et 0,60 et des erreurs relatives de 10,29% et 10,89%. Les estimations de la fraction glycosylée de la caséine κ sont en revanche de mauvaise qualité. De même pour les protéines sériques, les équations restent à améliorer : les estimations obtenues par les méthodes testées ne sont pas assez précises.

Tableau 4 Tableau récapitulatif des résultats.

Composition en protéines (g/100g de lait)	**Méthode choisie**	**Ecart-type résiduel relatif** (en %)	**R^2**
Caséines	dérivée + PLS1	3,72	0,89
κ-CN glycosylée	dérivée + PLS1	24,12	0,38
κ-CN	dérivée + PLS1	10,89	0,60
α_{S2}-CN	AG 1 tour + PLS1	10,29	0,65
α_{S1}-CN	PLS1	6,32	0,71
β-CN	dérivée + PLS1	6,04	0,77
Protéines sériques	AG 2 tours + PLS1	9,35	0,63
α-LA	dérivée + PLS1	10,9	0,48
β-LG	AG 1 tour + PLS1	15,29	0,45

4. Conclusion

Notre étude montre que la spectroscopie MIR permet de prédire correctement les concentrations des principales caséines dans le lait de vache. Pour les protéines sériques, les équations restent toutefois à améliorer. Nous avons pu observer que la régression PLS appliquée à la dérivée du spectre est une méthode donnant de bons résultats. La sélection de longueurs d'ondes par algorithmes génétiques apporte également quelques améliorations pour certaines protéines. Les méthodes de régularisation, en revanche, ont peu d'effet sur notre jeu de données.
Les résultats sont comparables à ceux de Rutten *et al.* (2011), hormis pour la β-lactoglobuline qui pour ces auteurs présentait une erreur d'estimation nettement plus faible. Les différences observées pourraient provenir de la méthode de référence utilisée qui n'est pas la même dans les deux études (LC-MS versus électrophorèse capillaire) et pourrait générer des valeurs quantitatives plus ou moins précises selon le cas.
Afin d'améliorer l'estimation de la composition protéique par spectrométrie MIR, il semble avant tout chose nécessaire d'affiner les valeurs quantitatives obtenues par notre méthode de référence en travaillant sur un nombre plus important d'individus, surtout pour des composants comme les protéines où une forte variabilité de structure peut exister (variants génétiques).

5. Remerciements

Cette étude a été financée par l'ANR, Apis-Gène, le CNIEL, France Génétique Elevage, FranceAgriMer et le Ministère de l'Agriculture. Les auteurs remercient les domaines expérimentaux INRA de Mirecourt et du Pin pour leur aide technique ainsi que le comité de pilotage du programme PhénoFinlait pour les discussions constructives.

Bibliographie

Bertrand D. & Dufour E. (2006). *La spectrométrie infrarouge et ses applications analytiques*. Second ed., Tec&Doc Lavoisier, Paris.

Bonfatti V. (2011). *Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows*. J. Dairy Science 94, 5776-5785.

Chen T. & Martin E. (2009). *Bayesian linear regression and variable selection for spectroscopic calibration.* Analytica Chimica Acta 631, 13-21.

Chiquet J. (2009). *Analyse de données prostate : quelques méthodes de régularisation l1*. http://stat.genopole.cnrs.fr/_media/members/jchiquet/teachings/l1_reg.pdf
Debry G. (2001). *Lait, nutrition et santé* – Tec&Doc Lavoisier, Paris.

De Marchi M. & al. (2009). *Prediction of protein composition of individual cow milk using mid-infrared spectroscopy.* Ital.J.Anim.Sci 8(2),399-401.

Ferrand M. & al. (2010) *Application d'un algorithme génétique en spectrométrie moyen infrarouge pour estimer le profil en acides gras du lait de chèvre*. Agrostat 2010.

Ferrand M. & al. (2010). *Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression.* Chemometr. Intell. Lab. Syst. 106, 183-189.

FOSS (1998). *Reference Manual of Milkoscan FT120 (Type 71200)*. Denmark.

Hastie T. & al (2009) *Fast Regularization Paths via Coordinate Descent*. http://www-stat.stanford.edu/~hastie/TALKS/glmnet.pdf

Huquet B. (2009). *Prédiction de la composition du lait en acides gras à partir de données spectrométriques. Quelle méthode privilégier ?* Mémoire de fin d'études pour l'obtention du diplôme d'Agronomie Approfondie, Agrocampus Rennes.

Grosclaude F. (1988). *Le polymorphisme génétique des principales lactoprotéines bovines. Relations avec la quantité, la composition et les aptitudes fromagères du lait*. INRA Prod. Anim., 1 (1), 5-17.

Jarvis R. M. & Goodacre R. (2005). *Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data.* Bioinformatics 21 (7), 860-868.

Lalanne C. (2008-2009), *Approche méthodologique pour l'intégration de données de neuroimagerie et de génétique.* Mémoire de Master 2 Bioinformatique et Biostatistiques, Université Paris Sud 11.

Leardi R. & Lupiañez G. (2002). *Genetic algorithms applied to feature selection in PLS regression: how and when to use them.* Anal.Chim. Acta. 461, 189-200.

Li Q. & Lin N. (2010). *The Bayesian Elastic Net*. Bayesian Analysis 5 (1), 151-170.

Rutten M.J.M. & al. (2011). *Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra.* Journal of Dairy Science, 94:4183-88.

Rutten M.J.M. & al. (2011). *Predicting bovine milk protein composition based on Fourier transform infrared spectra.* Journal of Dairy Science, 94:5683-5690.

Soyeurt H. & al. (2011). *Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries*. Journal of Dairy Science, 94:1657-67.

Tenenhaus M. (1998). *La régression PLS: théorie et pratique*. Editions TECHNIP.

Tibshirani R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, Series B, 58(1):267-288.

Zou H. & Hastie T. (2004). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society, 67, part 2, pp301-320.

Zou H. & Hastie T. (2003). *Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays.* http://webdocs.cs.ualberta.ca/~mahdavif/ReadingGroup/Papers/10.1.1.9.6188.pdf

Session 5 : Maîtrise des Procédés I / *Process Control I*

Modélisation de la dynamique microbienne dans l'industrie alimentaire
Une approche basée sur des principes de plans d'expériences pour la microbiologie prévisionnelle

Modeling microbial dynamics in food processes
An experiment design approach to predictive microbiology

Eva Van Derlinden & Jan Van Impe

BioTeC - Chemical and Biochemical Process Technology and Control,
Department of Chemical Engineering, Katholieke Universiteit Leuven
E-mail : *eva.vanderlinden@cit.kuleuven.be, jan.vanimpe@cit.kuleuven.be*

Résumé
Pour assurer la sécurité et la qualité microbienne des aliments dans toute la chaîne alimentaire, connaissance de l'influence de l'environnement sur l'évolution microbienne est indispensable. A partir d'information expérimentale, la microbiologie prévisionnelle développe des modèles mathématiques pour évaluer la dynamique microbienne dans des produits alimentaires.
La qualité du modèle détermine largement comment les prévisions des modèles rapprochent la dynamique microbienne réelle pendant une étape du processus et pendant une période spécifique. La construction du modèle, l'identification des paramètres et la validation du modèle sont des procédures *data-driven*. Dans ce contexte, le choix rigoureux des expériences est extrêmement important. En appliquant des techniques mathématiques basées sur des principes statistiques, on construit des expériences specifiques relevantes pour l'industrie alimentaire, qui améliorent le procédé de construction du modèle et la qualité des paramètres.
Mots-clés: microbiologie prévsionnelle, estimation des paramètres, modèles cinétiques, modèles probabilistes, plan d'expériences

Abstract
To ensure microbial food safety and quality throughout the food chain, knowledge about the influence of the environment on microbial behavior is indispensable. Based on experimental information, predictive microbiology develops mathematical models to enable the evaluation of microbial dynamics in real food products.
How closely model predictions approximate microbial dynamics during a specific processing step and time span is highly determined by the model accuracy. As both model building, and subsequent parameter estimation and model validation, are data-driven steps, careful selection of the experiments is of high importance. By applying statistically based mathematical techniques, experiments relevant for food decontamination processes can be selected to optimize the model building procedure and improve the parameter estimation accuracy.
Keywords: predictive microbiology, parameter estimation, kinetic models, probabilistic models, experiment design

1 Introduction

To ensure microbial food safety and quality throughout the complete food chain, sufficient and accurate knowledge about the influence of environmental conditions on microbial behavior is indispensable. Based on experimental information, predictive microbiology develops mathematical models to predict microbial behavior given certain (dynamic) environmental conditions. Ultimately, these models enable the prediction of microbial behavior in real food products. Predictive models have important applications in risk assessments and HACCP. The implementation of predictive models has been improved by their integration in software packages (e.g., Combase [US-UK] and Sym'Previus [FR]), useful for academia, government and food industry. In addition, predictive models are an essential tool for risk control during optimization of various bio-engineering processes.

Accurate predictions of the microbial evolution ask for (*i*) a model structure that accurately describes the selected system, and (*ii*) reliable model parameter values with good statistical quality. The selection of the explanatory and response variables and the model structure, as well as the model parameter estimation are data-driven processes. Therefore, efficient and accurate model building requires highly informative experimental studies, i.e., the static and/or dynamic experiments should cover the complete targeted range of the explanatory variables, and the combination of all experiments performed should enclose sufficient and accurate information.

2 Kinetic growth rate models

Secondary predictive models are being developed to describe the influence of changing environmental conditions on growth and/or inactivation dynamics. With respect to modeling the effect of the extrinsic and intrinsic conditions on the microbial growth rates, models are mainly square root-type or cardinal parameter-type models. These model structures have the advantage that they can be easily extended towards additional environmental factors. As these model structures are generally assumed valid for a wide range of microorganisms, focus is often on the estimation of the related model parameters. Here, two mathematical techniques/approaches to increase and/or optimize the information contained in a (series of) experiment(s) are illustrated.

Design of experiments (DOE) is an experimental approach that enables to determine the relation between (environmental) factors, their interactions and statistical properties. In the domain of predictive microbiology, the technique of DOE is mostly used: (1) to scan an extended region for which a probabilistic model is built, (2) to collect data in an extended region to build response surface models, and (3) to collect data in a specified region to obtain accurate and reliable parameter estimates of existing models. The potential of this approach is illustrated for two secondary models describing the growth rate as a function of different environmental conditions. Next to the most often applied full factorial design, other designs (i.e., fractional factorial, central composite, Latin-square and Box-Behnken design) are evaluated with respect to the accurate and efficient estimation of the model parameters. As can be expected, full factorial designs perform best. Latin-square designs and Box-Behnken designs also yield acceptable parameter estimates while significantly reducing the total number of experiments. The poorest performance was observed for the central composite designs and randomly selected experiments (Mertens et al. 2012).

Optimal experiment design for parameter estimation (OED/PE) is a mathematical technique that enables to pick a small set of highly informative, static and/or dynamic experiments, resulting in unique and accurate parameter estimates. OED/PE assumes that the model structure taken is valid. When applying dynamic experiments, this approach also guarantees parameter estimates that are valid under varying, more realistic conditions. The significant reduction in the experimental burden when using the OED/PE approach is shown using the Cardinal Temperature Model with Inflection (CTMI) as a case study. Focus is on the efficient and accurate estimation of the model parameters (T_{min}, T_{opt}, T_{max} and μ_{opt}). The obtained model parameter values are characterized by a small uncertainty error and yield a good result during validation. More information can be found in Van Derlinden et al. (2008, 2010, 2012).

3 Probabilistic models

Generally, kinetic models have found wide acceptance since they perform well under conditions that permit rapid population development. However, care should be taken that predictions from kinetic models, due to their semi-mechanistic or empirical basis, are not made beyond the interpolation region. Since no growth conditions are usually omitted from the model fitting process, conditions close to the growth/no growth boundary, which are often of industrial interest, may not lie within this region. Moreover, when a microbial population experiences progressively harsher conditions and moves towards conditions that will eventually preclude growth, variability increases significantly, and kinetic models will fail to provide accurate descriptions. Under these circumstances, it is more useful to consider the probability that growth is likely to occur at all, rather than the growth rate. This type of approach is now widely used to develop probabilistic models that define combinations of environmental factors representing the boundary between growth and no growth, which are, in fact, a quantitative description of the hurdle concept.

With respect to the food industry, growth/no growth models (G/NG) are particularly relevant to pathogens with a low infection dose, as their ability to initiate growth implies that they have the potential to be harmful to the consumer. With respect to food poisoning, a similar approach can be adopted to develop toxin/no toxin production models. Next to this aspect, G/NG models are valuable for the adaptation of food recipes and/or the development of new food products, i.e., these models enable to define whether the new product formulation will support the outgrowth of specific food pathogens and/or food spoilage organisms.

Growth/no growth models are empirical models which do not include any mechanistic information about the underlying relationship between the microbial behavior and the chemical and physical food characteristics. Their transferability is specifically limited because of three factors. (*i*) The general mechanisms behind the synergistic relation between different environmental conditions are not yet fully understood. As a result, G/NG models are only valid for the specific environment and microorganism for which they have been constructed. (*ii*) G/NG boundaries are known to depend on the initial inoculation level. A decrease in the inoculum size lowers the growth probability. Possibly, this can be explained by the distribution of physiological cell states as is observed for initial lag times. (*iii*) The predicted response is highly related to the time span considered for defining the observed growth or no growth. When times beyond the experimental range are considered, growth at non-supporting conditions or a higher growth percentage can be observed as a result of the resuscitation of injured cells. When selecting the experiment duration, a trade-off has to be made between practical feasibility and a realistic

shelf life.

Due to the purely data-driven approach of G/NG models, the quality of the experimental data is of utmost important, i.e., the accuracy of the G/NG boundary is determined by the number of experimental data collected. Especially at conditions approaching the G/NG boundary, a higher number of repetitions will yield a more accurate approximation of the growth probability. As a consequence, highly accurate models require a demanding experimental scheme, e.g., a 5% growth probability requires at least 20 replicates. In practice, the experimental scheme selected should include a balance between practicality, model accuracy, and what is relevant for the food industry.

The importance of specific properties of probabilistic G/NG models and the relation with their validity is illustrated using case studies relevant for food safety and food quality. For instance, the results obtained by Mertens et al. (2010) show that the experimental set-up taken has a significant effect on the outcome of the experimental study and thus the overall model validity. This study focuses on the growth potential of the spoilage yeast *Zygosaccharomyces bailii* in a solid(like) model system that resembles acidic sauces. Next to the effect of the solid environment, also the effect of pH, glycerol and acetic acid was investigated. Growth was defined by optical density. Depending on the location where optical density is measured, a colony might be locally present or not, which makes it difficult to make an overall and objective distinction between growth and no growth. In this case, the initial cell number and the percentage of cells that is able to survive the conditions and finally grow will determine the outcome of the optical density measurements.

4 Conclusion

Overall, how closely the model approximates microbial dynamics during the selected process and time span is highly determined by the model accuracy. As both the model building procedure, and the subsequent parameter estimation and model validation process, are data-driven steps, careful selection of the experimental set-up and scheme is of high importance. The model extrapolation region and the accuracy of the prediction are mainly determined by the experimental region tackled and the number of experiments that are combined. By applying statistically based mathematical techniques, experiments relevant for food decontamination processes can be selected to finally optimize the model building procedure and improve the parameter estimation accuracy.

5 Acknowledgements

This research is supported in part by projects OT/09/25 and PFV/10/002 (OPTEC Optimization in Engineering) of the Research Council of the KULeuven, project KP/09/005 (SCORES4CHEM) of the KULeuven Industrial Research Fund, and the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian Federal Science Policy Office. E. Van Derlinden was supported by the postdoctoral grant PDMK/10/122 of the KULeuven Research Fund. J.F. Van Impe holds the chair Safety Engineering sponsored by the Belgian Chemistry and Life Sciences Federation essenscia.

References

Mertens, L., Van Derlinden, E., & Van Impe, J. F. (2012) Comparing experimental design schemes in predictive food microbiology: optimal parameter estimation of secondary models. *Submitted*.

Mertens, L., Van Derlinden, E., Dang, T. D. T., Cappuyns, A. M., Vermeulen, A., Debevere, J., Moldenaerts, P., Devlieghere, F., Geeraerd, A. H., & Van Impe, J. F. (2010) On the critical evaluation of growth/no growth assessment of *Zygosaccharomyces bailii* with optical density measurements: Liquid versus structured media. *Food Microbiology*, 28, 736-745.

Van Derlinden, E., Bernaerts, K., & Van Impe, J. F. (2008) Accurate estimation of cardinal growth temperatures of *Escherichia coli* from optimal dynamic experiments. *International Journal of Food Microbiology*, 128, 89-100.

Van Derlinden, E., Bernaerts, K., & Van Impe, J. F. (2010) Simultaneous versus sequential optimal experiment design for the identification of multi-parameter microbial growth kinetics as a function of temperature. *Journal of Theoretical Biology*, 264, 3q47-355.

Van Derlinden, E., & Van Impe, J. F. (2012) Modeling microbial kinetics as a function of temperature: Evaluation of dynamic experiments to identify the growth/inactivation interface. *Journal of Food Engineering*, 108, 201-210.

Contrôle non paramétrique de procédés par lots basé sur STATIS et la classification

Non parametric on line control of batch processes based on STATIS and clustering

Ndèye Niang[1], Gilbert Saporta[1], Flavio S. Fogliatto[2]

[1] Chaire de Statistique Appliquée & CEDRIC CNAM
292, rue Saint Martin, 75141 Paris Cedex 03, France,
ndeye.niang_keita@cnam.fr
gilbert.saporta@cnam.fr

[2] ffogliatto@producao.ufrgs.br

Résumé
Nous proposons une nouvelle approche du contrôle de qualité des procédés par lots basée sur la méthode STATIS et des cartes de contrôles non paramétriques à partir d'enveloppes convexes. Cette approche générale peut être utilisée pour le contrôle en fin de fabrication des procédés par lots ainsi que pour le contrôle en cours de fabrication après une étape de classification. La méthode proposée est illustrée sur des données réelles.
Mots-clés : Procédés par lots, Classification, Contrôle de qualité multivarié, STATIS.

Abstract
We propose a new non parametric quality control strategy for monitoring batch processes based on the three way method STATIS and convex hull peeling. This general approach allows off line monitoring of batch processes as well as on line one after a clustering step. A real example illustrates the proposed method.
Keywords : Batch process, Clustering, Multivariate quality control, STATIS

1. Introduction

Les procédés par lots sont largement utilisés dans le secteur industriel notamment dans l'industrie agroalimentaire, chimique ou pharmaceutique. Dans ces procédés, les matières premières sont introduites dans un ordre spécifique et subissent une série de transformations pendant une durée qui peut être fixe ou variable donnant alors lieu à des procédés à temps fixe ou à temps variable. Le produit final obtenu est ensuite analysé pour vérifier s'il correspond à des standards de qualité désirés. Le suivi du procédé s'effectue à travers un ensemble de variables caractéristiques du procédé prélevées par un échantillonnage en ligne au fur et à mesure de son déroulement. Les données se présentent sous la forme d'un tableau à trois entrées ou « cube » de données. Due à la nature multidimensionnelle des données issues de tels procédés, les cartes de contrôle multivariées sont alors les seules adéquates pour le contrôle de leur qualité.

La carte multivariée la plus fréquemment utilisée est la carte T^2 de Hotelling (Lowry& Montgomery (1995)). En général, ces cartes de contrôle sont basées sur l'hypothèse d'indépendance des observations et de multinormalité des caractéristiques du procédé. Mais dans la pratique ces hypothèses ne sont pas toujours vérifiées. De plus les cartes de contrôle classiques ne permettent pas un contrôle efficace lorsque les standards de qualité sont décrits par des profils ou courbes. Dans le cas de tels procédés, le contrôle s'effectue à travers des cartes multivariées basées sur une analyse en composantes principales particulière (multiway principal component analysis) Nomikos & MacGregor (1995). Ces cartes seront notées MPCA-CCs dans la suite.
L'application des MPCA-CCs pour le monitoring des procédés par lots a été initialement proposé par Jackson & Mudhokar (1979), et largement étudiée par la suite par Nomikos & MacGregor (1995), Kourti & MacGregor (1996) et MacGregor (1997). Elle suppose, en plus de la normalité des variables, que tous les lots aient la même durée et ne peut donc pas être directement utilisée pour le contrôle des procédés à temps variable, ni pour le contrôle de procédés par lots en cours de fabrication. De nombreuses méthodes ont été proposées pour adapter les MPCA-CCs aux cas cités ci-dessus: Nomikos (1995), Kassidas *et al*. (1998), Doan & Srinivasan (2008). Elles peuvent être globalement considérées comme des méthodes de prétraitement dont le but est de donner la même longueur à tous les lots afin d'appliquer ensuite les MPCA-CCs classiques. Cependant ces méthodes présentent toutes quelques limitations (Niang et al. 2009).
Une approche générale basée sur la méthode STATIS a été proposée (Niang et al. 2009) permettant le contrôle en fin de fabrication des procédés par lots à temps fixe (avec STATIS) et à temps variable (avec STATIS DUAL) sans aucun traitement préalable des données ni hypothèse sur la distribution des variables. Elle consiste d'abord à utiliser la méthode STATIS pour réduire la dimension des données puis à construire des cartes de contrôles non paramétriques à partir des enveloppes convexes obtenues directement sur les plans factoriels issus de l'application de la méthode STATIS.
Nous nous intéressons au contrôle en cours de fabrication qui consiste à suivre le procédé au fur et mesure de son déroulement pour détecter le plus tôt possible une sortie des limites de contrôle plutôt que d'attendre la fin du lot. Il s'agit donc de vérifier le comportement du procédé à chaque instant noté t. L'application de l'approche décrite précédemment permet d'établir une distribution de référence pour le comportement du procédé jusqu'à l'instant t ou de manière équivalente une carte de contrôle sur des tableaux partiels obtenus en sélectionnant les t premières lignes des tableaux de données. En principe il faudrait autant de cartes de contrôle que d'instants de mesures. La dernière carte est identique à la carte pour le contrôle off line. Mais en pratique ne sont intéressantes que celles qui correspondent à des instants de changement important dans l'évolution du procédé, ces instants définissent une partition de l'ensemble des instants de mesures.
Dans cet article, nous proposons une approche basée sur la classification sous contrainte de contiguïté pour déterminer ces instants. Après un rappel sur les cartes de contrôle non paramétriques basées sur STATIS, nous présentons dans la section 3 notre proposition pour le contrôle on line. La méthode proposée est ensuite illustrée sur des données réelles d'un procédé à temps fixe.

2. Cartes de contrôle non paramétriques basées sur STATIS

On dispose d'un historique de *N* lots de référence c'est à dire des lots ayant donné un produit de bonne qualité définissant ainsi une distribution de référence représentant le bon fonctionnement du procédé. Dans le cas des procédés à temps fixe, les données se présentent sous la forme de *N* tableaux à *p* variables prélevées à *T* instants (figure 1). On est donc en présence de plusieurs tableaux décrivant un ensemble d'individus sur *p* variables. Il est alors possible de les analyser directement sans aucun traitement préalable en utilisant la méthode STATIS. Nous expliquons plus en détail notre méthode de contrôle de qualité après avoir rappelé brièvement la méthode STATIS.

	LOT 1	*LOT 2*		*LOT N*	
X =	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_1**	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_2**	……..	*VARIABLES* $x_1,x_2,\ldots,x_p$ **X_N**	*instant 1* *instant 2* . . *instant T*

Figure 1- Matrice des données

2.1 STATIS

STATIS est une méthode d'analyse exploratoire simultanée de plusieurs tableaux de données recueillies à différentes occasions Escoufier(2006). A notre connaissance son utilisation en contrôle de qualité se limite aux travaux de Scepi (2002).
L'idée essentielle est la recherche d'une structure commune aux tableaux pour voir si les distances entre individus sont stables d'un tableau à l'autre. Elle fonctionne en trois étapes. D'abord on effectue une analyse globale dans laquelle on cherche à comparer la structure des tableaux sans pouvoir donner une explication fine des éventuelles différences entre tableaux. Cette étape est appelée *interstructure*.
L'étude fine s'effectue dans la deuxième analyse appelée *intrastructure*. Elle repose sur la détermination d'un résumé global des tableaux appelé compromis qui permet de trouver un espace commun de représentation. L'étude de l'évolution de chacun des individus des tableaux sur cet espace de représentation permet d'expliquer au niveau individuel les écarts mis en évidence par l'interstructure.
Plus précisément, on dispose de $\mathbf{X}_i$ (i = 1,…, N) matrices contenant T observations de p variables. Préalablement à l'analyse, les données sont centrées réduites. STATIS associe à chaque $\mathbf{X}_i$ la matrice ($T \times T$) des produits scalaires entre individus $\mathbf{W}_i = \mathbf{X}_i\mathbf{X}_i^{'}$, où $\mathbf{X}_i^{'}$ est la matrice transposée de $\mathbf{X}_i$. C'est un objet représentatif de $\mathbf{X}_i$; il contient tous les liens inter-individus et ses vecteurs propres sont les composantes principales de $\mathbf{X}_i$. Pour comparer deux tableaux $\mathbf{X}_i$ et $\mathbf{X}_{i'}$, on utilise le coefficient *RV* de corrélation vectorielle, Escoufier (2006) défini par:

$$RV_{ii'} = \mathrm{trace}(\mathbf{W}_i\mathbf{W}_{i'}) / \sqrt{\mathrm{trace}(\mathbf{W}_i)^2\,\mathrm{trace}(\mathbf{W}_{i'})^2} \qquad (1)$$

RV varie entre 0 et 1; plus il est proche de 1, plus les deux matrices $\mathbf{W}_i$ et $\mathbf{W}_{i'}$ sont similaires.

2.1.1 Interstructure

L'interstructure consiste à étudier graphiquement les ressemblances globales entre tableaux. Comme en ACP, les deux premiers vecteurs propres de la matrice S contenant les coefficients *RV* entre $\mathbf{W}_i$ et $\mathbf{W}_{i'}$ (i ,i'= 1,…, N) définissent le premier plan principal ce qui permet de visualiser les proximités

entre tableaux en y projetant les objets $\mathbf{W}_i$: la coordonnée du tableau $\mathbf{W}_i$ sur le *kième* axe factoriel est donnée par $c_i^k = \sqrt{\lambda_k} u_i^k$ où λ_k est la valeur propre associée au *kième* vecteur propre $\mathbf{u}^k$. Les coefficients *RV* étant positifs, $\mathbf{u}^1$ a ses composantes toutes de même signe, elles seront prises positives.

2.1.2 Intrastructure

L'étude de l'intrastructure consiste d'abord à rechercher le compromis qui résume au mieux l'ensemble de tableaux. La solution est une moyenne pondérée des objets $\mathbf{W}_i$ les coefficients α_i^1 étant les composantes du premier vecteur propre $\mathbf{u}^1$ normalisé:

$$\mathbf{W} = \sum_{i=1}^{N} \alpha_i^1 \mathbf{W}_i \qquad (2)$$

Les poids α_i^1 représentent alors le niveau d'accord entre les tableaux et le compromis. Cette définition du compromis confère à STATIS une propriété de robustesse vis à vis des valeurs aberrantes: plus un tableau est différent des autres, moins il a d'influence sur le compromis. Cette propriété est particulièrement intéressante en contrôle de qualité dont le but est la détection de valeurs anormales.

Les vecteurs propres de la matrice compromis $\mathbf{W}$ permettent ainsi d'obtenir l'espace de représentation commun à l'ensemble des tableaux. Il est alors possible de visualiser sur le premier plan principal des points artificiels B_t (t =1,…, T) appelés points compromis. Les coordonnées sur le *k-ième* axe factoriel sont les éléments du vecteur suivant:

$$\mathbf{z}_k = \sqrt{\delta_k}\,\mathbf{v}^k = (1/\sqrt{\delta_k})\mathbf{W}\mathbf{v}^k \qquad (3)$$

où δ_k est la valeur propre associée au *k-ième* vecteur propre $\mathbf{v}^k$.

De plus, il est possible de représenter les individus de tous les tableaux $\mathbf{W}_i$ en les projetant sur le plan compromis par la technique des points supplémentaires. Les différentes positions d'un individu selon les tableaux définissent sa trajectoire qui permet de mettre en évidence des écarts entre tableaux au niveau individuel. On peut donc avoir une représentation détaillée du comportement commun des lots à un instant donné.

2.2 Cartes de contrôle non paramétriques

Les cartes de contrôle non paramétriques que nous proposons sont basées sur des enveloppes convexes directement construites sur les plans principaux de l'interstructure de STATIS. Pour établir une région de contrôle de confiance $(1-\alpha)$ on utilise une proposition de Zani *et al.* (1998), comprenant les trois étapes suivantes :

- on détermine sur le plan de l'interstructure une région intérieure qui contient une proportion π^* des points. Elle est obtenue par lissage par une *B*-spline des contours de l'enveloppe convexe contenant les points, cette dernière étant obtenue par pelages successifs de l'enveloppe convexe contenant l'ensemble des points.

---

* π est égale à 50% des points dans Zani *et al*. (1998), mais on peut utiliser une plus grande proportion.

- Ensuite on détermine une estimation du centre de la région en prenant par exemple la moyenne arithmétique des observations dans la région.

- Finalement, la carte de contrôle est obtenue par dilatation de l'enveloppe lissée en multipliant la distance entre le centre et frontière de la région par un nombre *l* correspondant à la probabilité α de fausse alarme désirée.

Avec cette méthode, à partir des *N* lots de référence associés à *N* tableaux de dimension *T* x *p*, on obtient une carte de contrôle non paramétrique. Le but du contrôle en fin de fabrication est de vérifier la conformité des données d'un nouveau lot représenté par la matrice $\mathbf{X}_{N+1}$ avec des standards de qualité résumés, représentés par la carte de contrôle issue des lots de référence. Le contrôle du nouveau lot s'effectue alors en projetant la matrice $\mathbf{X}_{N+1}$ sur la carte. Le lot sera déclaré sous contrôle si la matrice se projette à l'intérieur de la région de contrôle. Dans le cas contraire, le lot sera hors contrôle. La section 4 présente les résultats de l'application de cette méthode à des données issues d'un procédé de polymérisation.

3- Contrôle on line

Rappelons qu'il consiste à suivre le procédé au fur et mesure de son déroulement pour détecter le plus tôt possible une sortie des limites de contrôle plutôt que d'attendre la fin du lot. La démarche que nous proposons consiste à établir, selon la méthode décrite en section 2, de manière séquentielle des cartes de contrôle non paramétriques basées sur STATIS appliquée à des tableaux partiels issus des tableaux de référence. Elle comporte donc une étape préalable de détermination de la longueur des séquences ou de la taille des tableaux partiels que nous proposons de réaliser à partir d'un partitionnement de l'ensemble des instants de mesures.

Nous disposons de *N* matrices $\mathbf{X}_i$ contenant *T* observations de *p* variables. Elles représentent le comportement de référence du procédé produisant des lots de bonne qualité. Le problème est donc de trouver une partition *P* des *T* instants de mesure commune à l'ensemble des lots avec une contrainte de conservation de la chronologie.

L'application d'une classification ascendante hiérarchique sous contrainte de contiguïté temporelle Murtagh (1985) sur chaque lot permet d'obtenir un ensemble de *N* partitions des *T* instants en *K* classes. La variabilité des lots de référence peut entraîner une variabilité dans les tailles des classes: la classe 1 d'une partition P_i peut contenir les instants de 1 à t alors que la classe 1 de la partition P_j contient les instants de 1 à t-1. Plus formellement, soit n_{ik} la taille de la classe k de la partition *Pi* associé au lot i, les instants de contrôle associés au lot i et notés t_{ik}, sont définis par $t_{ki} = \sum_{l=1}^{k} n_{il}$ avec k variant de 1 à K.

En considérant l'ensemble des N lots de référence, on obtient pour chaque k un ensemble de *N* valeurs t_{ki} (i= 1,…, *N*) qui peut être assimilé à une période « critique » pendant laquelle il faudrait surveiller le procédé. Nous proposons de choisir comme instants pour le contrôle on line les valeurs $t_k = \sup_i t_{ik}$.

Cela revient à effectuer le contrôle au dernier instant de la période critique augmentant ainsi la probabilité de détection des lots hors contrôle. En effet, il est usuel en contrôle de qualité de supposer que si une cause assignable produit un déréglage des caractéristiques du procédé, ce dernier persiste. Effectuer le contrôle à l'instant t_k permet donc de détecter un plus grand nombre de lots qui ont pu être déréglés antérieurement à t_k . Lorsque les instants de mesure sont proches les uns des autres, cela

n'affectera pas beaucoup la période opérationnelle moyenne. Dans le cas contraire d'autres stratégies prenant en compte l'écart entre les instants de mesures devraient être considérées.
En appliquant la méthode proposée en 2.2 aux K ensembles de *N* tableaux obtenus en sélectionnant successivement les t_k premiers instants des tableaux de référence, on construit alors K cartes de contrôle. Plus précisément pour chaque instant de contrôle, on applique STATIS aux N tableaux à t_k individus et on obtient la carte de contrôle non paramétrique à partir du plan factoriel représentant l'interstructure.
Le contrôle d'un nouveau lot en cours de fabrication consiste ensuite à projeter les tableaux de taille t_k associés au lot en cours de fabrication sur les cartes correspondantes.

4- Application

Nous illustrons notre proposition sur des données réelles utilisées dans la littérature des procédés par lots par Nomikos, Mc Gregor ou Eriksson et al. par exemple. Les données sont issues d'un procédé de polymérisation et sont composées de 18 lots de référence sélectionnés comme représentant le comportement normal souhaité du procédé. Pour chaque lot, 10 variables ont été prélevées à 100 instants. Les variables x_1 x_2 x_3 x_6 et x_7 sont des mesures de température, x_4 x_8 et x_9 sont des mesures de pression et les variables x_5 et x_{10} représentent des vitesses d'écoulement de matières ajoutées au réacteur. On dispose de plus d'un ensemble de 11 lots supplémentaires pour tester les performances des méthodes. Il contient 4 lots de bonne qualité et 7 mauvais lots. Nous avons appliqué STATIS et les régions de contrôle avec un niveau de confiance de 99% sont construites sur les plans factoriels de l'interstructure.
La figure 2 montre les résultats du contrôle en fin de fabrication. Tous les 7 mauvais lots ont été signalés hors contrôle avec cependant un plus fort signal pour 6 d'entre eux (fig.2.a). Le mauvais lot proche de la limite a été diagnostiqué comme ayant un comportement différent des 6 autres et n'est en général pas détecté comme étant hors contrôle (Eriksson *et al*). La carte de contrôle (fig.2.b) montre les résultats pour les bons lots. 3 bons lots parmi les 4 sont signalés sous contrôle. On constate cependant une fausse alerte comme dans Eriksson et al.

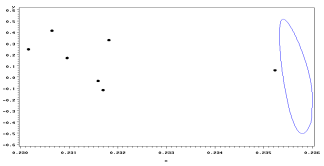


Fig. 2.a. Mauvais lots

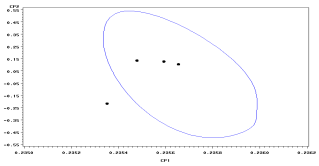


Fig. 2.b. Bons lots

Conclusion

Nous avons proposé une méthode pour le contrôle de qualité des procédés par lots en fin et en cours de fabrication basée sur la méthode STATIS. Le suivi du procédé est effectué à travers des cartes de contrôles non paramétriques utilisant toutes les observations disponibles pour le contrôle en fin de fabrication, et une partie des observations séquentiellement pour le contrôle en cours de fabrication. La méthode est illustrée sur des données réelles.
Des évaluations plus formelles des performances de la méthode sont en cours ainsi que des études comparatives avec d'autres méthodes proposées dans la littérature.

Bibliographie

Doan, X-T., Srinivasan, R. (2008) Online monitoring of multi-phase batch processes using phase-*based* multivariate statistical control. *Computers and Chemical Engineering*, 32: 230-243

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S.(2001).*Multi- and Megavariate Data Analysis.* Umetrics

Escoufier, Y. 2006. Operator related to a data matrix: a survey. *Proceedings in Computational Statistics* Rizzi A. *et al*. (eds), 285-297 Physica-Verlag.

Jackson, J.E. and Mudholkar, G.S. (1979) Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21 (3), 341–34.

Kassidas, A., MacGregor, J.F. and Taylor, P.A. (1998) Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44, 864–875.

Kourti, T. and MacGregor, J.F. (1996) Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology*, 28 (4), 409–428.

Lowry, C.A. and Montgomery, D.C. (1995) A review of multivariate control charts. *IIEE Transactions*, 27 (6), 800–810.

MacGregor, J.F. (1997) Using on-line process data to improve quality: challenges for statisticians. *International Statistical Review*. 65 (3), 309–323.

Murtagh, F. (1985) A Survey of Algorithm for Contiguity-constrained clustering and Related problems. The computer journal, 28(1), 82-88

Niang, N., Fogliatto F. and Saporta, G. (2009) Batch Process Monitoring by Three-way Data Analysis Approach, *ASMDA'09*, Vilnius, July 2009, pp.294-298

Nomikos, P. and MacGregor, J.F. (1995) Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1), p.41–59.

Scepi, G. (2002) Parametric and non parametric multivariate quality control charts. In *Multivariate Total Quality Control*, Physica-Verlag , Lauro C. *et al*. (eds), 163–189.

Zani, S., Riani, M. and Corbellini, A. (1998) Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28, 257-270.

Amélioration de la puissance du test de la capabilité d'un processus ne possédant qu'une limite de tolérance en présence d'erreurs de mesure

Improvement of the value of the capability test for a one-sided process with measurement errors

Daniel Grau

IUT de Bayonne, 17 place Paul Bert, 64100 Bayonne, CNRS UMR 5142
E-mail : daniel.grau@univ-pau.fr

Résumé
Dans l'industrie, il est assez fréquent qu'une seule tolérance soit définie pour la caractéristique d'intérêt d'un produit. Dans cette situation les indices $C_p^u(u,v)$ et $C_p^l(u,v)$ peuvent être utilisés pour mesurer la performance du processus. L'obtention de ces indices est cependant soumise à deux types d'incertitude. L'incertitude due au choix de l'échantillon et l'incertitude due aux erreurs de mesure. Si la littérature concernant les indices de capabilité prend systématiquement en compte la première incertitude, il est rare que ce soit le cas pour la seconde. Or négliger les erreurs de mesure inhérentes à tout processus conduit à sous estimer la capabilité du processus et donc à des tests statistiques rejetant des processus capables. Afin d'améliorer la puissance du test nous proposons d'utiliser une valeur critique ajustée. Un exemple tiré de l'industrie agro-alimentaire illustre la méthode proposée.
Mots-clés : indices de capabilité, erreurs de mesure, tolérance unique
Abstract
In the manufacturing industry, it is quite common that only one tolerance should be defined for the characteristic of interest of a product. In that situation $C_p^u(u,v)$ and $C_p^l(u,v)$ indices can be used to measure the process performance. However, these indices undergo two types of uncertainty. The uncertainty due to the choice of the sample and the one due to measurement errors. If the literature concerning the capability indices systematically takes into account the first uncertainty, it is unusual that this happens for the second. Yet, ignoring the measurement errors inherent in any process leads to underestimate the process capability and thus statistical tests rejecting capable processes. To improve the power of the test we suggest the use of an adjusted critical value. An example from the food industry illustrates the proposed method.
Keywords : capability indices, measurement errors, one-sided tolerance

1. Introduction

Les indices de capabilité qui permettent de mesurer la performance d'un processus en fonction des contraintes fixées par le cahier des charges sont largement utilisés dans l'industrie. Le premier indice

C_p mesure la capabilité potentielle du processus indépendamment de sa position moyenne dans l'intervalle de tolérance. Cette position est prise en compte par l'indice C_{pk} introduit par Kane (1986). L'indice C_{pm} proposé par Chan, Cheng and Spiring (1988) prend en compte la déviation du processus par rapport à la cible. Combinant les indices précédents, Pearn, Kotz and Johnson (1992) proposent l'indice C_{pmk}. Enfin, pour un processus ayant des limites de tolérance inférieure *LSL* et supérieure *USL*, et une cible *T* située au milieu de l'intervalle de tolérance *m* = (*LSL* + *USL*)/2, Vännman (1995) propose une famille d'indices incluant les indices usuels sous la forme suivante,

$$C_p(u,v) = \frac{d - u|\mu - m|}{3\sqrt{\sigma^2 + v(\mu - T)^2}},$$

où $d = (USL - LSL)/2$ représente la moitié de l'intervalle de tolérance, μ est la moyenne, σ est l'écart-type, et *u* et *v* sont deux paramètres positifs ou nuls. Bien que les cas de tolérances symétriques ($T = m$) soient courants dans l'industrie, les situations où les tolérances ne sont pas symétriques ($T \neq m$) sont rencontrées assez fréquemment. Ceci se produit, soit lorsqu'une déviation est considérée comme plus grave dans une direction que dans la direction opposée, soit en présence de tolérances symétriques au départ, mais avec des données non gaussiennes nécessitant une transformation rendant ainsi les tolérances asymétriques. Dans ce cas Chen et Pearn (2001) proposent d'utiliser la famille

$$C''_p(u,v) = \frac{d^* - uA^*}{3\sqrt{\sigma^2 + vA^2}},$$

où $A = \max\{d(\mu - T)/D_u, d(T - \mu)/D_l\}$, $A^* = \max\{d^*(\mu - T)/D_u, d^*(T - \mu)/D_l\}$, $D_u = USL - T$, $D_l = T - LSL$, et $d^* = \min\{D_u, D_l\}$.

Dans l'étude qui suit nous nous intéressons à la situation particulière où le risque d'une déviation du processus dans une direction est considéré comme si peu important par l'utilisateur qu'il est amené à ne définir qu'une seule limite de tolérance. Pour pouvoir mesurer la performance du processus par un indice similaire à ceux utilisés dans le cas de deux tolérances, Grau (2009) propose que l'utilisateur quantifie approximativement ce risque. Si le risque est considéré comme *k* fois moins grave dans la direction opposée à la borne de tolérance, Grau (2009) propose d'utiliser les familles

$$C_p^u(u,v) = \frac{D_u - uA_u^*}{3\sqrt{\sigma^2 + vA_u^{*2}}}, \text{ et } C_p^l(u,v) = \frac{D_l - uA_l^*}{3\sqrt{\sigma^2 + vA_l^{*2}}}, \quad (1)$$

pour des bornes de tolérance supérieure et inférieure, avec $A_u^* = \max(\mu - T, (T - \mu)/k)$, et $A_l^* = \max((\mu - T)/k, T - \mu)$. Notons que la lettre *u* utilisée en exposant est l'abréviation de 'upper', et est donc indépendante du premier des deux paramètres (u,v) de la famille d'indices. Le choix de la constante k (≥ 1) étant approximatif, les indices $C_p^u(u,v)$ et $C_p^l(u,v)$ ont été construits de telle sorte qu'ils soient indépendants de k lorsque la moyenne dévie vers la limite de tolérance. Le choix k = 1 revient à considérer comme identiques les risques de déviation du processus à droite et à gauche de la cible. Si le risque d'une déviation est considéré comme nul, à gauche de la cible par exemple, il suffit de donner une valeur infinie à k. Dans ce cas, lorsque $\mu < T$ nous avons $C_p^u(u,v) = D_u/(3\sigma)$ qui est indépendant de μ, et qui n'est rien d'autre que la capabilité potentielle C_p. Pour $C_p^u(u,v)$ et $C_p^l(u,v)$ Grau (2011a) propose de choisir les paramètres *u* et *v* en fonction de l'importance que l'utilisateur accorde au centrage du processus et/ou à la proportion de non conformes.

Les indices de capabilité calculés à partir d'échantillons ne sont pas les véritables valeurs mais seulement des estimations, ce qui entraine une première source d'incertitude. Une deuxième source d'incertitude, souvent ignorée, provient des erreurs de mesure.

Dans l'étude qui suit nous nous contentons de développer les résultats relatifs à $C_p^l(u,v)$, ceux concernant $C_p^u(u,v)$ s'en déduisant facilement. Au paragraphe 2 nous mettons en évidence l'effet des erreurs de mesure sur $C_p^l(u,v)$. Au paragraphe 3 nous donnons la distribution d'échantillonnage de cet indice. Les paragraphes 4 et 5 sont consacrés à la détermination des valeurs critiques et des valeurs critiques ajustées lorsque des erreurs de mesure sont présentes. Enfin un exemple réel dans l'industrie agro-alimentaire illustre les résultats obtenus dans le dernier paragraphe.

2. Effet des erreurs de mesure

Mittag (1997) est le premier auteur à mettre en évidence les effets des erreurs de mesure sur les indices de capabilité. Les erreurs de mesure étant considérées comme une variable aléatoire $M \sim N(0,\sigma_M^2)$, Mittag (1997) définit le degré d'erreur de contamination par

$$\tau = \frac{\sigma_M}{\sigma}.$$

Soit $X \sim N(\mu,\sigma^2)$ la caractéristique d'intérêt du processus. En fait la variable G (avec les erreurs de mesure) est observée et non la véritable variable X. On admet de plus que X et M sont liées additivement, $G = X + M$, et que X et M sont stochastiquement indépendantes. Nous avons donc $G \sim N(\mu,\sigma_G^2 = \sigma^2 + \sigma_M^2)$ et l'indice de capabilité $C_p^{lG}(u,v)$ est obtenu en substituant σ_G à σ dans l'équation (1).

$$C_p^{lG}(u,v) = \frac{D_l - uA_l^*}{3\sqrt{\sigma_G^2 + vA_l^{*2}}}.$$

La relation entre $C_p^l(u,v)$, la véritable capabilité du processus, et $C_p^{lG}(u,v)$, la capabilité empirique du processus, s'exprime sous la forme

$$C_p^{lG}(u,v) = \frac{\sqrt{1+v\xi^{l2}}}{\sqrt{1+\tau^2+v\xi^{l2}}} C_p^l(u,v) \qquad (2)$$

où $\xi^l = \max(\xi/k, -\xi)$ et $\xi = (\mu - T)/\sigma$. Il est clair que le ratio $C_p^{lG}(u,v)/C_p^l(u,v)$ est une fonction décroissante de τ. Donc les erreurs de mesure conduisent à une sous-estimation de la véritable capabilité du processus.

3. Distribution d'échantillonnage de $C_p^{lG}(u,v)$

Une pratique courante dans l'industrie pour estimer la capabilité d'un processus consiste, lorsque le processus est considéré comme stable, à prélever r échantillons $(G_{i1}, G_{i2}, ..., G_{in_i})$ de taille variable n_i, qui sont supposés suivre une loi $N(\mu,\sigma_G^2)$. Soit $\bar{G}_i = \sum_{j=1}^{n_i} G_{ij}/n_i$ et $S_{G_i} = \left[n_i^{-1} \sum_{j=1}^{n_i} (G_{ij} - \bar{G}_i)^2 \right]^{1/2}$ la moyenne et l'écart-type du i^{ème} échantillon, et $N = \sum_{i=1}^{r} n_i$ le nombre total d'observations. Nous considérons l'estimateur naturel de $C_p^{lG}(u,v)$ suivant

$$\hat{C}_p^{lG}(u,v) = \frac{D_l - u\hat{A}_l^{*G}}{3\sqrt{S_G^2 + v\hat{A}_l^{*G^2}}},$$

où $\hat{A}_l^{*G} = \max\left(\left(\bar{\bar{G}} - T\right)/k, T - \bar{\bar{G}}\right)$, $\bar{\bar{G}} = \sum_{i=1}^{r} n_i \bar{G}_i / N$, et $S_G^2 = \sum_{i=1}^{r} n_i S_{G_i}^2 / N$. Appliquant la même technique utilisée par Pearn, Lin et Chen (2001) pour obtenir la fonction de répartition de $\hat{C}''_{pmk} = \hat{C}''_p(1,1)$, pour $(u,v) \neq (0,0)$ nous obtenons

$$F_{\hat{C}_p^{lG}(u,v)}(x) = 1 - \int_0^{KG(x)} H_G(x,t)dt \text{ , pour x > 0,} \tag{3}$$

où $K_G(x) = 3\sqrt{N}C_p^{lG}(0,0)\Big/\left(u + 3x\sqrt{v}\right)$,

$H_G(x,t) = F_K\left(\left(\frac{3\sqrt{N}C_p^{lG}(0,0) - ut}{3x}\right)^2 - vt^2\right) f_{Z_l^G}(t)$, avec $F_K(x)$ la fonction de répartition d'une χ^2_{N-r},

$f_{Z_l^G}(t) = k\phi\left(kt - \sqrt{N}\xi_G\right) + \phi\left(t + \sqrt{N}\xi_G\right)$, où $\xi_G = (\mu - T)/\sigma_G$ et $\phi(x)$ représente la densité de probabilité d'une loi N(0,1).

4. Test de capabilité

Nous considérons les hypothèses de test suivantes :

$H_0 : C_p^l(u,v) \leq c$ Le processus n'est pas capable,

$H_1 : C_p^l(u,v) > c$ Le processus est capable.

Si l'estimation $\hat{c}_p^l(u,v)$ de la capabilité $C_p^l(u,v)$ est plus grande que la valeur critique c_0, nous rejetons l'hypothèse nulle et concluons que le processus est capable avec une erreur α. c et α étant donnés, la valeur critique c_0 est obtenue en résolvant l'équation $\alpha = P\left(\hat{C}_p^l(u,v) > c_0 \mid C_p^l(u,v) = c\right)$, et la puissance du test peut être calculée à partir de l'équation $\pi(C_p^l(u,v)) = P\left(\hat{C}_p^l(u,v) > c_0 \mid C_p^l(u,v)\right)$. Lorsque nous sommes en présence d'erreurs de mesure, ce n'est pas $\hat{c}_p^l(u,v)$ qui est calculé, mais $\hat{c}_p^{lG}(u,v)$. Dans ces conditions, le niveau et la puissance du test que nous notons α_G et π_G sont définis par $\alpha_G = P\left(\hat{C}_p^{lG}(u,v) > c_0 \mid C_p^l(u,v) = c\right)$ et $\pi_G(C_p^l(u,v)) = P\left(\hat{C}_p^{lG}(u,v) > c_0 \mid C_p^l(u,v)\right)$. Les discussions précédentes ont montré que nous sous-estimons la véritable capabilité si nous utilisons $\hat{C}_p^{lG}(u,v)$ au lieu de $\hat{C}_p^l(u,v)$. La probabilité que $\hat{C}_p^{lG}(u,v)$ soit plus grand que c_0 sera moins grande que celle obtenue en utilisant $\hat{C}_p^l(u,v)$. Donc α_G sera plus petit que α et π_G sera plus petit que π. Pour illustrer les variations de π_G en fonction du degré de contamination τ, nous avons considéré le cas particulier u = 0.5, v = 1.5, k = 3 et à l'aide du logiciel Maple, tracé les courbes π_G pour différentes valeurs de $C_p^l(0,0)$ et $C_p^l(u,v)$. La Figure 1 représente π_G pour $\tau \in [0,1]$ avec r = 1, N = 50, $\alpha = 0.05$, $C_p^l(u,v)$ = $c(0.20)(c + 1)$, c = 1.00 et $C_p^l(0,0)$ = $C_p^l(u,v)$ + 0.33, puis c = 1.50 et $C_p^l(0,0)$ = $C_p^l(u,v)$. Nous constatons que π_G décroît lorsque τ croît et que le taux de décroissance est plus important pour les grandes valeurs de c. La présence d'erreurs de mesure peut avoir un effet très important sur π_G. Par exemple, pour c = 1.5 et $C_p^l(0,0)$

= $C_p^l(u,v)$ = 2.3 (Figure 1.b), π_G est à peu près égal à 1 lorsqu'il n'y a pas d'erreurs de mesure. Cependant, quand τ = 1, π_G est à peu près égal à 0.41 pour $\xi > 0$, et à peu près égal 0.15 pour $\xi < 0$.

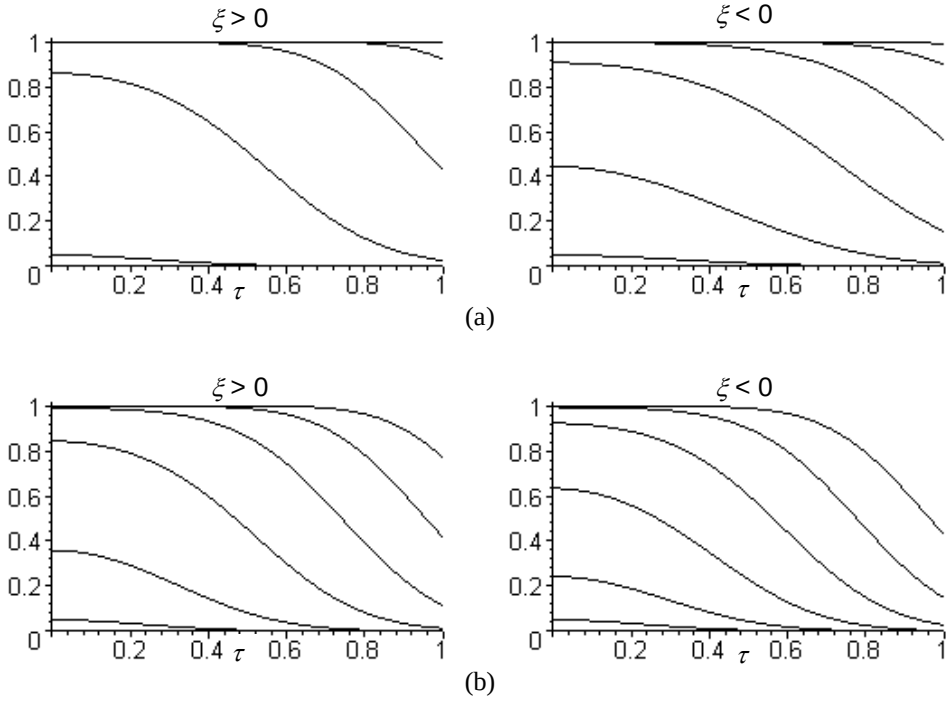


Figure 1: Courbes π_G en fonction de τ avec u = 0.5, v = 1.5, k = 3, r = 1, N = 50, α = 0.05, $C_p^l(u,v)$ = c(0.20)(c + 1) (de bas en haut) pour (a) c =1 et $C_p^l(0,0) = C_p^l(u,v) + 0.33$; (b) c =1.5 et $C_p^l(0,0) = C_p^l(u,v)$ □□□

5. Valeur critique ajustée

Nous avons vu au paragraphe précédent que le niveau et la puissance du test diminuent lorsque les erreurs de mesure croissent. La présence d'erreurs de mesure a donc pour conséquence une plus grande difficulté à considérer comme capable un processus qui est effectivement capable, et une plus grande difficulté à détecter un processus non capable. Pour améliorer le test nous proposons de déterminer une valeur critique ajustée c_0^A plus petite que la valeur critique c_0. Soient $\alpha_A = P\left(\hat{C}_p^{lG}(u,v) > c_0^A \mid C_p^l(u,v) = c\right)$ et $\pi_A(C_p^l(u,v)) = P\left(\hat{C}_p^{lG}(u,v) > c_0^A \mid C_p^l(u,v)\right)$ le niveau et la puissance du test en utilisant la valeur critique ajustée. Puisque $c_0^A < c_0$, $P\left(\hat{C}_p^{uG}(u,v) > c_0^A\right)$ est plus grand que $P\left(\hat{C}_p^{uG}(u,v) > c_0\right)$, et donc π_A et α_A augmentent. Pour conserver le niveau initial du test

nous posons $\alpha_A = \alpha$, et résolvons l'équation $\alpha = P\left(\hat{C}_p^{lG}(u,v) > c_0^A \mid C_p^l(u,v) = c\right)$ pour obtenir c_0^A. D'après (3), nous devons résoudre

$$\alpha = \int_0^{KG(c_0^A)} H_G(c_0^A, t)dt\,, \tag{4}$$

équation faisant intervenir ξ_G et $C_p^{lG}(0,0)$ qui sont inconnus puisque μ et σ_G sont inconnus. $C_p^{lG}(0,0)$ est lié à $C_p^l(0,0)$ puisque $C_p^{lG}(0,0) = C_p^l(0,0)/\sqrt{1+\tau^2}$ d'après (2). D'autre part Grau (2011b) montre que

$$C_p^l(0,0) = \sqrt{1+\tau^2}\left(\frac{\sqrt{1+v\xi_G^{l\,2}}\sqrt{1+v\xi^{l\,2}}}{\sqrt{1+\tau^2+v\xi^{l\,2}}}C_p^l(u,v) + u\xi_G^l/3\right), \tag{5}$$

où $\xi_G^l = \max\left(\xi_G/k, -\xi_G\right)$ et $\xi^l = \max\left(\xi/k, -\xi\right) = \max\left(\xi_G\sqrt{1+\tau^2}/k, -\xi_G\sqrt{1+\tau^2}\right)$. En conséquence pour $C_p^l(u,v) = c$, seul ξ_G est inconnu dans (4) et doit être estimé. Avec cette estimation et $C_p^l(u,v) = c$, la valeur ainsi obtenue pour $C_p^l(0,0)$ dans (5) est notée C_{p1}^l. Elle permet d'obtenir la valeur critique ajustée c_0^A et ensuite de calculer la puissance du test à partir de l'équation

$$\pi_A(C_p^l(u,v)) = P\left(\hat{C}_p^{lG}(u,v) > c_0^A \mid C_p^l(u,v),\, C_p^l(0,0) = C_{p1}^l\right) = \int_0^{K1(c_0^A)} H_1(c_0^A, t)dt\,,$$

où $K_1(x) = 3\sqrt{N}C_{p1}^l \Big/ \left[\left(u + 3x\sqrt{v}\right)\sqrt{1+\tau^2}\right]$ d'après (2),

$$H_1(x,t) = F_K\left(\left(\frac{3\sqrt{N}C_{p1}^l/\sqrt{1+\tau^2} - ut}{3x}\right)^2 - vt^2\right) f_{Z_{l1}^G}(t)\,,$$

$$f_{Z_{l1}^G}(t) = k\phi\left(kt - \sqrt{N}\xi_{G1}\right) + \phi\left(t + \sqrt{N}\xi_{G1}\right),$$

$$\xi_{G1} = \begin{cases} k\xi_{G1}^l = k\xi_1^l/\sqrt{1+\tau^2} & si \quad \hat{\xi}_G > 0 \\ -\xi_{G1}^l = -\xi_1^l/\sqrt{1+\tau^2} & si \quad \hat{\xi}_G < 0 \end{cases},$$

et $\xi_1^l = \dfrac{-C_{p1}^l u/3 + \sqrt{\left(C_{p1}^l\right)^2(u/3)^2 + \left(\left(C_{p1}^l\right)^2 - \left(C_p^l(u,v)\right)^2\right)\left(v\left(C_p^l(u,v)\right)^2 - (u/3)^2\right)}}{v\left(C_p^l(u,v)\right)^2 - (u/3)^2}$, voir Grau (2011b).

A l'aide du logiciel Maple, il est possible de calculer la valeur critique ajustée et la puissance du test. La figure 2 représente les variations de π_A en fonction de $\tau \in [0,1]$, avec r = 1, N = 50, α = 0.05, $C_p^l(u,v)$ = c(0.20)(c + 1), c = 1.00 et $C_p^l(0,0)$ = $C_p^l(u,v)$ + 0.33, puis c = 1.50 et $C_p^l(0,0)$ = $C_p^l(u,v)$. Dans la Figure 2 nous constatons que π_A décroît lorsque τ croît. Cependant si nous comparons les Figures 1 et 2, nous constatons que la puissance obtenue à partir de la valeur critique ajustée décroît moins rapidement. Ainsi la puissance du test a été améliorée.

6. Exemple

Pour illustrer les résultats précédents nous utilisons les données issues d'un contrôle des poids de barres de nougat dans l'entreprise Chabert & Guillot située à Montélimar. A la sortie de la production

le nougat se présente sous forme de bloc de 520 mm de longueur, 260 mm de largeur et 60 mm d'épaisseur. Dans ce bloc sont découpées une quarantaine de tranches d'une épaisseur de 13mm pour obtenir des barres de 13x260x60 mm, qui sont vendues pour un poids de 200g. La réglementation sur les produits préemballés impose un contrôle des poids de telle sorte que la moyenne des poids soit au moins égale à 200g, que le poids de toutes les barres soit supérieur à 182g, et qu'il n'y ait pas plus de 2% des barres qui aient un poids compris entre 182 et 191g. Cette dernière condition étant difficile à respecter, une trieuse pondérale éjecte les barres pesant moins de 191g, qui sont ensuite recyclées. Un opérateur situé au niveau de la trieuse pondérale peut ajuster la largeur de la coupe. Compte tenu des contraintes législatives, physiques et financières, la cible *T* est fixée à 212g, et la tolérance inférieure est égale à 191g. Par conséquent $D_l = 21$. Aucune tolérance supérieure n'est fixée, mais étant donnés les coûts annuels de surdosage, le risque d'une déviation à droite ne peut être considéré comme nul. Il a été décidé qu'une déviation à droite est 3 fois moins grave qu'à gauche. En conséquence le risque *k* est égal à 3. Pour déterminer la capabilité du processus, 20 blocs ont été sélectionnés de façon aléatoire. Dû à la déformation systématique au début et à la fin de chaque bloc, seulement 36 barres ont été retenues pour déterminer le poids moyen et l'écart-type du poids d'une barre. La représentation graphique des données, et les tests usuels d'ajustement permettent de penser que les poids suivent une distribution normale. Pour choisir un indice $C_p^l(u,v)$, Grau (2011a) propose de sélectionner le couple (u,v) en fonction d'une déviation du processus et d'une proportion de non-conformes au-delà desquelles le processus ne puisse plus être considéré comme capable. Pour l'entreprise, la moyenne du processus ne doit pas être inférieure à 200g, c'est-à-dire que la moyenne ne doit pas s'éloigner de plus de 57% de la distance entre la cible et la tolérance inférieure. D'autre part, afin d'avoir une mesure facilement interprétable, le processus est considéré comme capable dès que $C_p^l(u,v)$ est supérieur ou égal à 1. A partir des résultats obtenus par Grau (2011a), pour des valeurs de *u* et *v* définies avec une précision de 0.1, les couples pour lesquels la moyenne ne s'éloigne pas de plus de 57% de la distance entre la cible et la tolérance inférieure sont les couples $(u,v) = (0.1,0.3)$, $(0.4,0.2)$ et $(0.8,0.1)$. Pour ces trois couples la proportion maximale de barres non conformes lorsque $C_p^l(u,v) = 1$ est respectivement 6037, 3645 et 1681 ppm (parts par million), ou encore 0.6037%, 0.3645% et 0.1681%. Compte tenu des coûts assez faibles de recyclage (poids inférieur à 191g), l'entreprise a décidé de choisir l'indice $C_p^l(0.1,0.3)$. Ceci signifie que lorsque l'indice $C_p^l(u,v)$ est supérieur ou égal à 1, on est sûr que la moyenne du processus n'est pas inférieure à 200g et que la proportion de barres non conformes est inférieure à 0.6037%. Pour déterminer si le processus est capable avec un degré d'erreur de contamination τ égal à 0.18 obtenu par une étude de répétabilité et de reproductibilité, nous posons $c = 1$ et $\alpha = 0.05$. A partir de l'échantillon de $N = 720$ observations, nous obtenons $\bar{\bar{G}} = 209.994$, $S_G = 4.418$, $\hat{\xi}_G = -0.454$ et $\hat{c}_p^{lG}(0.1,0.3) = 1.523$. A l'aide du logiciel Maple, nous obtenons la valeur critique $c_0^A = 1.043$. Puisque $\hat{c}_p^{lG}(0.1,0.3) > c_0^A$, nous pouvons conclure que le processus est capable. A titre d'illustration de l'intérêt de prendre en compte les erreurs de mesure, considérons un échantillon tel que $\bar{\bar{G}} = 209.100$ et $S_G = 6.388$. Nous avons toujours $\hat{\xi}_G = -0.454$, et donc $c_0^A = 1.043$. Par contre, si nous ignorons les erreurs de mesure, la valeur critique prend la valeur $c_0 = 1.058$. Puisque $c_0^A = 1.043 < \hat{c}_p^{lG}(0.1,0.3) = 1.049 < c_0 = 1.058$, en tenant compte des erreurs de mesure nous concluons que le processus est capable, mais nous n'obtenons pas la même conclusion si nous les ignorons.

7. Conclusion

Pour des processus ne possédant qu'une limite de tolérance, Grau (2009) propose d'utiliser les indices $C_p^u(u,v)$ et $C_p^l(u,v)$. Le choix du couple (u,v) peut être effectué en fonction de la déviation maximale admissible par rapport à la cible et/ou la proportion maximale admissible de non conformes, Grau (2011a). Dans cette étude nous montrons que la présence d'erreurs de mesure peut avoir un impact important sur les conclusions du test de la capabilité du processus. Nous proposons donc d'utiliser une valeur critique ajustée permettant d'améliorer la puissance du test. Enfin un exemple réel tiré de l'industrie agro-alimentaire permet d'illustrer la méthode proposée.

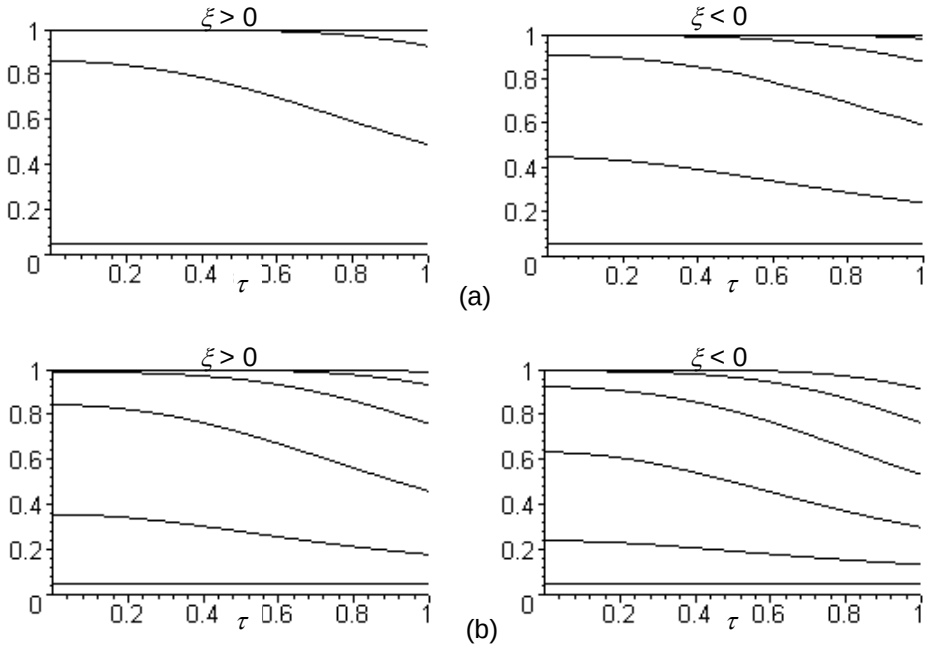


Figure 2: Courbes π_A en fonction de τ avec $u = 0.5$, $v = 1.5$, $k = 3$, $r = 1$, $N = 50$, $\alpha = 0.05$, $C_p^l(u,v) = c(0.20)(c + 1)$ (de bas en haut) pour (a) $c = 1$ et $C_p^l(0,0) = C_p^l(u,v) + 0.33$; (b) $c = 1.5$ et $C_p^l(0,0) = C_p^l(u,v)$ □□□

Bibliographie

Kane, V.E. (1986). Process capability indices, *Journal of Quality Technology*, 18(1), 41-52.

Chan, L. K., Cheng, S. W. et Spiring, F. A. (1988). A New Measure of Process Capability : C_{pm}. *Journal of Quality Technology*, 20(3), 162-175.

Pearn, W. L., Kotz, S. and Johnson, N. L. (1992). Distributional and inferential properties of process capability indices. *Journal of Quality Technology*, 24(4), 216-231.

Vännman, K. (1995). A unified approach to capability indices. *Statistica Sinica*, 5(2), 805-820.

Chen, K. S. and Pearn, W. L. (2001). Capability indices for processes with asymmetric tolerances. *Journal of the Chinese Institute of Engineers*, 24(5), 559-568.

Grau, D. (2009). New Process capability indices for one-sided tolerances. *Quality Technology and Quantitative Management*, 6(2), 107-124.

Grau, D. (2011a). Process yield, process centering, and capability indices for one-sided tolerance processes. *Quality Technology and Quantitative Management*, in press.

Mittag, H. J. (1997). Measurement error effects on the performance of process capability indices. *Frontiers Statistical Quality Control*, 5, 195-206.

Pearn W. L., Lin P. C., Chen K. S. (2001) Estimating process capability index C''_{pmk} for asymmetric tolerances: Distributional properties, *Metrika* 54(3), 261-279.

Grau, D. (2011b). Testing capability indices for one-sided processes with measurement errors, soumis à publication.

Fiabilité des résultats de méthodes analytiques : une approche Bayesienne de la validation des méthodes analytiques

Reliability of analytical methods' results : a Bayesian approach to analytical method validation

Eric Rozet[1], Bernadette Govaerts[2], Pierre Lebrun[1], Bruno Boulanger[3], Eric Ziemons[1] & Philippe Hubert[1]

[1] *Laboratory of Analytical Chemistry, Department of Pharmacy, Université de Liège, 1 Av. de l'Hôpital, Bat. B36, 4000 Liège, Belgique.*
E-mail : eric.rozet@ulg.ac.be

[2] *Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgique.*

[3] *Arlenda SA, Liège, Belgique.*

Abstract
Of core importance at the end of the validation is the evaluation of the reliability of the individual results that will be generated during the routine application of the method. Regulatory guidelines provide a general framework to assess the validity of a method, but none address the issue of results reliability. In this study, a Bayesian approach is proposed to address this concern. By providing the minimum reliability probability $(\pi_{\min})$ needed for the subsequent routine application of the method, as well as specifications or acceptance limits $(\pm\lambda)$, the proposed Bayesian approach provides the effective probability of obtaining reliable future analytical results over the whole concentration range investigated. This is summarized in a single graph: the reliability profile. This Bayesian reliability profile is also compared to two frequentist approaches. Furthermore, the applicability of the Bayesian reliability profile is shown using as example the validation of a bioanalytical SPE-HPLC-UV method.
Keywords : Results reliability; validation; reliability profile; Bayesian approach

1. Introduction

Evaluation of the reliability of analytical results obtained by quantitative analytical methods is essential in order to appraise the trustworthiness of decisions made using them, especially where health and financial risks are involved. It is assumed that this evaluation will be made during method validation studies. Owing to the key role of analytical results in such decision making, regulatory expectations are expressed in numerous documents that are issued corresponding to the given industrial sector [1-8]. Nonetheless, these documents focus only on analytical method performance criteria such as the evaluation of systematic error (or trueness) and random error (or precision), linearity, limit of quantification, etc. These documents do not focus on the main aim of any quantitative analytical method, i.e. the method's results. The decision to declare a method fit for its purpose should fundamentally be made on the reliability of the results it generates.

Some improvements in defining and assessing method validation have been made in order to return to the essence of quantitative analytical methods by focusing on total error [9-17]. In fact, total error is the measure of the global amount of analytical error that is linked to each result generated by a method. However, this is not enough and a further step needs to be performed by evaluating the probability of obtaining reliable results over the concentration range investigated. Previous research related to the evaluation of this probability has been published [18-20]. However, these reports focus only on a single concentration evaluation and no information is provided regarding the reliability probability over the whole concentration range investigated, thus leading to a gross reliability evaluation.

Reliability of analytical results obtained by quantitative analytical methods should be evaluated favourably over the intended concentration range. Therefore, the reliability of analytical results obtained by a quantitative analytical method could be defined as:

"the probability (π) *of an analytical method to provide analytical results* (X) *within predefined acceptance limits* $(\pm\lambda)$ *around their reference or conventional true concentration values* (μ_T) *over a defined concentration range and under given environmental and operating conditions*."

This can be expressed by:

$$\pi = P(-\lambda < X - \mu_T < \lambda) \qquad \text{Eq. 1}$$

Consequently, the objective of the validation phase can be summarised to evaluate whether the reliability probability π that each future result will fall within predefined acceptance limits (λ) is greater than or equal to a minimum claimed level $\pi_{\min}$ [18-20]. The statistical problem here is two-fold: the probability π needs to be estimated and the uncertainty in its estimation must be taken into account when comparing it to $\pi_{\min}$. This is not an easy problem to solve since it has no exact small sample solution in frequentist statistics [18-20].

This study thus aims at providing a thorough methodology for the evaluation of reliability of analytical results obtained by a quantitative analytical method. The main objective is to provide a reliability profile for the full concentration range of the analyte of interest. This profile will give the probability of obtaining results within defined specifications or acceptance limits over the whole concentration range studied. From this perspective, the Bayesian approach toward inference offers many advantages. The Bayesian framework motivating the evaluation of the reliability of analytical results obtained by a quantitative analytical method will first be detailed. Then simulations will be performed to compare its performances with two available frequentist approximations of π [18,20]. Finally, its applicability will then be illustrated with the analysis of results reliability of a real case. The application studied is a new bioanalytical method dedicated to the determination of ketoglutaric acid (KG) and hydroxymethylfurfural (HMF) in human plasma by double-cartridge SPE-HPLC-UV.

2. Method validation: principles, model and actual reliability approaches

2.1 Design and statistical model of a validation study

Analytical method validation are designed in such a way that sources of variability that will be encountered during the future routine use of the method are included such as operator, equipments or days. The combination of these sources of variability is called runs or series. Usually at least three runs are performed. Each run is composed of calibration standards used to fit a response function and validation standards prepared in the matrix where the analyte will be found in routine experiments. Calibration standards at several concentration levels are analyzed at least in duplicates in each run. Additionally three to six independent replicates of validation standards also at several i^{th} concentration levels $\mu_{T,i}$ (*i*: 1 to I) are analysed in the same run. These validation standards are used to assess the adequacy of the response function and the validity of the whole analytical method under study. Let suppose that for each of the i^{th} concentration level of the validation standards, the number of runs is *J* and that in each run *K* replicates are performed. The validation experiments can be described, for each of the i^{th} concentration level studied, by a one way random Analysis Of Variance (ANOVA) model with runs (or series) as random factor:

$$X_{i,jk} = \mu_i + \alpha_{i,j} + \varepsilon_{i,jk}, \quad \alpha_{i,j} \sim N\left(0, \sigma^2_{\alpha,i}\right), \quad \varepsilon_{i,jk} \sim N\left(0, \sigma^2_{\varepsilon,i}\right) \qquad \text{Eq. 2}$$

where μ_i is the overall mean of the i^{th} concentration level studied of the validation standard, $\mu_i + \alpha_{i,j}$ is the mean in run *j* (*j*: 1 to J), $\varepsilon_{i,jk}$ is the residual error, $\sigma^2_{\alpha,i}$ is the run-to-run variance, and $\sigma^2_{\varepsilon,i}$ is the within-run or repeatability variance, both for the i^{th} concentration level.

The overall variability of the analytical method is measured by the intermediate precision variance $\sigma^2_{I.P.,i} = \sigma^2_{\alpha,i} + \sigma^2_{\varepsilon,i}$. All these parameters of the variance components model can be estimated by classical ANOVA methods [21].

2.2 Actual reliability probability estimators

Based on this one way random ANOVA model, two propositions have been made to compute the reliability probability π. They are both frequentist approximations and can only be applied separately to each of the *I* concentration levels providing discrete estimation of the reliability probability only at the $\mu_{T,i}$ concentration levels tested with the validation standards.

The first frequentist approximation, adapted from Dewé et al. [18], is to compute based on the formula of statistical tolerance intervals the probability of obtaining results included within the specified acceptance limits [-λ, + λ]. Statistical tolerance intervals are intervals within which it is expected that each future result has a user-defined probability of falling [22,23]. If using β-expectation tolerance interval [24] this is computed for each concentration level *i* as:

$$\pi_i^{Beti} = P\left[X_i > \mu_{T,i} - \lambda\right] + P\left[X_i < \mu_{T,i} + \lambda\right]$$

$$= P\left[t(f) > \frac{\left(\mu_{T,i} - \lambda\right) - \overline{X}_i}{\hat{\sigma}_{I.P.,i}\sqrt{1 + \frac{K\hat{R}_i + 1}{N\left(\hat{R}_i + 1\right)}}}\right] + P\left[t(f) < \frac{\left(\mu_{T,i} + \lambda\right) - \overline{X}_i}{\hat{\sigma}_{I.P.,i}\sqrt{1 + \frac{K\hat{R}_i + 1}{N\left(\hat{R}_i + 1\right)}}}\right] \quad \text{Eq. 3.}$$

where J is the number of runs and K the number of replicates by series, $N=JK$. $\overline{X}_i$ is the mean concentration of the results obtained by the method for the i^{th} concentration level and $\hat{\sigma}_{I.P.,i}$ is the intermediate precision standard deviation for each i^{th} concentration level. $t(f)$ is a student distribution with f degrees of freedom computed based on the Satterthwaite approximation [25] and $\hat{R}_i$ is the ratio between the run-to-run variance and the within-run (or repeatability) variance of each concentration level.

The second approach estimates by maximum likelihood the reliability probability [20]:

$$\pi_i^{ML} = P\left[Z > \frac{\left(\mu_{T,i} - \lambda\right) - \overline{X}_i}{\hat{\sigma}_{I.P.,i}}\right] + P\left[Z < \frac{\left(\mu_{T,i} + \lambda\right) - \overline{X}_i}{\hat{\sigma}_{I.P.,i}}\right] \quad \text{Eq. 4}$$

where Z is a standard normal variable.

3. Bayesian framework

The novel proposition made in this paper relies on the commonly applied designs of analytical method validation studies described in the previous section.[1,3,9]. However, rather than evaluating the method performance independently for each concentration, this approach models this reliability probability over the whole concentration range over which the analytical method is studied. In this context model Eq. 2 is rewritten as the following linear model with random slopes and intercepts:

$$X_{ijk} = \beta_0 + \beta_1\mu_{T,i} + u_{0,j} + u_{1,j}\mu_{T,i} + \varepsilon_{ijk} \quad \text{Eq. 5.}$$

where the subscripts i stands for the I concentration levels of the validation standards, j for the J number of series or runs and k for the K number of replicates per run. $\mu_{T,i}$ is i^{th} concentration level of the validation standard and is considered as a reference or conventional true value. $\boldsymbol{\theta} = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$ are the fixed effects. In analytical chemistry they represent the constant bias and proportional bias, respectively. This form of model for the bias function is acceptable for analytical methods. Indeed the "linearity" validation criterion expresses that, on average, results obtained by analytical methods should ideally be strictly proportional to the true concentration [3,14]. Nonetheless, in practice, strict proportionality is never observed and the linear regression proposed is a widely observed situation (see

for example [14]). Additionally, $\mathbf{U}_j = \begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix}$ are the random effects of the j^{th} runs and are also assumed coming from a normal distribution:

$$\mathbf{U}_j \sim iN(\mathbf{0}, \sigma_u^2 \underset{2x2}{\mathbf{\Sigma}}) \qquad \text{Eq. 6.}$$

Finally, ε_{ijk} is the residual error assumed independent and coming form a normal distribution of variance σ_i^2. This variance is also given as being dependent on the concentration level *i*. Practically speaking, this phenomenon is frequently observed in real life situations [26]. The general form of this variance function is a power of the concentration:

$$\sigma_i = \sigma(\mu_{T,i})^{\gamma} \qquad \text{Eq. 7.}$$

The logarithmic transformation of the variance function is, however, preferred in order to ensure that only positive values of variances will be obtained.

$$Log_{10}(\sigma_i) = Log_{10}(\sigma) + \gamma Log_{10}(\mu_{T,i}) \qquad \text{Eq. 8.}$$

Finally the following vague priors are defined:

$$\boldsymbol{\theta} \sim N\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \mathbf{\Gamma}\right) \qquad \text{Eq. 9.}$$

$$\mathbf{\Gamma}^{-1} = \mathbf{0}$$

$$\mathbf{\Sigma} \sim \text{Wishart}(0.0001\mathbf{I}_2, 2)$$

where $\mathbf{I}_2$ represents the 2 x 2 identity matrix and $\mathbf{\Gamma}^{-1} = \mathbf{0}$ denotes a matrix of 0s that represents a vague prior of $\boldsymbol{\theta}$.

$$\gamma \sim N(0, 0.0001)$$

$$\tau = \frac{1}{\sigma} \sim Gamma(0.0001, 0.0001)$$

Having specified the regulatory or client acceptance limits (λ), the main aim is to obtain the reliability probability (π) as a function of the reference concentration $\mu_{T,i}$. When the Bayesian model specification is completed, MCMC sampling can be performed, using freely available software such as Winbugs (for example, using R2Winbugs package from R), which allows us to obtain the posterior distribution of each parameter. From these posterior distributions one can then easily obtain the predictive distribution of the reliability probability π for any concentration level $\mu_{T,i}$. Indeed, from the posterior distribution of the parameters obtained with the MCMC sampling, the posterior reliability probability π of results following the normal distribution defined in Eq. 5 lying inside the acceptance limits $[-\lambda\,;\,\lambda]$ is easily obtained analytically.

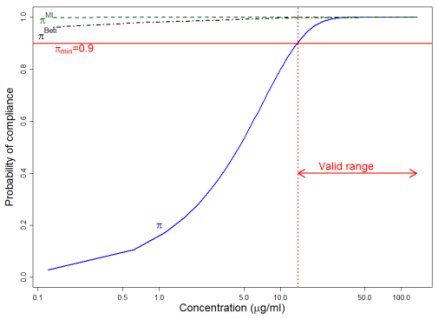


Figure 1: Reliability profile for ketoglutaric acid (KG) depicting on the y-axis the reliability probability or probability of compliance i.e. the probability to obtain future analytical results within pre-specified acceptance limits λ=±20% with respect to the reference concentration of analytes on the x-axis. Continuous (blue) curve: Bayesian reliability profile π; Dashed (green) line: frequentist reliability profile π^{ML} (Eq. 4). Dotted-Dashed (black) line: frequentist reliability profile π^{Beti} (Eq. 3). Continuous horizontal (red) line: minimum reliability probability $\pi_{\min}$ set at 90%.

Finally, a graph, similar to Figure 1, representing the reliability of the results obtained by a quantitative analytical method over the whole concentration range studied can be obtained, and the concentration range over which the method is sufficiently reliable can be determined by comparing the posterior reliability probability (π) to a minimum reliability value ($\pi_{\min}$), for e.g. 95%.

4. Simulations

In order to evaluate the performance of the proposed Bayesian framework to compute the reliability probability π over the whole concentration range studied, it is compared to the two frequentist approaches π^{Beti} (Eq. 3) and π^{ML} (Eq. 4) that allows to compute the probability only on specifically defined concentration levels.
Simulation scenarios

4.1. Simulation scenarios

All the simulations are performed assuming the linear mixed model represented in Eq. 5 in four cases over a concentration range arbitrarily defined from 60 to 120%:

- **Case 1:** Absence of systematic error and large intermediate precision variance (which corresponds for the 100% concentration level to a relative bias=0% and Intermediate precision RSD=15.9%):

$$X_{ijk} = 0 + 1\mu_{T,i} + u_{0,j} + u_{1,j}\mu_{T,i} + \varepsilon_{ijk} \text{ with } \mathbf{U}_j \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.03^2 & 0\\0 & 0.012^2\end{pmatrix}\right] \text{ and } \sigma_i = 1(\mu_{T,i})^{0.6}.$$

- **Case 2:** A proportional systematic error and large intermediate precision variance (which corresponds for the 100% concentration level to a relative bias=10% and Intermediate precision RSD=15.9%):

$$X_{ijk} = 0 + 1.1\mu_{T,i} + u_{0,j} + u_{1,j}\mu_{T,i} + \varepsilon_{ijk} \text{ with } \mathbf{U}_j \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.03^2 & 0\\0 & 0.012^2\end{pmatrix}\right] \text{ and } \sigma_i = 1(\mu_{T,i})0.6$$

- **Case 3**: Absence of systematic error and small intermediate precision variance (which corresponds for the 100% concentration level to a relative bias=0% and Intermediate precision RSD=6.5%):

$$X_{ijk} = 0 + 1\mu_{T,i} + u_{0,j} + u_{1,j}\mu_{T,i} + \varepsilon_{ijk} \text{ with } \mathbf{U}_j \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.03^2 & 0\\0 & 0.012^2\end{pmatrix}\right] \text{ and } \sigma_i = 1(\mu_{T,i})0.4$$

- **Case 4:** A proportional systematic error and small intermediate precision variance (which corresponds for the 100% concentration level to a relative bias=10% and Intermediate precision RSD=6.5%):

$$X_{ijk} = 0 + 1.1\mu_{T,i} + u_{0,j} + u_{1,j}\mu_{T,i} + \varepsilon_{ijk} \quad \text{with} \quad \mathbf{U}_j \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.03^2 & 0\\0 & 0.012^2\end{pmatrix}\right] \quad \text{and}$$

$\sigma_i = 1(\mu_{T,i})0.4$.

For the two frequentist approaches four concentration levels have been selected: 60, 80, 100 and 120%. The acceptance limits have been settled at $\lambda = \pm 20\%$ around the defined reference concentration levels. Additionally, the true probability of having results falling within these acceptance limits have been computed. The number of series and repetitions have been defined at *J*=4 and *K*=4, respectively. 2000 simulations were performed for each case.

4.2. Simulations results

Figures 2a to 2d show the results of the simulations for the proposed Bayesian reliability profile. These figures show the median reliability probability for the Bayesian approach (π-blue continuous line) as well as for the two frequentist approaches: π^{ML} (dashed green line) and π^{Beti} (doted-dashed black line). The true reliability probability is depicted on these figures with the continuous red line. On Figure 2 the two frequentist approaches are quite close to the true reliability probability, with better performances for the π^{ML} approach.

Figure 2 also shows that the Bayesian reliability probability π is also very close to the true reliability probability. One important additional feature that is added onto Figures 2a to 2d is the three shaded regions. These shaded regions represent the interval between the minimum and maximum probability values obtained by each approach during the simulations. As can be seen, the two frequentist approaches have always larger width than the Bayesian approach. These results suggest that, although the two frequentist approaches will on average provide adequate estimation of the true reliability probability (especially for π^{ML}), the Bayesian reliability profile will always provide more precise estimation of this reliability profile, closer to the true reliability probability. This means that when assessing the reliability of a quantitative analytical method, the estimation of the reliability probability given by our proposed Bayesian profile has greater probability to be closer to the real reliability

probability than the two frequentist approximations. The decision about the reliability of an analytical method using the proposed Bayesian approach will hence be more accurate.

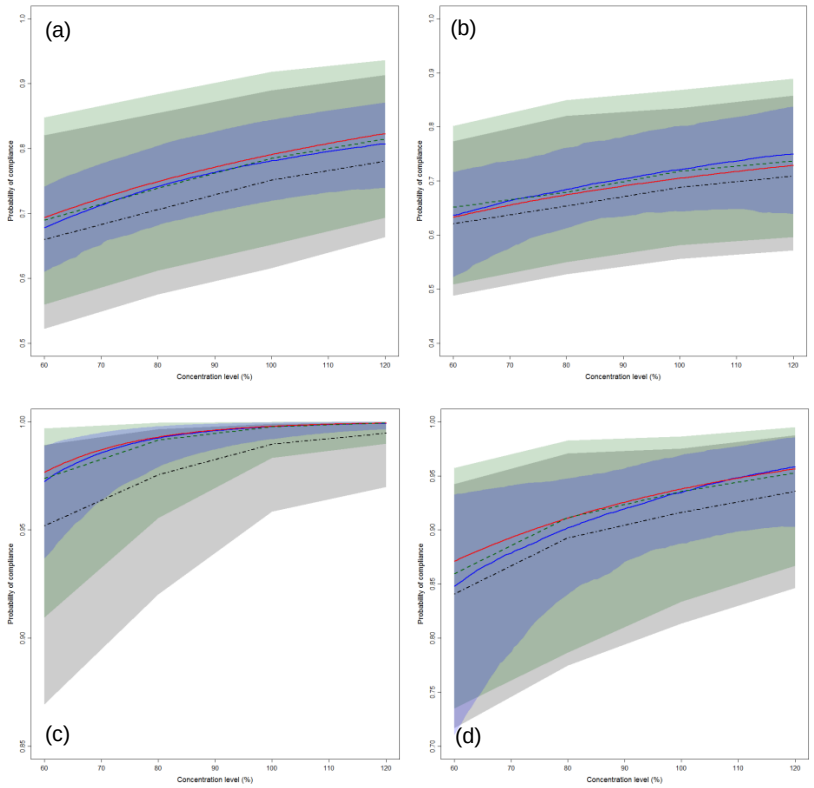


Figure 2: Reliability profiles obtained for the simulations for (a) case 1, (b) case 2, (c) case 3, (d) case 4,. The y-axis depicts the reliability probability or probability of compliance i.e. the probability to obtain future analytical results within pre-specified acceptance limits settled at λ=±20% with respect to the reference concentration of analytes on the x-axis. Continuous (blue) curve: Bayesian reliability profile π; Dashed (green) line: frequentist reliability profile π^{ML} (Eq. 4). Dotted-Dashed (black) line: frequentist reliability profile π^{Beti} (Eq. 3). Continuous (red) line: true reliability probability. The shaded regions represent the interval between the minimum and maximum probability values obtained by each approach (π, π^{Beti}, π^{ML}) during the simulations..

5. Real case study

As stated in previous sections, it is argued here that using all the validation criteria, including results accuracy, is not the ultimate aim in assessing the validity of analytical method. Linearity, precision, trueness and accuracy are needed, but the final target is to evaluate the reliability of the results

generated by the quantitative analytical method. The whole Bayesian procedure described in Section 3 is applied, by defining the acceptance limits within which the results must fall. These acceptance limits (λ) was set at a maximum of +/-20% around the reference concentration values ($\mu_{T,i}$) of the validation standards. The minimum reliability criterion ($\pi_{\min}$) was set at 90%, meaning that the minimum probability of obtaining future measurements falling within the specification limits during routine analysis is 0.90. Having set these requirements, the reliability profile can be worked out, as illustrated in Figure 1 for KG.

In this profile, the minimum reliability criterion ($\pi_{\min}$) of 90% is shown by the continous (red) horizontal lines. The concentration levels with at least 90% reliability define the valid concentration range for KG and consequently represent the lower and upper quantification limits. The valid concentration range using the Bayesian algorithm extends from 14.1 to 133.3 µg/ml for KG. In Figure 1, the two frequentist reliability probabilities (π^{Beti} and π^{ML}) are also drawn for comparison purposes.

As can be seen on Figure 1, the two frequentist approaches would define the analytical method as reliable over the whole concentration range tested, while the Bayesian one reduces the valid concentration range. On the light of the results obtained from the previous simulations, the Bayesian reliability profile provides a more precise evaluation of the reliability probability and the frequentist approaches could result in overconfident decisions about the validity and reliability of the SPE-HPLC-UV method for the quantification of KG in human plasma. The Bayesian reliability profile, by including the uncertainty of the parameters over the whole concentration range studied ultimately makes the decision of validity more trustworthy and should lead to an increase in reliability for final users of the results such as clients, patients and so on.

6. Conclusion

In this study, a novel Bayesian proposition was made in order to evaluate the reliability of results obtained by analytical methods over a defined concentration range. The usual validation criteria, i.e. trueness (systematic error), precision (random error) and accuracy (total error), are only intermediate steps in assessing this reliability. They are then combined and summarized in a single value: the probability of obtaining reliable results over the whole concentration range investigated (π). This leads to a reliability profile over which the lower and upper quantification limits are appropriately located and thus the analytical method valid concentration range can be obtained by comparison with a minimum reliability probability ($\pi_{\min}$).

The simulations performed to compare this novel approach with two frequentist ones showed that the Bayesian approach provide accurate and more precise estimation of the reliability probability, a point that is essential considering the highly important health and financial decisions that will be made using the validated methods in routine applications. Furthermore, the Bayesian approach is the first attempt at allowing the evaluation of the reliability of these results over the entire concentration range investigated. The applicability of this approach to real validation studies was tested here with the validation of a complex bioanalytical method. Finally, the Bayesian approach is in full agreement with the regulatory validation guidelines since all the required validation criteria are included. The major difference is that the validity of the analytical method is decided by proving that "the probability (π) of an analytical method to provide analytical results (X) within predefined acceptance limits $(\pm\lambda)$

around their reference or conventional true concentration values (μ_T) over a defined concentration range and under given environmental and operating conditions" is acceptable.
Finally, it is shown in this work that the determination of quantitative analytical results reliability is the final aim of any quantitative analytical method validation. Indeed, using these reliability profiles the analyst can see how far the results generated by the analytical method will be reliable for its future intended use. The consumer or client risk linked to the use of the results generated by the analytical procedure is known and managed. This probability is never known by the other actual analytical method validation decision methodologies.

Acknowledgment

A research grant from the Belgium National Fund for Scientific Research (FRS-FNRS) to E. Rozet is gratefully acknowledged.

Bibliographie

[1] Guidance for industry: Bioanalytical Method Validation, US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Rockville, May 2001.
[2] M. Thompson, S.L.R. Ellison, Wood R., Pure Appl. Chem. 74 (2002) 835-855.
[3] International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use, Topic Q2 (R1): Validation of Analytical Procedures: Text and Methodology, Geneva, 2005.
[4] EP21-A, Estimation of Total Analytical Error for Clinical Laboratory Methods; Approved Guideline. Clinical and Laboratory Standards Institute (CLSI 21-A, Wayne, PA, 2003).
[5] ISO/CEI 17025: General requirements for the competence of testing and calibration laboratories, International Organization for Standardization (ISO), Geneva, 2005.
[6] ISO 15189: Medical laboratories - Particular requirements for quality and competence, International Organization for Standardization (ISO), Geneva, 2007.
[7] Method Validation and Quality Control procedures for Pesticide Residues, in Doc. SANCO/2007/3131, 31 October 2007.
[8] 2002/657/EC: Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC Concerning the Performance of Analytical Methods and the Interpretation of Results. Off. J. Eur. Commun. L221 (2002) 8.
[9] Ph. Hubert, J.J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, J. Pharm. Biomed. Anal. 36 (2004) 579-586.
[10] Ph. Hubert, J.J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, E. Rozet, J. Pharm. Biomed. Anal. 45 (2007) 70-81.

[11] Ph. Hubert, J.J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, E. Rozet, J. Pharm. Biomed. Anal. 45 (2007) 82-96.
[12] E. Rozet, C. Hubert, A. Ceccato, W. Dewé, E. Ziemons, F. Moonen, K. Michail, R. Wintersteiger, B. Streel, B. Boulanger, Ph. Hubert, J. Chromatogr. A 1158 (2007) 126-137.
[13] B. Boulanger, E. Rozet, F. Moonen, S. Rudaz, Ph. Hubert, J. Chromatogr. B 877 (2009) 2235-2243.
[14] E. Rozet, A. Ceccato, C. Hubert, E. Ziemons, R. Oprean, S. Rudaz, B. Boulanger, Ph. Hubert, J. Chromatogr. A, 1158 (2007) 111-125.
[15] A. Bouabidi, E. Rozet, M. Fillet, E. Ziemons, E. Chapuzet, B. Mertens, R. Klinkenberg, A. Ceccato, M. Talbi, B. Streel, A. Bouklouze, B. Boulanger, Ph. Hubert, J. Chromatogr. A 1217 (2010) 3180-3192.
[16] D. Hoffman, R. Kringle, Pharm Res. 24 (2007) 1157-1164.
[17] D. Hoffman, R. Kringle, J. Biopharm. Stat., 15 (2005) 283-293.
[18] W. Dewé, B. Govaerts, B. Boulanger, E. Rozet, P. Chiap, Ph. Hubert, Chemometr. Intell. Lab. Syst. 85 (2007) 262-268.
[19] B. Boulanger, W. Dewé, A. Gilbert, B. Govaerts, M. Maumy, Chemometr. Intell. Lab. Syst. 86 (2007) 198–207.
[20] B. Govaerts, W. Dewé, M. Maumy, B. Boulanger, Qual. Reliab. Engng. Int. 24 (2008) 667-680.
[21] Searle S.R., Casella. G. and McCulloch C.E., *Variance components* (1992), Wiley.
[22] Guttman I., *Statistical Tolerance Regions: Classical and Bayesian*, Hafner, Darian, 1969.
[23] Hahn G.J., Meeker W.Q., *Statistical Intervals: A Guide for Practitioners*, Wiley: New York, 1991.
[24] R.W. Mee, Technometrics, 26 (1984) 251-254.
[25] F.E. Satterthwaite, Psychometrika, 6 (1941) 309-316.
[26] W. Horwitz, L.R. Kamps, K.W. Boyer, J. Assoc. Off. Anal. Chem. 63 (1980) 1344-1354.

Session 6 : Analyse de Risque II / *Risk Analysis II*

Optimisation de la décision de surveillance : cas de la fabrication des lardons

Optimisation of surveillance decision: application to the diced bacon process

Natalie Commeau[1,2,3] & Marie Cornu[4] & Eric Parent[1]

[1] *UMR 518 INRA-MIA, 16 rue Claude Bernard 75005 Paris, France*
E-mail : *natalie.commeau@agroparistech.fr*

[2] *Laboratoire de sécurité des aliments, ANSES, 23 av du Général de Gaulle, 94706 Maisons-Alfort, France*
[3] *AgroParisTech ENGREF, 19 avenue du Maine, 75732 Paris, France*
[4] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), DEI, SECRE, LME, Cadarache, France*

Résumé
Nous proposons une application de la théorie bayésienne de la décision dans le domaine de la microbiologie des aliments. La construction d'une fonction de coût basée sur les décisions que peut prendre une entreprise après avoir examiné les résultats d'un plan d'échantillonnage à deux classes est proposée. Les coûts sont basés sur 1) les décisions que l'entreprise peut prendre et 2) les éventuelles pénalités que le client impose à l'entreprise si la qualité des aliments produits n'est pas satisfaisante. Le but de ce travail est de montrer une manière de déterminer la taille d'un échantillon de manière à minimiser l'espérance des coûts pour l'entreprise.
Mots-clés : théorie bayésienne de la décision, plan d'échantillonnage, fonction de coût, microbiologie des aliments

Abstract
We propose to use the Bayesian theory of decision in the field of food microbiology. A way to construct a loss function linked to the decisions a plant has to make after having studied the results of a two-class sampling plan is detailed. The costs of the loss function are based on 1) the decisions taken by the plant and on 2) the possible fines the client imposes to the plant if the quality of the food is not good enough. The aim of this work is to show a way to determine a sample-size for the sampling plan so that the expected costs are minimal for the plant.
Keywords : Bayesian decision theory, sampling plan, loss function, food microbiology

1 Introduction

In the food industry, food business operators are concerned with contamination in their plant, to ensure that "foodstuffs comply with the relevant microbiological criteria" (European regulation (EC) No.2073/2005, Article 3). Sampling plans are one tool used to assess the microbiological

contamination of the food at one stage of the process. In particular, two-class attribute sampling plans estimate the proportion of units in which a given micro-organism is detectable. This proportion is often referred to as the "apparent prevalence", or more simply the prevalence.
For each subpart of the production (e.g. a lot, or a day/month/year of production), n units are sampled and tested for presence of the micro-organism. A (crude) estimate of the apparent prevalence is then x/n, where x is the number of "positive" units. Moreover, such a sampling plan can also be used as a decision tool. If x is lower than or equal to c the maximum allowable number of positive units, the lot can be sent to the client or the trust in the process is reinforced, whereas if x is higher than c then another decision is taken (e.g. the lot is rejected, or the process has to be corrected). The difficulty of this decision tool is to set n and c properly. We propose a technique using the Bayesian decision theory (see Berger, 1985) to determine these numbers for a given scenario so as to minimise the costs of the plant and we apply it to a plant processing diced bacon (a typically French product made of pork breast) and the *L. monocytogenes* contamination.

2 Bayesian decision theory

Making decision when there is no uncertainty is relatively simple but it becomes more difficult when there are uncertainties. The decision maker has to choose, for example, between an option with not much uncertainty and another with lots of uncertainty. The second option could be better or worse then the first option but the decision maker does not know by which. To solve this problem, probability and the Bayesian decision theory are used.
The decision theory under uncertainty aims to take decision by taking into account all the available information. The theory supposes that the set $\mathcal{D}$ of all possible decisions d as well as the set Ψ of all possible values for states of nature ψ are defined. The latter is also called the parameter, with prior distribution π. The *loss function* defined for all $(d, \psi) \in \mathcal{D} \times \Psi$ links a loss when decision d is taken and ψ turns out to be the true state of nature. This loss function is a key element of the decision theory.
A natural method to find the best decision in the presence of uncertainty is to find the decision d which minimizes the expected loss function (Berger, 1985). In this work, decision is taken after observation $x \in \mathcal{X}$ has been observed. The *decision rule* is a function from $\mathfrak{X}$ into $\mathcal{D}$. The distribution of X conditionally to ψ is denoted $f(X = x|\psi)$.
The most natural expected loss is the one involving uncertainty, since θ is unknown when the decision is taken. The *Bayesian expected loss* of an action d is

$$\rho(d|x) = \int_{\Psi} L(\delta(x) = d, \psi)\pi(\psi)d\psi.$$

The aim of our work is to find $d^\star = arg\min_d \rho(d|x)$. In fact, $d^\star$ also minimizes the *Bayes risk* of a decision rule $\delta : r(\delta) = \int_{\mathfrak{X}} \int_{\ominus} L(\delta(x), \psi)\pi(\psi)f(x|\psi)dxd\psi$.
Here, the data has not been observed yet and the decision maker has first to choose the experimental set-up e which will give x_e. The best decision rule is the one which minimizes $r(\delta_e) = \int \int L(\delta_e(x_e), \psi)\pi(\psi)f(x_e|\psi)dxd\psi$.

3 Application to a diced bacon process

We focused on a specific plant which processes diced bacon and sells all its production to a unique client, but most of the exercise below is easily transferrable to another case. In this example, the (relatively) classical situation in which a sampling plan is used as a decision tool to accept or reject a lot is not appropriate as it does not fit to the way this plant manages its contamination by *L. monocytogenes*. Actually, this type of decision (accepting or rejecting a lot) is indeed classical in statistical studies but not so much universal in the practice of the management of microbial contamination by the food business operators, for various reasons, in particular the fact that the lot is not always an appropriate level of aggregation when dealing with an ubiquitous micro-organism (as discussed by Commeau et al., in press).
On the basis of exchanges with the stakeholders, we assume that this plant (in interaction with its client, and the regulators) is mostly concerned by its monthly prevalence, which is defined as the proportion in which *L. monocytogenes* is detectable in the entire diced bacon production over a given period of one month. Each month, this prevalence is denoted ψ. We assume that the between-month variability of ψ follows a Beta distribution with parameters a and b. A month-to-month temporal dependency probably exists but is not explicitly modeled.
We also assume (still on the basis of exchanges with the stakeholders) that a high apparent prevalence ψ has negative consequences as follows :

- low if $\psi \in \Psi_0 = [0; \psi_0]$;
- medium if $\psi \in \Psi_1 =]\psi_0; \psi_1]$;
- high if $\psi \in \Psi_2 =]\psi_1; 1]$,

where ψ_0 and ψ_1 are two defined levels of prevalence. Of course, $\psi_0 < \psi_1$.
At every period, the plant analyses n samples. The number x of positive samples among n follows a Binomial distribution with parameters n and ψ. The decision maker wants to find the best set-up $e = n$ which minimizes its loss function. Each period, a decision d is taken among the followings:

- d_0: contamination is under control, no correction needed;
- d_1: contamination is a little bit too high, a slight correction is needed;
- d_2: contamination is too high, a big correction is needed.

We assume that the negative consequence can be quantified in euros as follows:

- for the decision taken by the stakeholder : 0 for d_0, C_1 for d_1 and C_2 for d_2;
- for the cost of a fine to be paid to the client : 0 if $\psi \in \Psi_0$, K_1 if $\psi \in \Psi_1$ and K_2 if $\psi \in \Psi_2$.

When decision d_0 has been taken, the fine is 0 if $\psi \in \Psi_0$, K_1 if $\psi \in \Psi_1$ and K_2 if $\psi \in \Psi_2$. As decisions d_1 and d_2 should normally reduce the prevalence, the client reduces the fine when these decisions are taken: when d_1 is chosen, the fine is lowered by $1 - \alpha$ and by $1 - \beta$ when d_2 is chosen, where α and β are two known numbers and $0 < \alpha < \beta$.
To all these costs, the sampling cost (c for each detection analysis) is added. The loss function $L(\delta(x), \psi)$ is described on Table 1.

Decisions ψ	d_0	d_1	d_2
$\psi \in \Psi_0$	$0+cn$	$C_1 + cn$	$C_2 + cn$
$\psi \in \Psi_1$	$K_1 + cn$	$C_1 + \alpha K_1 + cn$	$C_2 + \beta K_1 + cn$
$\psi \in \Psi_2$	$K_2 + cn$	$C_1 + \alpha K_2 + cn$	$C_2 + \beta K_2 + cn$

Table 1: Values taken by the cost function L depending on the decision d taken by the plant and the prevalence ψ of the lot.

As $f(x|\psi) = \mathcal{B}in(n, \psi)$ and $\pi(\psi) = \mathcal{B}eta(a, b)$, the posterior distribution of ψ is $\pi(\psi|x) = \mathcal{B}eta(a + x, b + n - x)$ and the marginal distribution of x is:

$$\begin{aligned} f(x) &= \frac{\pi(\psi) f(x|\psi)}{\pi(\psi|x)} \\ &= \frac{\Gamma(a+b)\Gamma(a+x)\Gamma(b+n-x)\Gamma(n+1)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)\Gamma(n-x+1)\Gamma(x+1)}. \end{aligned}$$

In this example, it is easy to calculate L and r. It appears that:

$$\begin{aligned} \delta_n(x) = d_0 &\quad \Leftrightarrow x \leq c_1 \\ \delta_n(x) = d_1 &\quad \Leftrightarrow c_1 < x \leq c_2 \\ \delta_n(x) = d_2 &\quad \Leftrightarrow x > c_2, \end{aligned}$$

where c_1 and c_2 are integers between 0 and n and $c_1 < c_2$. For a given value of n, $L(\psi, \delta_n(x) = d_0) < L(\psi, \delta_n(x) = d_1)$ when $x = c_1$ and $L(\psi, \delta_n(x) = d_0) > L(\psi, \delta_n(x) = d_1)$, when $x = c_1 + 1$. The value of c_2 is calculated in the same way but d_0 is replaced by d_1 and d_1 by d_2. As a consequence, c_1 and c_2 are functions of n. Here, Bayesian risk is equal to:

$$\begin{aligned} r(\delta_n) &= cn + \sum_{x=0}^{c_1} (K_1 P_1 + K_2 P_2) f(x) \\ &+ \sum_{x=c_1+1}^{c_2} (C_1 + (\alpha K_1 P_1 + \alpha K_2 P_2)) f(x) + \sum_{x=c_2+1}^{n} (C_2 + (\beta K_1 P_1 + \beta K_2 P_2)) f(x), \end{aligned}$$

where $P_i = \mathbb{P}(\psi \in \Psi_i | x, n)$, $i = 1, 2$.
We asked an expert to estimate costs c, C_1, C_2, K_1 and K_2. The task was the following:

- define the slight and the big corrections needed to lower the prevalence;
- describe the fines the client charges the plant;
- give a cost to each correction and each kind of fine.

For instance, the expert said that a slight correction would be that the plant cleans more the workshop during a week and also controls 20 lots of its suppliers during a month. The cost of this action is $C_1 = 4250$ euros. We gave values to coefficients α and β and to parameters of distributions a and b (see Table 2). Table 3 shows values of c_1 and c_2 for different values of n. When n increases, c_1 (resp. c_2) gets closer to ψ_0 (resp.ψ_1). This is undoubtedly due to the

c	C_1	C_2	K_1	K_2	α	β	ψ_0	ψ_1	a	b
20	4250	14000	6200	90050	0.3	0.15	0.2	0.6	2	3

Table 2: Values of different costs (c for detection analysis, C_1 and C_2 for decision d_1 and d_2, K_1 and K_2 for fines), coefficients α and β, prevalence levels ψ_0 and ψ_1, distribution parameters for the prior of prevalence ψ a and b.

fact that more information is collected when n increases so the prevalence ψ over the period is estimated better. As consequence, the decision taken is close to the best decision if ψ was known (so no uncertainty). The best decision if ψ is knows is $\delta'(\psi) = d_i$ when $\psi \in \Psi_i$ for $i = 0, \ldots, 2$. The Bayesian risk depends only on n so the value of n which minimizes $r_n(\delta)$ is

n	c_1	c_2
5	0	4
10	2	7
50	15	32
100	27	63
1000	225	609

Table 3: Values of c_1 and c_2 for different sample sizes n.

determined by calculating $r_n(\delta)$ when n varies. Table 4 indicates $r_n(\delta)$ for increasing values of n. With the numerical values given in Tab. 2, the minimum is reached for $n = 17$, where $c_1 = 5$ and $c_2 = 12$.

n	$r_n\delta \times 10^3$
5	9966
15	9661
16	9655
17	9652
18	9657
19	9657
30	9733

Table 4: Values of c_1 and c_2 for different sample sizes n.

This exercise illustrates the type of calculations that can be performed under this theory. This is a very powerful tool to conceptualize the types of decisions taken by a plant, even if the elicitation is difficult, given that most producers often maintain their sampling plans as they were under previous regulations without questioning them, and are not trained to justify them with such arguments.

By experience, under this framework, the two most important and difficult issues to be discussed in detail with the plant are on the one hand the duration of the period over which it is

relevant to consider the prevalence, and on the other hand the nature (and costs) of the negative consequences, both of them being linked together.

Bibliography

Commission Regulation (EC) No 2073/2005. (2005) On microbiological criteria for foodstuffs. *Official Journal of the European Union*; L338:1-26.

Berger, J.O. (1985). Statistical decision theory and Bayesian analysis. Springer Verlag, New York, second edition.

Commeau, N. Cornu, M. Albert, I. Denis, J.-B. and Parent, E. (2012). Listeria in food, risk assessment accounting for between and within batch variability: a Bayesian modeling. *Risk Analysis*, in press.

Robert, C.P. The Bayesian choice (2001). Springer

Analyse d'incertitude d'un modèle de culture : démarche et illustration sur deux cas d'étude.

Uncertainty analysis of a crop of culture: approach and illustration of two case studies.

François Brun[1], Nathalie Keussayan[2], Arnaud Bensadoun[3], Jacques-Eric Bergez[2], Bernard Lacroix[4], Philippe Debaeke[2], Luc Champolivier[3], Jean-Pierre Palleau[3] & Emmanuelle Mestries[3], Daniel Wallach[2].

[1] *ACTA, INRA UMR 1248 AGIR, B.P. 52627, 31326 Castanet Tolosan, France*
E-mail : *francois.brun@acta.asso.fr*
[2] *INRA, UMR 1248 AGIR, B.P. 52627, 31326 Castanet Tolosan, France*
[3] *CETIOM, UMR 1248 AGIR, B.P. 52627, 31326 Castanet Tolosan, France*
[4] *ARVALIS - Institut du végétal, 6 chemin de la côte vieille 31450 Baziège, France*

Résumé
Les modèles de système sont des outils de plus en plus utilisés en agronomie. En particulier, les modèles de culture visent à représenter le fonctionnement du système étudié en simulant la dynamique de développement et de croissance des cultures. Ces modèles restent des représentations simplifiées d'une réalité bien plus complexe qu'est un agro-système. Ainsi, il serait intéressant d'associer un indice de fiabilité aux simulations réalisées afin de prendre en compte l'incertitude lors de leur utilisation, notamment pour comparer des scénarios et prendre des décisions. D'abord, nous présentons une démarche d'analyse d'incertitude pour calculer et associer cet indice de fiabilité aux simulations. Puis, nous illustrons la mise en pratique concrète de cette démarche, ainsi que les difficultés rencontrées sur deux cas d'étude : un modèle de culture du maïs utilisé pour comparer les stratégies d'irrigation (MODERATO) et un de tournesol utilisé pour l'évaluation variétale (SUNFLO).
Mots-clés : Analyse d'incertitude, modèle de culture, paramètre, méthode bayésienne.

Abstract
The system models are tools increasingly used in agronomy. In particular, the crop models designed to represent the evolution of the system studied by simulating the dynamics of development and growth of crops. These models remain simplified representations of reality of the very complex agricultural system. So, it would be interesting to associate a reliability index to simulations in order to take into account the uncertainty in their use, to compare scenarios. First, we present a procedure to realize the uncertainty analysis, to compute and to match the index of reliability to simulations. Then, we illustrate the practical implementation of this approach and the difficulties encountered on two case studies: a model of corn used to compare irrigation strategies (MODERATO) and one on sunflower used for the varietal evaluation (SUNFLO).
Keywords : Uncertainty analysis, crop model, parameter, residual variance, Bayesian method.

1. Introduction

Les modèles de système pour l'agronomie s'imposent de plus en plus comme des outils incontournables pour les chercheurs et ingénieurs du développement agricole. Par modèle de système on entend un modèle qui considère explicitement l'objet modélisé comme un ensemble d'éléments et de processus qui interagissent, et qui est basé au moins pour partie sur le comportement de ces éléments. Ainsi, les modèles de cultures considèrent la croissance et le développement de la plante en interaction avec le sol et l'atmosphère. Ils sont notamment utiliser pour comparer des scénarios.

Ces modèles de culture représentent de manière simplifiée des systèmes vivants très complexes qui font intervenir un très grand nombre de processus. L'expérience montre qu'ils présentent des niveaux d'erreurs de prédiction élevés, mais l'analyse des incertitudes sur les prédictions reste relativement difficile à appréhender, puis à exploiter en routine pour de nouvelle prédiction. Les explications sont nombreuses : leur nature (dynamique), leur complexité (nombreux paramètres jusqu'à 200 dans certains cas) et les structures de données pour leur analyse ou leur utilisation (corrélation spatiale, temporelle). Pourtant, quelque soit leur état actuel, un point essentiel est de connaître le niveau de fiabilité des prédictions de ces modèles car les utilisateurs des modèles ont besoin de connaître le niveau de précision des modèles afin de prendre en considération cette information dans l'analyse des résultats.

Dans ce travail, nous proposons une démarche pour quantifier l'incertitude des sorties du modèle et l'appliquons à deux cas d'étude particuliers : un modèle bio-décisionnel de culture du maïs (MODERATO utilisé pour la recherche de stratégies optimales de conduite de l'irrigation en volume limité et un modèle de culture du tournesol (SUNFLO) utilisé pour simuler la réponse des variétés de tournesol à l'environnement et à la conduite de culture. Nous illustrons la mise en pratique concrète de cette démarche sur nos deux cas d'étude, ainsi que les difficultés rencontrées. Une attention particulière est apportée à la présentation des résultats à destination des utilisateurs.

2. Proposition d'une démarche d'analyse d'incertitude

Afin de mener l'analyse d'incertitude, nous proposons la démarche suivante présentée dans le tableau 1.

Etapes	Tâches
Définition des besoins et des contraintes	1) explicitation des variables d'intérêt (par rapport à utilisation)
	2) choix d'indicateurs d'incertitude pour les variables d'intérêt
	3) identification des sources d'incertitude
	4) caractérisation des informations disponibles
Analyse d'incertitude	5) quantification des sources d'incertitudes
	6) propagation de l'incertitude (Distribution de chaque variable d'intérêt)
	7) « meilleure réponse » (valeur moyenne de la variable d'intérêt)
	8) valeur des indicateurs d'incertitude
Analyse des résultats Vérification des hypothèses	9) analyse des contributions des différentes sources d'incertitude
	10) vérification avec des données
	11) explicitation et analyse des hypothèses

Tableau 1. Proposition d'une démarche opérationnelle et générique pour associer un niveau d'incertitude aux sorties d'un modèle.

Elle est proche de propositions d'autres auteurs (de Rocquigny et al., 2008 avec des applications dans le monde industriel ou Warren-Hicks et al., 2010 avec des applications en environnement). Cette démarche est assez générique et met en avant la nécessité de bien préciser différents éléments nécessaire pour préciser l'objectif de l'analyse d'incertitude et faciliter la vérification des résultats. Nous proposons une vue relativement linéaire, mais il faut bien considérer la mise en œuvre de cette démarche comme un travail de modélisation itératif avec des boucles de progrès pouvant concerner les différents étapes.

3. Description des deux cas d'étude

Les deux modèles considérés sont des modèles de culture qui simulent jour après jour la progression de l'enracinement, l'élaboration de la surface foliaire et de la biomasse aérienne de la culture en fonction des contraintes de température, de rayonnement et d'eau (et d'azote pour SUNFLO) subies par la culture. Les données d'entrée utiles au modèle sont liées au milieu (sol, climat), à la conduite de culture (date de semis, fertilisation azotée, irrigation,…) et au génotype (phénologie notamment).

3.1 MODERATO. Modèle bio-décisionnel de culture du maïs

Ce modèle est utilisé pour la recherche de stratégies optimales de conduite de l'irrigation du maïs en volume limité en évaluant les conséquences de ses stratégies en prenant en compte l'incertitude climatique (Bergez et al., 2001). Le modèle est composé d'un modèle de culture et d'un modèle de décision. Le modèle de culture est celui décrit dans Wallach et al. (2001) avec une légère modification des formalismes pour la sénescence.
Il permet de prédire le rendement du Maïs, mais aussi d'autres variables permettant le fonctionnement du modèle de décision ou encore des variables utiles pour estimer les paramètres (Biomasse et Indice de Surface foliaire (LAI)). Le génotype n'est décrit que par quelques variables d'entrée pour décrire sa phénologie.

3.2 SUNFLO. Modèle de culture du tournesol

Le modèle de culture SUNFLO est utilisé pour la simulation de la réponse des variétés de tournesol à l'environnement et à la conduite de culture (Figure 1). Il a été développé dans le cadre de l'UMT Tournesol INRA-CETIOM (Casadebaig et al., 2011). Il est codé sous RECORD-VLE (Quesnel et al., 2009).
Il permet ainsi de prédire le rendement et la teneur en huile du tournesol à l'échelle d'une parcelle et calcule des indicateurs de stress subis par la culture.
SUNFLO a été développé pour représenter de manière dynamique l'interaction entre un génotype, son milieu (sol, climat) et la conduite culturale. Son originalité tient au fait qu'il permet de tenir compte des différences entre les génotypes sur différents critères (leur phénologie, leur architecture, leur comportement face au stress hydrique et la manière dont elles remplissent leur graines) et au fait que ces caractéristiques phénotypiques sont accessibles car mesurables simplement dans les essais d'évaluation des génotypes au champ ou en serre (protocole d'évaluation de la tolérance au stress hydrique).

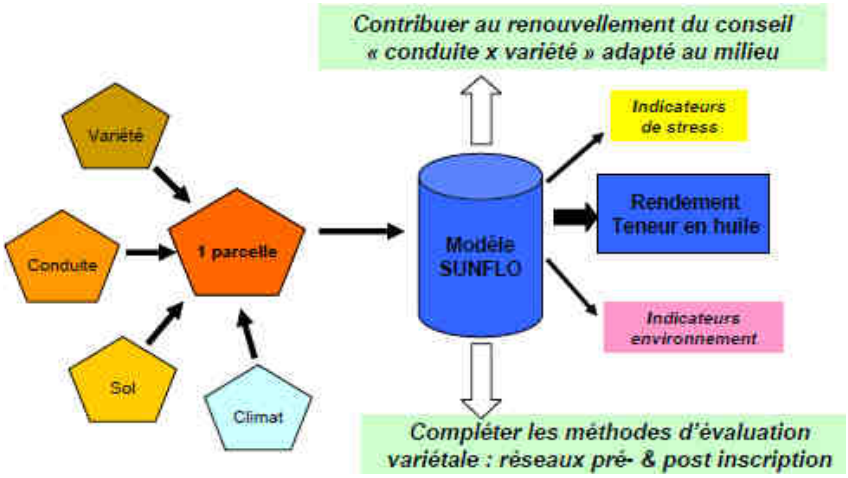


Figure 1: Représentation du fonctionnement du modèle de culture du tournesol SUNFLO

4. Mise en pratique sur les cas d'étude

Dans cette partie, nous illustrerons les différentes étapes à l'aide des travaux réalisés sur les deux cas d'étude MODERATO et SUNFLO.

4.1 Définition des besoins et des contraintes

4.1.1 Explicitation des variables d'intérêt

Il s'agit de préciser le cas d'utilisation du modèle et de formaliser les variables d'intérêt qui intéressent réellement les utilisateurs pour comparer leur scénarios.

Pour MODERATO, les variables d'intérêt sont présentés dans la table 2 : il s'agit de variable concernant la sortie principale du modèle qu'est le rendement. Ces différentes variables intéressent les utilisateurs pour appréhender notamment le comportement moyen ou sa variabilité.

	description
$C^{random\ year}$	Rendement, climat aléatoire. $p_{i,t,k}^{(s)}$ avec i choisi aléatoirement {1,…,49}
$C^{year\ i}$	Rendement, climat d'une année particulière i. $p_{i,t,k}^{(s)}$ avec i connu
C^{ave}	Rendement moyenné sur les années. $(1/49)\sum_{i=1}^{49} p_{i,t,k}^{(s)}$
C^{sd}	Ecart type du rendement. $\sqrt{\frac{1}{48}\sum_{i=1}^{49}\left[\left(p_{i,t,k}^{(s)} - (1/49)\sum_{i=1}^{49} p_{i,t,k}^{(s)}\right)^2\right]}$
C^{poor}	Nombre d'année avec des rendements sous le seuil de 6t/ha. $\sum_{i=1}^{49} 1\{p_{i,t,k}^{(s)} < 6\}$

Table 2. Variables d'intérêt sur le rendement pour évaluer les stratégies d'irrigation. $p_{i,t,k}^{(s)}$ est la prédiction pour la stratégie s, l'année i, le vecteur de paramètre $\theta^{(t)}$ et l'erreur résiduelle $\varepsilon_k^{(t)}$.

Pour SUNFLO, sur le tournesol, dans un premier temps nous nous intéressons au rendement moyen et à la teneur en huile moyenne (équivalent à C^{ave} de la table 2). Les indicateurs sur la variabilité interannuelle intéressent aussi les utilisateurs, mais ils restent à définir précisément.

4.1.2 Choix d'indicateurs d'incertitude pour les variables d'intérêt

Pour chacune des variables d'intérêt, nous aurons comme indicateur d'incertitude complet la distribution des valeurs issue de l'analyse. Par ailleurs, nous proposons d'utiliser des intervalles de confiance comme indicateur synthétique qui sont plus lisibles pour un utilisateur final.

4.1.3 Identification des sources d'incertitude

Les principales hypothèses que nous faisons sont de considérer comme source d'incertitude (Figure 2):
- les conditions climatiques à venir (au sens « variabilité ») (X_{incert}).
- les valeurs des paramètres des modèles (au sens « manque de connaissance » essentiellement)(θ_{incert}).
- la variance résiduelle liée au fait que le modèle n'explique pas tout (au sens « manque de connaissance » essentiellement)(ε_r).

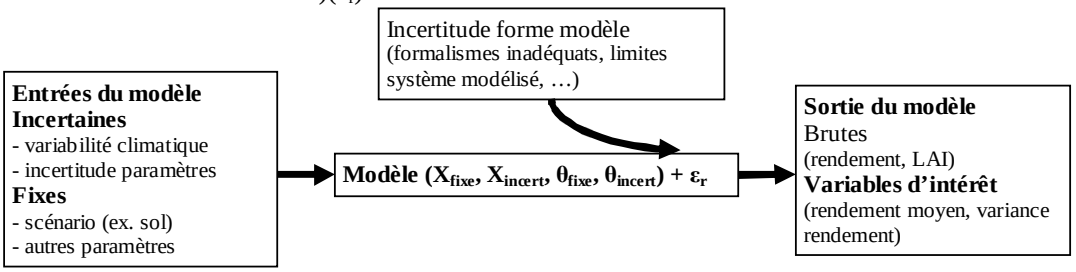


Figure 2. Modélisation de l'incertitude pour nos deux cas d'étude

4.1.4 Caractérisation des informations disponibles

Dans les deux cas, nous avons deux sources d'information principale à notre disposition pour quantifier les sources d'incertitude et évaluer la qualité des indicateurs :
- la littérature scientifique, avec notamment, des valeurs disponibles pour différents paramètres des modèles.
- les données d'expérimentation, avec dans les deux cas des variables observés « statiques » (comme le rendement ou la teneur en huile) et des suivis dynamiques (biomasse aérienne et indice foliaire).

Pour MODERATO, on des données sur du maïs cultivé dans 11 sites différents dans le sud ouest France avec des réserves utiles allant de 40 à 588 mm et 1 à 4 ans de mesures (entre 1986 et 1997). Pour chacun des site-année, il y a 2 à 10 traitements d'irrigation, ce qui nous fait un total de 81 site-année-irrigation avec le rendement et la biomasse finale, ainsi que dans certains cas un suivi dynamique des biomasses et des indices foliaires. Pour le modèle, on a les données d'entrée (variété, date de semis, sol, météo et calendrier d'irrigation) pour ces 81 unités de simulation.

Dans le cas de SUNFLO, nous avons un premier jeu de des données avec 167 combinaisons site-année-conduite-génotypes avec 167 rendements et 167 teneurs en huile ainsi qu'un suivi dynamique et 663 mesures d'indice foliaire et 416 de biomasse. Par ailleurs, on a en plus un jeu de données d'expérimentations sur certains processus du modèle, menées dans le but de paramétrer directement les génotypes : il s'agit de mesure en conditions contrôlées (expansion des feuilles et transpiration) ou au champ (architecture de la plante par exemple en condition potentielle).

4.2 Analyse d'incertitude

4.2.1 Quantification des sources d'incertitudes

Pour quantifier l'incertitude climatique, nous utilisons la variabilité climatique observé dans des séries climatiques du passé. Ainsi, pour MODERATO, nous utiliserons à titre d'exemple, 49 années de la station météo de Blagnac (près de Toulouse).

Pour quantifier l'incertitude sur les paramètres, nous avons la démarche suivante :
1- Autant que possible, nous utilisons la littérature disponible afin d'obtenir la distribution des paramètres.
2- Quand elle n'est pas disponible, nous définissons des bornes à partir de l'expertise des concepteurs des modèles.
3- L'ensemble de cette information, nous donne l'**information *a priori***.
4- Nous utilisons les données expérimentales disponibles **pour affiner cette quantification**.

Pour affiner la quantification de l'incertitude sur les paramètres, nous estimons les paramètres à partir des données expérimentales disponibles.
Une première possibilité est l'approche bayésienne qui permet d'obtenir la distribution de certains paramètres des modèles à partir des données disponibles. C'est de loin l'étape la plus complexe et longue en temps de calcul dans la démarche. L'approche bayésienne part d'une distribution préalable des paramètres du modèle, et mises à jour sur la base des données d'étalonnage. Le résultat de l'étalonnage bayésienne est une distribution conjointe des paramètres du modèle et les variances de l'erreur résiduelle, appelée la distribution a posteriori. Il ya encore seulement quelques exemples d'estimation bayésienne de paramètres pour les modèles de cultures (Iizumi et al, 2009b; Ceglar et al, 2011), ou des modèles de forêt (Van Oijen et al, 2005). Pour cela, nous utilisons un l'algorithme Metropolis-Hastings within Gibbs codé sous R (pour plus de détail, voir Wallach et al, soumis).
Pour MODERATO, 15 paramètres sont estimés ainsi, les 14 autres étant fixé. La Table 3 résume les distributions a priori et a posteriori pour ces paramètres.

Abréviation	Description [Unité]	Distrib. a priori (uniforme)			Distrib. a posteriori	
		Binf	Bsup	e.t.	moyenne	e.t.
a2sen	Stress hydrique sur sénescence [-]	0	1	0.29	0.25	0.042
a3sen	Stress hydrique sur sénescence [-]	1	2	0.29	1.79	0.093
himax	Indice de récolte (*) [-]	0.45	0.55	0.029	0.51	0.021
p1logi	Paramètre pour l'indice foliaire (*) [-]	0.65	0.99	0.098	0.66	0.013
p1sen	Paramètre de sénescence [-]	0.0011	0.0021	0.00028	0.0019	0.0002
p2logi	Paramètre pour l'indice foliaire [(℃ days)- 1]	0.007	0.013	0.00087	0.0086	0.00007
p2sen	Paramètre de sénescence [-]	4.2	7.8	1.04	5.9	0.11
r1hi	Transpiration sur indice de récolte [-]	1	2	0.29	1.8	0.13
r1rue	Transpiration sur conversion du rayonnement [-]	0.0001	1	0.029	0.83	0.093
r1sf	Transpiration sur indice foliaire (*) [-]	0.4	1.2	0.23	0.96	0.19
r2hi	Transpiration sur indice de récolte [-]	1	2	0.29	1.1	0.068
r2rue	Transpiration sur conversion du rayonnement [-]	0.0001	1	0.029	0.95	0.040
r2sf	Transpiration sur indice foliaire (*) [-]	0.4	1.2	0.23	0.63	0.12
rue1	Conversion du rayonnement en biomasse (*) [g/MJ]	3	4	0.29	3.0	0.019
rue2	Conversion du rayonnement en biomasse (*) [g/MJ]	3	4	0.29	3.03	0.037
σ_1^2	Erreur résiduelle du rendement (variance)				1.3	0.11
σ_2^2	Erreur résiduelle de la biomasse (variance)				2.3	0.11
σ_3^2	Erreur résiduelle de l'indice foliaire (variance)				0.81	0.032

Table 3. Distribution des paramètres de MODERATO. e.t. : ecart type.

Pour SUNFLO, cette étape relativement couteuse en temps n'a pas encore été à son terme et nécessite encore des calculs.

Une deuxième possibilité est d'estimer directement les distributions des paramètres à partir de données, ce qui est possible lorsque l'on a des données sur les processus concernés. Cela a été réalisé uniquement pour SUNFLO et les 13 paramètres génotypiques. Dans la Table 4, on retrouve un extrait des valeurs et de leur variation pour deux paramètres de stress à titre d'illustration.

Variétés	a_LE	sd a_LE	CV a_LE	a_TR	sd a_TR	CV a_TR
Airelle	-2.7250	0.6870	0.2521	-6.8300	0.9040	0.1324
Euroflor	-15.5720	4.4810	0.2878	-6.1410	1.3100	0.2133
Frankasol	-6.7880	2.3310	0.3434	-7.2350	0.4700	0.0650
Heliasol	-5.2190	1.0260	0.1966	-5.2230	0.9250	0.1771
INRA6501	-4.5970	0.6110	0.1329	-7.2930	1.1030	0.1512
Melody	-3.8100	0.2940	0.0772	-5.6510	0.4730	0.0837
Prodisol	-4.2520	0.4770	0.1122	-7.1320	0.7090	0.0994

Table 4 : Extrait des moyennes, écart-types et coefficients de variation des paramètres de stress par génotype (au total, plus d'une vingtaine de génotype paramètre de cette manière).

Pour quantifier l'incertitude sur lié à la structure du modèle, à savoir l'erreur résiduelle, nous utilisons les résultats de l'estimation bayésienne qui nous permet d'accéder à la distribution de l'erreur résiduelle en même temps qu'à la distribution conjointe des paramètres (Table 3, les trois dernières lignes). Souvent on ne prend pas en compte l'erreur résiduelle, qui n'est pourtant pas négligeable dans notre cas, pour le calcul d'incertitude sur les prédictions, mais uniquement l'effet des incertitude sur les paramètres du modèle (par exemple Iizumi et al., 2009). Ce point nous semble particulièrement intéressant à mentionner.

4.2.2 Propagation de l'incertitude

Enfin, nous combinons ces différentes sources d'incertitude dans un plan d'expérience afin de quantifier l'incertitude sur les sorties du modèle sur nos variables d'intérêt.

Pour SUNFLO et MODERATO, nous effectuons les simulations sur les unités de simulations correspondant aux données pour vérifier les calculs, mais ensuite nous pouvons réaliser des simulations pour de nouvelles situations. C'est ce qui est fait avec MODERATO pour lequel nous avons défini 3 stratégies différentes d'irrigation, utilisant le même volume limité d'eau d'irrigation (125 mm). Pour expliquer de manière simple : la stratégie « Floraison » vise à répartir l'irrigation autour de la floraison (en fonction de la phénologie), la stratégie « Tardif » vise à retarder le plus possible l'irrigation afin d'avoir de l'eau en fin de saison, la stratégie « Précoce » vise à utiliser l'eau dès les premiers événements de stress hydrique.

4.2.3 « Meilleure réponse » (valeur moyenne de la variable d'intérêt)

Pour MODERATO, on retrouve les valeurs moyennes des différentes des variables considéré dans la Table 5. Pour SUNFLO, les résultats ne sont pas encore disponibles.

4.2.3 Valeur des indicateurs d'incertitude

Un exemple de sortie pour MODERATO est dans la Figure 3. On y voit la dynamique de la surface foliaire qui sert à calibrer le modèle avec l'intervalle de confiance à 90% sur deux unités de simulation et les données correspondante. Les distributions cumulatives se différencient fortement pour les 3 premières variables d'intérêt (rendement pour une année climatique non connu « by year », pour une année donne « 1997 » ou en moyenne sur 49 années) avec une incertitude qui se réduit pour la stratégie d'irrigation « floraison » (Figure 3, à droite).

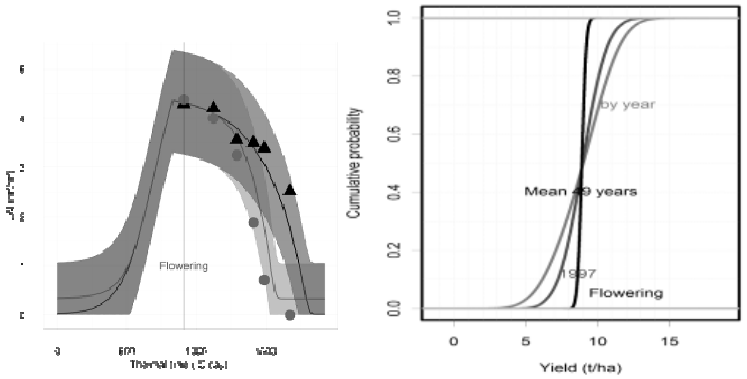


Figure 3. Exemple de résultats pour MODERATO. A gauche : Evolution de la surface foliaire (LAI) et intervalle de confiance à 90% pour deux niveaux d'irrigation (les points correspondent à des mesures). A droite : fonctions de distribution cumulative pour différents critères d'intérêt.

Un des principaux résultats à mentionner est le fait que suivant la variable d'intérêt, et donc, l'utilisation du modèle, nous pouvons obtenir des incertitudes sur les prédictions avec des ordres de grandeur très différents (table 5).

Variables d'intérêt	stratégies	Valeur moyenne	Ecart-Type	Intervalle de confiance à 90%
$C^{random\ year}$ [t/ha]	Floraison	8.9	2.0	5.5-12.1
	Tardif	7.9	2.3	4.1-11.6
	Précoce	7.6	2.8	2.9-11.9
C^{1997} [t/ha]	Floraison	8.9	1.4	-
	Tardif	7.9	1.4	-
	Précoce	7.6	1.3	-
C^{ave} [t/ha]	Floraison	8.9	0.2	8.5- 9.3
	Tardif	7.9	0.3	7.5-8.3
	Précoce	7.6	0.2	7.2-8.0
C^{sd} [t/ha]	Floraison	2.0	0.2	1.7-2.4
	Tardif	2.3	0.2	2.0 -2.7
	Précoce	2.8	0.2	2.5-3.1
C^{poor} [years out of 49]	Floraison	4.2	1.8	1-7
	Tardif	10.9	2.5	7 -15
	Précoce	14.8	2.4	11-19

Table 5. Résultats pour les intervalles de confiance pour les différentes variables d'intérêt.

Ainsi, l'incertitude se réduit fortement lorsque l'on cherche à prédire non pas une année climatique donnée (C^{1997}) mais un rendement moyen (C^{ave}), qui semble une variable d'intérêt particulièrement

pertinent pour comparer des stratégies : ce qui nous intéresse c'est bien l'espérance du rendement (C^{ave}) et éventuellement la variabilité (par exemple C^{sd} ou C^{poor}) pour prendre sa décision.

4.3 Analyse des résultats et vérification des hypothèses de travail

4.3.1 Analyse des contributions des différentes sources d'incertitude

Le principe est de quantifier le poids relatifs des différentes sources d'incertitude prises en compte. Nous n'avons pas mené cette analyse pour le moment.

3.3.2 Vérification avec des données

Par ailleurs, les données disponibles, nous permettent dans certains cas de vérifier la vraisemblance de l'incertitude calculée (par exemple les intervalles de confiance) et, donc, de vérifier que les hypothèses faites conduisent à des résultats réalistes malgré le fait que ces hypothèses restent grossières et peu satisfaisante dans le cas de ces modèles dynamiques complexes (Table 6 pour les vérifications sur MODERATO).

Variables	Pourcentage dans IC50	Pourcentage dans IC90
Rendement	53%	91%
Indice Foliaire (LAI)	51%	89%
Biomasse	65%	95%

Table 6. Pour MODERATO, pourcentage de valeurs mesurées dans les intervalles de confiance à 50% (IC50) ou à 90% (IC90).

4.3.3 Explicitation et analyse des hypothèses

Il s'agit notamment des hypothèses sur les sources d'incertitudes et la forme de la modélisation de l'erreur (loi normale, centrée) : nous savons que ce ne sont pas des hypothèses complètement vraisemblables, néanmoins, il n'est pas évident de proposer autre chose de simple et les résultats en terme d'incertitude semblent réalistes.

4. Conclusion

La démarche proposée présente l'intérêt d'inciter à bien expliciter l'objectif de l'analyse d'incertitude (variables d'intérêt et indicateur d'incertitude) ainsi que les cas d'utilisation du modèle et à mettre en face les informations disponibles afin de choisir les méthodes pour quantifier les sources d'incertitude en conséquence. Enfin, il faut penser aussi à vérifier la cohérence des résultats autant que possible.

Par rapport à préciser l'utilisation du modèle, on se rend compte que l'on n'a pas les mêmes niveaux d'incertitude en fonction des variables d'intérêt considérés d'un même modèle (cf. 4.2.3).

La quantification des sources d'incertitude semble être de loin la partie la plus délicate et chronophage, comme cela a déjà été souligné par d'autres auteurs : il faut faire les choix méthodologiques en conséquence pour pouvoir la réaliser.

Financement

Ces travaux ont été réalisés dans le cadre du projet «Associer un niveau d'erreur aux prédictions des modèles mathématiques pour l'agronomie et l'élevage » (www.modelia.org) mené par l'ACTA et ses partenaires (2010-2012) et financé par le « Compte d'affectation spécial pour le développement agricole et rural » du Ministère de l'Agriculture et de la Pêche.

Bibliographie

Bergez, J.E., Debaeke, P., Deumier, J.M., Lacroix, B., Leenhardt, D., Leroy, P. & Wallach, D. (2001). MODERATO: an object-oriented decision tool for designing maize irrigation schedules. *Ecological Modelling*, 137, 43-60.

Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A , Champolivier, L., Debaeke, P. (2011). SUNFLO, a model to simulate genotype-specific performance of sunflower crop in contrasting environments. *Agricultural Forest Meteorology*, 151, 163-178.

Ceglar, A., Crepinsek, Z., Kajfez-Bogataj, L. & Pogacar, T. (2011). The simulation of phenological development in dynamic crop model: The Bayesian comparison of different methods. *Agricultural and Forest Meteorology*, 151, 101-115.

Iizumi, T., Yokozawa, M. & Nishimori, M. (2009). Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. *Agricultural and Forest Meteorology* 149, 333-348.

Quesnel, G., Duboz, R. & Ramat, E (2009). The Virtual Laboratory Environment - An Operational Framework for Multi-Modelling, Simulation and Analysis of Complex Systems. *Simulation Modelling Practice and Theory*, 17, 641-653.

de Rocquigny, E., Devictor, N. & Tarantola, S. (2008). Uncertainty in Industrial Practice: A Guide to Quantitative Uncertainty Management. Wiley-Blackwell (an imprint of John Wiley & Sons Ltd).

Wallach, D., Goffinet, B., Bergez, J.E., Debaeke, P., Leenhardt, D. & Aubertot, J.N. (2001). Parameter estimation for crop models: a new approach and application to a corn model. *Agronomy Journal*, 93, 757-766.

Wallach, D., Keussayan, N., Brun, F., Lacroix, B. & Bergez, J.E. (soumission début 2012). Using a crop model to evaluate irrigation strategies, with emphasis on uncertainty.

Warren-Hicks, W.J. & Hart, A. (2010). Application of Uncertainty Analysis to Ecological Risks of Pesticides (Environmental Chemistry & Toxicology). CRC Press 1st ed. (April 7, 2010). 228 pp.

Estimation de l'incertitude dans les analyses de cycle de vie en élevage : apport de l'analyse de sensibilité, limites du modèle
Uncertainty estimation in life cycle analysis: contribution of sensitivity analysis, limits of the model

Marion Ferrand[1], Vincent Manneville, Sindy Moreau, Elise Lorinquer, Thierry Charroin, Alicia Charpiot, Armelle Gac, Carlos Lopez, François Brun

[1] *Institut de l'Elevage,149 rue de Bercy, 75595 Paris cedex 12, France*
E-mail : *Marion.Ferrand@idele.fr*

Résumé

Les analyses de cycle de vie intègrent rarement les incertitudes sur les facteurs d'émission et de caractérisation. Leur nombre et le manque de connaissances sur leurs variabilités en sont les principales causes. Pour cibler les facteurs les plus influents sur les émissions de gaz à effet de serre, une analyse de sensibilité de Morris a été réalisée. Sur les 10 paramètres étudiés, seuls quatre apparaissent fortement influents : les émissions de CH4 en bâtiments et au stockage. Ce résultat est utilisable pour mieux orienter des expérimentations futures. Par ailleurs, cela permet d'alléger l'analyse d'incertitude finale par méthode de Monte Carlo si seuls les facteurs influents sont retenus.

Mots-clés : analyse de cycle de vie, incertitude, analyse de sensibilité, méthode de Morris, méthode de Monte-Carlo

Abstract

The life cycle assessments rarely include the uncertainties on the emission and characterization factors. Their number and the lack of knowledge of their variability are the main causes. To identify the most influential parameters on greenhouse gases emissions, a Morris sensitivity analysis was carried out. On the 10 parameters studied, only four appear to be highly influential. This result can be used to better lead future experimentations. Besides, it allows simplifying final uncertainty analysis by Monte Carlo method if only the influential factors are retained.

Keywords : life cycle analysis, uncertainty, sensitivity analysis, resampling methods

1. Introduction

L'analyse de cycle de vie (ACV) est une méthode pour l'évaluation environnementale d'un produit ou d'un système prenant en compte plusieurs impacts environnementaux. En ACV, le terme d'incertitude est souvent utilisé au sens large et recouvre à la fois des notions d'incertitude et de variabilité. Les incertitudes rencontrées portent principalement sur les facteurs d'émissions qui permettent de convertir les données techniques en flux d'éléments ((N, P, …) et sur les facteurs de caractérisation convertissant les flux en impacts sur l'environnement (changement climatique, acidification …).

L'Institut de l'Elevage a initié en 2009, une analyse environnementale multicritère sur les exploitations des Réseaux d'Elevage pour disposer d'une première quantification des impacts environnementaux et identifier les leviers d'action qui permettraient de réduire les impacts de l'élevage sur l'environnement (Gac *et al.,* 2010). Pour réaliser l'ACV de la filière bovin lait, près de 250 facteurs d'émission et de caractérisation[1] sont nécessaires. Limiter le nombre de paramètres à caractériser est donc primordial pour estimer l'incertitude sur l'impact final, les distributions de ces paramètres étant rarement connues et le coût pour obtenir les données nécessaires très élevé.

Pour faire un premier screening dans les facteurs d'émission et caractérisation à conserver pour l'analyse d'incertitude, l'analyse de sensibilité est particulièrement bien adaptée. Cette méthode permet de sélectionner les paramètres contribuant le plus aux impacts et d'orienter les recherches bibliographiques ou les expérimentations à mettre en place. Parmi, les différents types d'analyse de sensibilité, l'analyse de Morris (Morris, 1991 - Saltelli, 2000) est utilisée pour les modèles comportant beaucoup de paramètres à explorer. Cette méthode repose sur un plan d'expérience numérique où un seul paramètre varie à la fois. Ces derniers sont classés en trois groupes selon leurs effets : peu influents, moyennement influents, et très influents. Les interactions et la linéarité des effets sont également évaluées par cette méthode. Par la suite, l'incertitude sur les impacts finaux est obtenue par simulation de Monte-Carlo en fixant les paramètres du premier groupe à leurs valeurs nominales alors que seuls les facteurs les plus influents des second et troisième groupes sont dotés d'une distribution de probabilité.

2. Matériel et Méthode

2.1 Les données

Les « Réseaux d'élevage pour le conseil et la prospective (RECP) » sont un dispositif partenarial mis en place dans les années 1980 associant des éleveurs volontaires, les chambres d'agriculture et l'Institut de l'Elevage. Il est basé sur un suivi pluriannuel d'un échantillon de 1420 exploitations herbivores réparties sur l'ensemble du territoire.
Ce dispositif a pour objectif la description des systèmes d'élevage herbivore et l'élaboration de références techniques et économiques (élaboration de cas-type par système) à destination des éleveurs et des conseillers de terrain ou à vocation collective. Il s'agit aussi d'un outil de recherche appliquée : source de connaissances et d'expertise sur les systèmes de production régionaux et nationaux.

[1] Ces facteurs d'émission et de caractérisation constituent les paramètres de notre modèle et seront appelés paramètres par la suite

La base de données rassemble les données collectées en élevage ainsi qu'un ensemble de données calculées. Les informations stockées se structurent autour des moyens de production, du fonctionnement global de l'exploitation, des performances zootechniques, des résultats économiques et d'indicateurs agro-environnementaux. Sur le plan agro-environnemental, le bilan des minéraux a été intégré dans le système d'information dès 1996, et le volet sur les consommations d'énergie a été récemment consolidé grâce à une collecte de données qui se systématise depuis 2007. La prise en compte de la durabilité est en effet un objectif d'évolution du dispositif des Réseaux d'Elevage depuis quelques années (Charroin *et al.*, 2005).
L'évaluation de l'incertitude a porté sur les exploitations spécialisées en production laitière. Elle a été menée pour l'année 2008 sur un échantillon d'exploitations homogènes en termes de données renseignées. Au final, l'analyse des résultats porte sur un échantillon de 405 exploitations bovines laitières spécialisées.

2.2 L'analyse de cycle de vie

2.2.1 Principes

L'analyse de cycle de vie est une méthode, standardisée ISO 14044, qui est utilisée pour qualifier et quantifier les impacts sur l'environnement d'un produit en prenant en compte tout son cycle vie de la production à la destruction. Cette analyse se fait dans un système dont le périmètre est bien délimité. Par exemple, pour connaître l'impact de la production agricole d'un litre de lait, le système défini sera l'exploitation laitière mais pour connaître l'impact d'un litre de lait produit et distribué, on intégrera également dans le système l'usine d'embouteillage et la surface de vente. L'ACV conduite dans cette étude s'arrête aux portes de l'exploitation.

Pour mener une ACV, une fois le système défini, différents indicateurs environnementaux sont calculés à partir de données de bases. Dans le cadre de ses travaux d'ACV, l'Institut de l'Elevage retient 3 niveaux d'indicateurs (Figure 1) :

- **Les critères de pratique et de pression** peuvent être calculés directement à partir des données techniques d'une exploitation agricole, il s'agit des bilans d'azote, phosphore et potassium (N, P, K), des consommations d'énergie et de la biodiversité
- **Les indicateurs d'émission** ou de transfert représentent les flux d'éléments (NO_3, PO_4, C, gaz à effet de serre (GES) …) vers l'environnement. Ils sont calculés à partir des variables techniques ou des indicateurs de pratique et de pression en utilisant **des facteurs d'émission** qui permettent de convertir ces données en flux. Les facteurs d'émissions sont des paramètres issus de la bibliographie ou d'expérimentations.
- **Les indicateurs d'impact** correspondent à l'impact du produit/de l'exploitation étudié(e) sur l'environnement. Ces indicateurs sont calculés à partir des indicateurs d'émission à l'aide **des facteurs de caractérisation.** Dans cette étude, nous nous centrerons sur l'impact des émissions de GES (CH_4, CO_2, N_2O) sur le changement climatique.

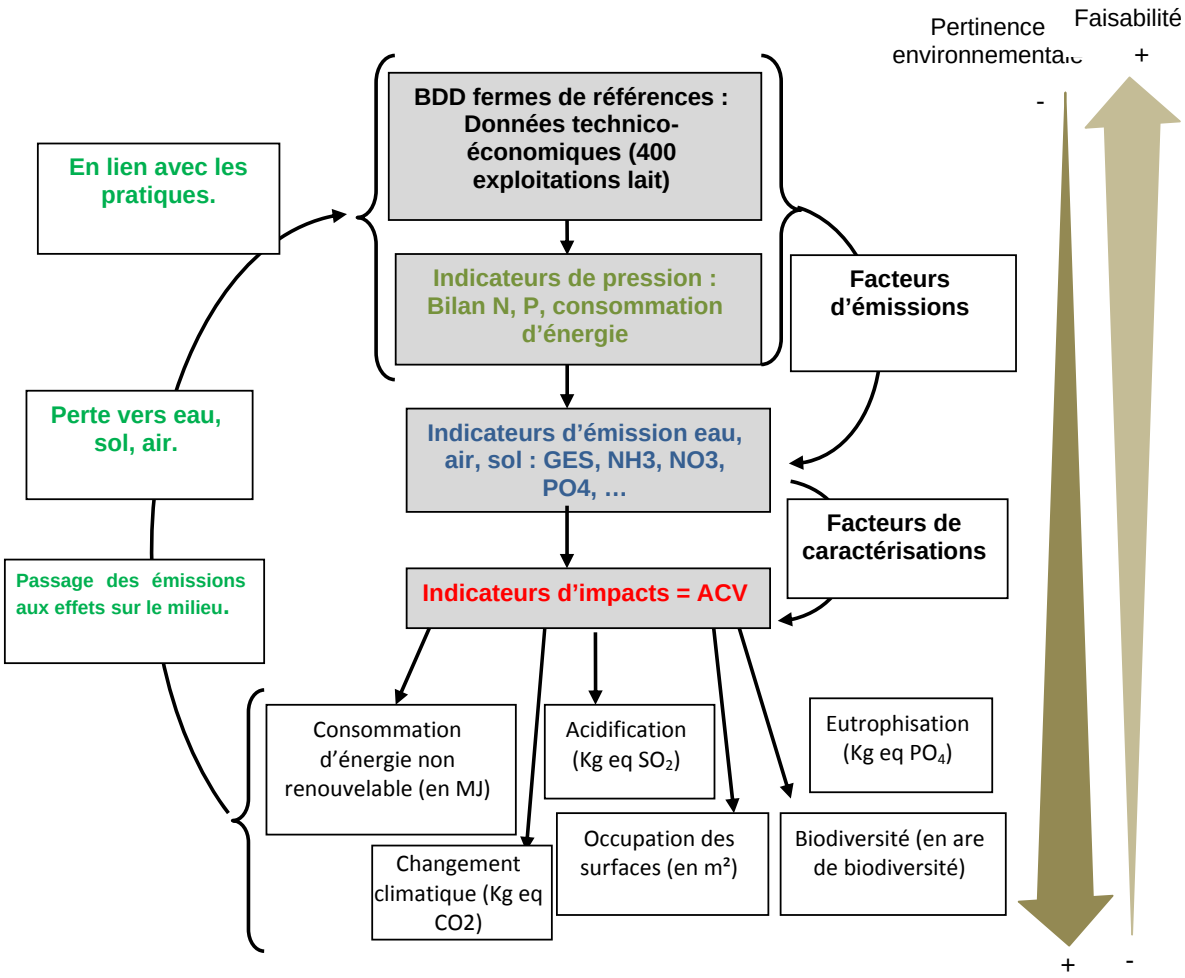


Figure 1. Les quatre étapes utilisées pour l'évaluation environnementale multicritère des exploitations d'élevage d'herbivores

2.2.2 L'incertitude en ACV

En ACV, le terme d'incertitude est souvent utilisé au sens large et recouvre à la fois des notions d'incertitude et de variabilité (Huijbregts, 1998). Quand l'incertitude est liée à un manque de connaissance sur la vraie valeur d'une quantité, elle peut être réduite par de nouvelles mesures plus fidèles (precise) et exactes (accurate). Si la variabilité est liée à la nature hétérogène des valeurs, elle ne peut pas être réduite, elle peut seulement être mieux évaluée par un échantillonnage adéquat. Les incertitudes sur les facteurs d'émission et de caractérisation (paramètres) sont très peu décrites dans les études d'ACV et rarement prises en compte. Les principales études où les incertitudes ont été considérées montrent des fourchettes d'estimation pouvant atteindre -50 à 100 % sur les émissions azotés (NH_3, NO, N_2O, N_2, NO_3) (Payraudeau *et al.*, 2007). Ces larges fourchettes sont liées à un

manque de connaissance sur les paramètres, les expérimentations référencées ne couvrant que quelques situations dans des contextes différents (Vigne *et al., 2011).*
Pour mieux évaluer l'incertitude, il semble nécessaire d'approfondir les connaissances disponibles sur les facteurs d'émissions en ciblant les paramètres les plus influents à l'aide d'une analyse de sensibilité.

2.3 L'analyse de sensibilité

L'analyse de sensibilité est une méthode complémentaire à l'analyse d'incertitudes. Elle intervient généralement après pour répartir l'incertitude du modèle entre les différentes sources (paramètres et variables d'entrée) mais peut aussi être utilisée préalablement pour limiter le nombre de paramètres à introduire dans l'analyse d'incertitude.

Saltelli (2000) distingue trois types d'analyse de sensibilité:

- les méthodes de « screening » moins coûteuses en temps mais aussi moins informatives sont utilisées sur les modèles où le nombre de paramètres est important et/ou le temps de simulation est long. L'analyse de Morris est classée dans ce groupe.
- Les méthodes dites locales basées sur des dérivées ne concernent que les modèles comportant peu de facteurs et étant linéaires.
- Les méthodes dites globales basées sur la décomposition de la variance sont très informatives mais sont peu adaptées aux modèles complexes car tous les facteurs varient de manière simultanée. Il s'agit notamment des indices de Sobol et de la méthode FAST.

Pour les modèles complexes et dont les paramètres sont mal connus, comme c'est le cas en analyse environnementale, l'analyse de sensibilité utilisée doit prendre en compte les interactions entre paramètres, la non-linéarité des effets, ne pas nécessiter de connaissances précises sur la distribution des paramètres et avoir des temps de calcul réduits (Ravalico, 2005). L'analyse de Morris même si elle ne permet pas de distinguer les interactions des effets linéaires est la méthode la plus adaptée à notre cas du fait du nombre de paramètres et de la durée des simulations (environ une minute par simulation). Les paramètres d'entrée vont donc être classés en 3 groupes d'effets :

- négligeables sur la sortie,
- linéaires et sans interactions,
- non linéaires ou avec interactions.

Pour cela, p+1 expériences répétées r fois vont être réalisées avec p le nombre de paramètres étudiés.
Pour chaque paramètre, un domaine d'étude (Ω) est défini à partir de ses bornes inférieure et supérieure. A chaque répétition, une valeur x comprise dans Ω est choisie aléatoirement et une variation Δ constante est appliquée à x.
L'effet élémentaire d'un paramètre i à un point x est égal à $d_i = \frac{f(x+\Delta) - f(x)}{\Delta}$ avec x et $x+\Delta \in \Omega$ et la fonction *f*() représentant le modèle ACV.

La moyenne des d_i sur r répétitions donne l'importance de l'effet en moyenne. L'écart-type des d_i indique si on est en présence d'interactions ou d'un effet non linéaire.

2.4. Mise en œuvre du modèle et de l'analyse de sensibilité

Le modèle utilisé est décrit dans le guide méthodologique (Schaefler, 2010). Le principe général est de calculer pour chacune des exploitations de la base de données des Réseaux d'élevage ses émissions de GES en faisant la somme des émissions des différents compartiments de l'exploitation (bâtiment, stockage, pâturage, épandage). Environ 250 paramètres sont appelés dans le modèle. Dans un premier temps, l'analyse de sensibilité n'a porté que sur 10 paramètres, les informations sur les autres paramètres n'étant pas encore disponibles. Les bornes minimales et maximales utilisées pour l'analyse de sensibilité sont déterminées d'après les différentes valeurs trouvées dans la bibliographie. Les valeurs prises sont listées dans la table 1. Le nombre de répétitions r a été fixé à 100, les résultats étant stables sur les différents essais effectués.

Paramètre	Valeur	Min	Max
FE CH4 bâtiment aire raclée	18	4	395
FE CH4 bâtiment litière accumulée	60	1,64	195
FE CH4 bâtiment caillebotis	305	209	492
FE CH4 lisier	35,5	0,32	44,8
FE NH3 bâtiment	0,12	0,03	0,37
FE N2O bâtiment aire raclée	0,00088	0	10,06
FE N2O bâtiment litière accumulée	0,71	0,68	1,4
FE N2O bâtiment caillebotis	0,48	0	5,86
FE N2O épandage	0,0157	0,0047	0,0391
Déposition atmosphérique	15	5	39

Table 1. Valeurs moyennes et bornes des paramètres pris en compte dans l'analyse de sensibilité

L'analyse de sensibilité est réalisée sous R 2.13.0. à l'aide du package sensitivity. Le modèle appelé est implémenté dans SAS version 9.2.2.

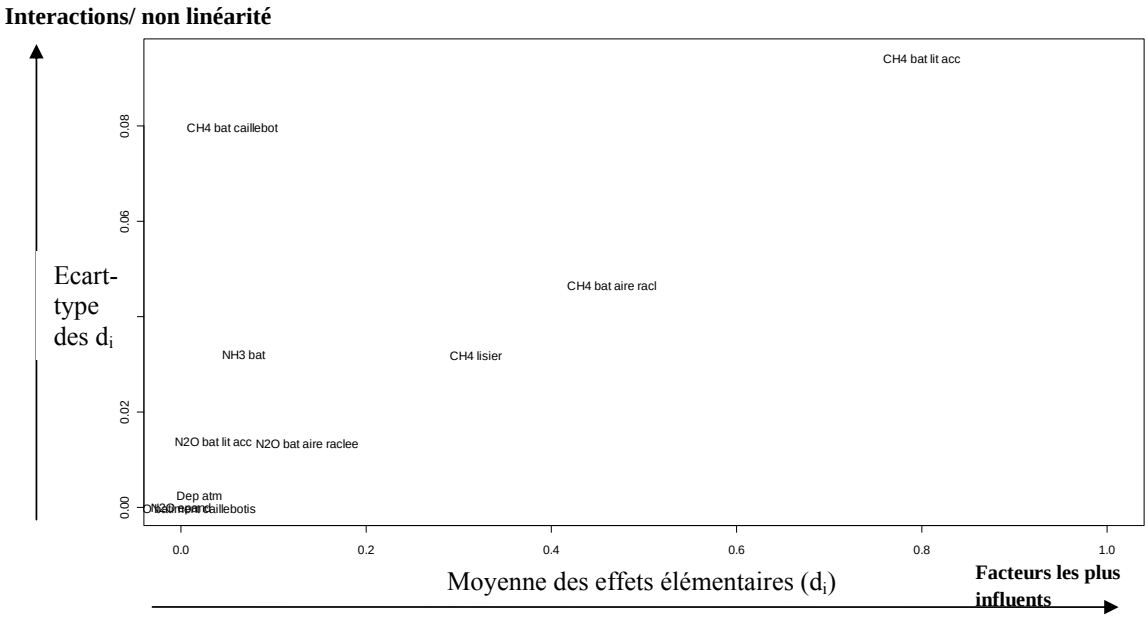
3. Résultats

Les résultats sont donnés dans le tableau 1 et visualisés graphiquement (figure 1) pour en faciliter la lecture. La moyenne des effets élémentaires (μ^*) est représentée par l'axe des abscisses. Plus un paramètre prend une valeur élevée, plus il a une influence sur les émissions de GES bruts de l'atelier lait ramenées aux 1000L. L'axe des ordonnées indique s'il y a des interactions ou des effets linéaires.

Paramètre	Moyenne des effets élémentaires	Ecart-type des effets élémentaires
FE CH4 bâtiment aire raclée	0,46	0,047
FE CH4 bâtiment litière accumulée	0,80	0,095
FE CH4 bâtiment caillebotis	0,06	0,080
FE CH4 stockage (lisier)	0,32	0,032
FE NH3 bâtiment	0,07	0,032

FE N2O bâtiment aire raclée	0,14	0,014
FE N2O bâtiment litière accumulée	0,03	0,014
FE N2O bâtiment caillebotis	0,01	0,000
FE N2O épandage	0,00	0,000
Déposition atmosphérique	0,02	0,003

Table 2. Résultats de l'analyse de sensibilité, 10 paramètres, 100 répétitions



Graphique 1. Visualisation des effets en fonction de leurs influences et de leurs interactions

Il ressort du graphe ci-dessus que les facteurs d'émission de CH4 en bâtiment et au stockage (lisier) sont les plus influents sur les émissions de GES bruts[2]. Les facteurs d'émission portant sur les gaz azotés (NH_3, N_2O) n'apparaissent pas influents quand à eux. Curieusement, l'effet du NH_3 qui n'est pas un GES n'est pas nul même s'il est faible.

Les écart-types des effets élémentaires des émissions de CH4 au bâtiment sont également élevés, ce qui signifie qu'on est vraisemblablement en présence d'un effet non-linéaire ou d'une interaction.

Pour l'analyse d'incertitude par méthode de Monte-Carlo, on ne retiendrait donc que quatre paramètres : les trois facteurs d'émission de CH4 au bâtiment[3] et celui du CH4 au stockage.

Une comparaison des résultats obtenus par analyse simplifiée (c'est-à-dire avec prise en compte de l'incertitude uniquement pour ces 4 paramètres) et par analyse exhaustive (en prenant en compte l'incertitude qu'on a sur les 10 paramètres) est en cours. Elle devrait nous permettre de quantifier la

[2] L'effet élémentaire pour les émissions de CH4 dans les bâtiments avec caillebotis est quasi nul, car très peu d'exploitations sont concernées.

[3] Même si l'effet élémentaire pour le CH4 dans les bâtiments avec caillebotis est quasi nul, il semble préférable de l'intégrer dans l'analyse d'incertitude pour ne pas créer une distorsion au niveau des quelques exploitations concernées.

différence sur l'incertitude finale entre une analyse simplifiée et une analyse exhaustive et confirmer l'intérêt d'une analyse de Morris en amont.

4. Discussion et conclusion

Le but de cette première étude était de tester l'intérêt d'une analyse de sensibilité en amont d'une analyse d'incertitude. La démarche n'est pas complète, une analyse de Monte-Carlo doit confirmer que cela n'entraîne pas de biais dans le calcul de l'incertitude. Si les résultats apparaissent non biaisés, le fait d'appliquer une analyse de Morris préalablement à l'analyse d'incertitudes permettra d'alléger le travail d'expérimentation et de bibliographie pour décrire les lois des différents paramètres.
Il semblerait également pertinent de mener les analyses de sensibilité module par module (bâtiment, pâturage, stockage) pour avoir un degré d'analyse un peu plus fin. Il est également envisager d'explorer des situations particulières en réalisant l'analyse de Morris sur des sous-groupes d'élevages.

D'autres points restent encore à explorer, notamment la prise en compte des corrélations entre paramètres. Actuellement les paramètres sont raisonnés indépendamment les uns et des autres sans tenir compte des corrélations existantes. La corrélation entre paramètres peut être prise en compte dans les méthodes de ré-échantillonnage en utilisant une matrice de corrélation ou de covariance. (Huijbregts 2003), mais pour cela il est nécessaire d'avoir des données brutes caractérisant tous les paramètres afin de définir cette matrice de corrélation. On peut vraisemblablement supposer que l'incertitude de nos sorties est surestimée, l'espace d'échantillonnage créé étant plus important que si les corrélations avaient été prises en compte ((Bjoörklund, 2002).

Les auteurs remercient les éleveurs qui participent au dispositif des Réseaux d'Elevage, les ingénieurs départementaux qui assurent le suivi et l'enregistrement des données des exploitations et les ingénieurs qui animent les équipes régionales.
Ces travaux ont été réalisés dans le cadre du projet «Associer un niveau d'erreur aux prédictions des modèles mathématiques pour l'agronomie et l'élevage » (www.modelia.org) mené par l'ACTA et ses partenaires (2010-2012) et financé par le « Compte d'affectation spécial pour le développement agricole et rural » du Ministère de l'Agriculture et de la Pêche, FranceAgriMer, le Cniel, Interbev, et Interreg.

Bibliographie

Bojacà, C.R, Schrevens E. (2010). Parameter uncertainty in LCA : stochastic sampling under correlation. *Int. J. Life Cycle Asses.*, 15:238-246.

Charroin, T., Palazon, R., Madeline, Y., Guillaumin, A., Tchakerian, E. (2005). *Le système d'information des Réseaux d'Elevage français sur l'approche globale de l'exploitation. Intérêt et enjeux dans une perspective de prise en compte de la durabilité.* Renc. Rech. Ruminants, 12.

Gac, A., Manneville, V., Raison, C., Charroin, T., Ferrand, M. (2010). *L'empreinte carbone des élevages d'herbivores : présentation de la méthodologie d'évaluation appliquée à des élevages spécialisés lait et viande.* Renc. Rech. Ruminants, 17:335-342.

Huijbregts, M.A. (1998). *Application of uncertainty and variability in LCA.* The International Journal of Life Cycle Assessment, 3(5) 273-280.

Huijbregts, M.A., Gilijamse W., Ragas A.M., Reijnders L. (2003). Evaluating uncertainty in environmental life-cycle assessment. A case study comparing two insulation options for a Dutch one-family dwelling. *Environ Sci Technol* 37(11) 2600-8.

Makowski, D., Monod, H. (2011). *Analyse statistique des risques agroenvironnementaux*. France : Springer.

Payraudeau, S., van der Werf, H.M.G., Vertès. F., 2007. Analysis of the uncertainty associated with the estimation of nitrogenous emissions from a group of farms. *Agricultural Systems*. Volume 94, 2, 416-430.

Ravalico, J. K., Maier Holger, R., Dandy Graeme, C., Norton, J. P., Croke, B. F. W. (2005). *A comparison of sensitivity analysis techniques for complex models for environment managemen*t. International Congress on Modelling and Simulation : advances and applications for management and decision making (pp.2533-2539), Melbourne, Andre Zerger & Robert M. Argent (eds.).

Saltelli, A., Chan, K., Scott, E.M. (2000). Sensitivity Analysis. Hester: John Wiley.

Schaeffler, E. (2010) Evaluation environnementale selon une apporche cycle de vie des exploitations laitières françaises. Mémoire de fin d'études ENITA Bordeaux 102p.

Vigne, M., Peyraud, J.L., Lecomte, P. (2011*). Impact du choix des coefficients énergétiques sur les résultats de l'analyse énergétique : Exemple de la consommation énergétique des élevages bovins laitiers réunionnais.* Renc. Rech. Ruminants, 18:167.

Session 7 : Sensométrie II / *Sensometrics II*

Une nouvelle proposition, l'Analyse Discriminante Multitableaux : STATIS-LDA

A new proposal, Multiway Discriminant Analysis: STATIS-LDA

Robert Sabatier[1], Myrtille Vivien[1] & Christelle Reynès[1]

[1] *Laboratoire de Physique Industrielle et Traitement de l'Information - EA2415 - Université Montpellier 1 - France*
E-mail : *sabatier@univ-montp1.fr*

Résumé
L'analyse des multitableaux (ou multiblocs) peut être abordée à l'aide d'un certain nombre de méthodologies plus ou moins diffusées, selon les disciplines et les pratiques. Par contre, il existe très peu de méthodes généralisant la discrimination à des multiblocs. Nous allons proposer une nouvelle approche, et son algorithme associé, pour résoudre ce problème de discrimination (dans le cas, bien sûr, où les groupes sont identiques pour tous les tableaux) qui utilise l'Analyse Factorielle Discriminante usuelle (au sens de Fisher) et l'approche STATIS. Un exemple d'application sera proposé.

Mots-clés : Multitableaux, Analyse Factorielle Discriminante Linéaire (LDA), STATIS, Compromis.

Abstract
Multiblock tables analysis can be performed thanks to several methodologies, more or less wildely known according to disciplines and customs. Conversely, there exist very few methods allowingto generalize classification to multiblocks. We propose a new approach and its associated algorithm to solve this classification task (when groups are obviously the same ones in all tables) which uses usual linear discriminant analysis (Fisher's linear discriminant) and STATIS approach. A first application will be proposed.
Keywords : Multiway tables, Linear Discriminant Analysis (LDA), STATIS, compromize

1 Introduction

En chimiométrie, analyse sensorielle, écologie, analyse d'images, metabonomics, etc... les données fournies sont souvent organisées sous la forme de multitableaux (ou multiblocs) et leur analyse peut être abordée à l'aide d'un certain nombre de méthodologies plus ou moins diffusées, selon les disciplines et les pratiques : PARAFAC/CANDECOMP (Harshmann, 1970), TUCKER3 (Tucker, 1963) sont les plus importantes pour les cubes de données (tous ces articles ont été largement modifiés et améliorés depuis leur parution). Pour les multitableaux, on utilise plus spécifiquement : les méthodes Procustes Généralisées (Gower, 1975), les Analyses

Canoniques Généralisées (Carroll, 1968) ou encore la Consensus PCA (CPCA) de Westerhuis *et al.* (1988). La méthode STATIS (Lavit, 1988), et ses différentes modifications, s'applique indifféremment aux deux cas. Pour une revue bibliographique plus fouillée voir Vivien (2002). Par ailleurs, la discrimination est un problème très courant en analysis de données. Or, il existe très peu de méthodes généralisant la discrimination à des multiblocs. Les plus citées sont celles de Guimet *et al.* (2005) et Louwerse *et al.* (1999). Dans cette problématique, les groupes sont identiques, quelle que soit la sous-matrice. Bien évidemment il existe une solution *simple* qui consiste à réaliser l'Analyse Factorielle Discriminante du *super-tableau* où l'on juxtapose (déplie) le multitableau. Cette approche n'est pas adéquate car elle ne préserve justement pas la structure du multitableau, empêchant une interprétation adéquate du modèle obtenu. Nous allons proposer une nouvelle approche, pour résoudre ce problème de discrimination qui utilise conjointement l'Analyse Factorielle Discriminante usuelle et l'approche STATIS. Un exemple d'application sera fourni.

2 Brefs rappels sur LDA et STATIS

2.1 Notations

Soit un multitableau X composé de K matrices (ou sous matrices ou blocs) $[X_1, \ldots, X_K]$, chacune de dimension $n \times p_k$ ($k \in \{1, \ldots, K\}$), dont les p_k variables (potentiellement $p_i \neq p_j$ si $i \neq j$) sont mesurées sur les mêmes n observations. Les variables de toutes les matrices X_k, sont considérées comme centrées pour la métrique dite "du poids des observations", D. Cette métrique induit un produit scalaire entre deux variables x et y par $(x, y)_D = x^{'}Dy$. $W_kD = X_kQ_kX_k^{'}D$ est une matrice $n \times n$, appelée "opérateur des observations", est le produit scalaire entre les observations au sens de la métrique Q_k. Cet opérateur est l'analogue de $V_kQ_k = X_k^{'}DX_kQ_k$ qui est "l'opérateur des variables" de X_k. Si Q_k est l'identité, cet opérateur est égal à la matrice de variance-covariance entre les p_k variables du tableau X_k. L'ACP du triplet (X_k, Q_k, D) est équivalent à la diagonalisation des opérateurs W_kD et V_kQ_k.
De plus, nous allons supposer que les n observations sont agrégées en I groupes ($I > 1$), et on notera par U_I la matrice $n \times I$ des indicatrices (ou du codage disjonctif complet).

2.2 La méthode STATIS

Le but de la méthode STATIS est de calculer un opérateur de consensus, appelé le *compromis*, pour ensuite analyser cet opérateur par une ACP, et projeter les observations et les variables de chaque sous-matrice sur les premières composantes de cette ACP.
L'*interstructure*, première étape de STATIS, consiste à réaliser le calcul des produits scalaires entre les K opérateurs W_kD puis en fournir une représentation graphique dans un espace de petite dimension (deux ou trois). La matrice $C = \{c_{k,k'} = tr(W_kDW_{k'}D)\}$, de dimension $K \times K$, est définie comme la matrice des produits scalaires, au sens d'Hilbert-Schmidt, des opérateurs. Enfin, la diagonalisation de la matrice C, génère K vecteurs propres normés $\{l_\alpha\}_{\alpha=1\ldots K}$, chacun de longueur K, associés aux valeurs propres $\{\lambda_\alpha\}$. Les éléments de la matrice C sont non négatifs (car la trace d'un produit d'opérateurs est positive). Il en découle que le graphique réalisé avec $(\sqrt{\lambda_\alpha}l_\alpha, \sqrt{\lambda_\beta}l_\beta)$ donne une représentation euclidienne des K opérateurs.
L'étape suivante de STATIS est le calcul du *compromis*. Il s'agit de trouver un opérateur W_cD, de même dimension que les précédents, qui soit un *consensus* entre les opérateurs W_kD, au sens d'un certain critère. Cet opérateur est choisi comme une combinaision linéaire des K opérateurs

: $W_cD = \sum_{k=1}^{K} \nu_k W_k D$, où $\nu = (\nu_1, \nu_2 \dots \nu_K)^{'}$ est un vecteur de poids des K opérateurs. Le vecteur ν est choisi pour maximiser $\|W_cD\|^2 = tr(W_cDW_cD)$. La solution optimale pour ν avec $(\nu^{'}\nu = 1)$ est donnée par le vecteur propre, associé à λ_1, issu de la diagonalisation de la matrice C. Or, on peut choisir ce vecteur avec toutes ses coordonnées positives (par application du théorème de Perron-Frobenius), il en découle que W_cD est également semi-défini positif.
La dernière étape de STATIS est l'*intrastructure*, qui consiste à représenter, sur un même graphique, les n observations données par le compromis, avec celles fournies par les K opérateurs W_kD.

2.3 La méthode LDA

Considérons une variable qualitative avec I groupes (ou catégories ou classes), dont U_I est la matrice du codage disjonctif complet. On notera $D_I = U_I^{'}DU_I$ la matrice diagonale des fréquences relatives des I groupes ; $G_k = D_I^{-1}U_I^{'}DX_k$ la matrice $(I \times p_k)$ des centres de gravité et $V_{B_k} = G_k^{'}D_IG_k$ la matrice des variances inter-groupes pour le tableau k. On rappelle que la matrice des variances intra-groupes,V_{W_k}, est définie par : $V_{W_k} = V_k - V_{B_k}$.
L'Analyse Factorielle Discriminante linéaire, au sens de Fisher (notée LDA), de X_k par rapport à U_I est la recherche d'une combinaison linéaire des variables $c = X_k a_{LDA}$, de variance inter-groupe maximale, la variance intra-groupe étant égale à 1. C'est-à-dire, solution du problème : $Max\{\frac{a_{LDA}^{'}V_{B_k}a_{LDA}}{a_{LDA}^{'}V_{W_k}a_{LDA}}\}$. La solution est donnée par la diagonalisation de $V_{W_k}^{-1}V_{B_k}$ (voir Saporta, 2006). On peut montrer que la recherche des a_{LDA} (les axes de la LDA) est déduite des axes (notés a_{ACP}) de l'ACP de (G_k, V_k^{-1}, D_I), par l'équation : $a_{LDA} = \frac{1}{\mu_{ACP}}V_k^{-1}a_{ACP}$.

3 La nouvelle méthodologie : STATIS-LDA

La méthode proposée consiste à utiliser conjointement la méthode STATIS et le triplet de l'AFD et ce, pour chaque tableau de données.

Définition : On appelle STATIS-LDA, la méthode STATIS appliquée aux K triplets : $\{(G_k, V_k^{-1}, D_I)\}_{k\in\{1,\dots,K\}}$.

Propriétés (admises):

- L'élément courant de la matrice C est : $c_{k,k'} = tr(V_{B_{k'k}}V_k^{-1}V_{B_{kk'}}V_l^{-1})$, où $V_{B_{kk'}} = G_k^{'-1}D_IG_{k'}$.
- Le compromis W_cD est combinaison linéaire pondérée des opérateurs associés à la LDA de X_k par rapport à U_I. Ce compromis fournit les axes de STATIS-LDA.
- L'ACP du compromis (diagonalisation de W_cD) est équivalente à l'ACP du triplet (G, Q, D_I), avec $G = [G_1, G_2, \dots, G_K]$ et $Q = diag\left[\nu_1 V_1^{-1}{}_{,n}\, u_2 V_2^{-1}, \dots, \nu_K V_K^{-1}\right]$. C'est-à-dire que cette analyse n'est pas l'AFD de X par rapport à U_I (même si tous les ν_k sont égaux entre eux).
- Si $K = 1$, STATIS-LDA est la LDA ordinaire. De plus, si dans chaque triplet au lieu d'utiliser V_k^{-1}, on utilise la matrice identité d'ordre k, on retrouve STATIS sur les matrices inter-groupes.

Dans cette ACP, on note que les pondérations ν_k sont déduites des produits scalaires entre opérateurs des LDA partielles (c'est STATIS), on aurait pu choisir une quelconque autre

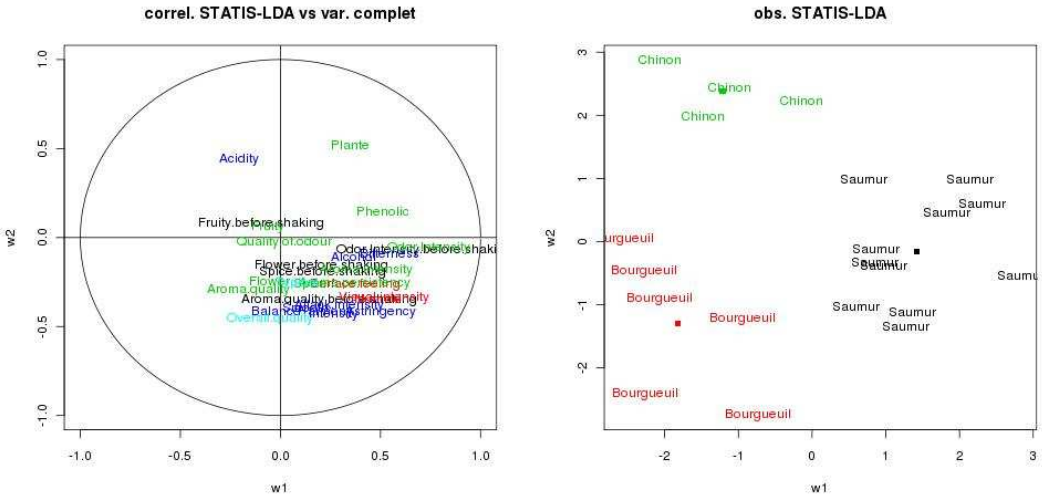


Figure 1: Représentation des variables et des observations (individus compromis) dans le STATIS-LDA appliqué aux données de vins.

pondération. On peut également, à la suite de cette analyse, réaliser des représentations graphiques et aides à l'interprétations, déduites aussi bien de la LDA que de STATIS : représenter les données et leurs centres de gravité, réaliser la validation croisée (leave-one-out, ou pas), représenter les groupes vus par le bloc k, faire de la prédiction,...

4 Application et comparaisons

L'exemple d'application est fourni par les données bien connues dites des *Vins de terroirs* de J. Pagès. Pour $n = 21$ vins (les observations), on a mesuré cinq groupes de variables ($K = 5$). Le premier comprend $p_1 = 5$ variables d'Olfaction avant agitation, le deuxième $p_2 = 3$ variables de Vision, le troisième $p_3 = 10$ variables d'Olfaction après agitation, le quatrième $p_4 = 9$ variables de Gustation, le cinquième $p_5 = 2$ les variables de jugement d'ensemble. Par ailleurs il y a trois groupes d'appellation : Saumur (11 vins), Chinon (4 vins) et Bourgueil (6 vins). Les 29 variables sont des moyennes de notes fournies par 36 juges (non utilisées ici) et sont centrées et reduites avec la pondération uniforme, et enfin la métrique Q_k de chaque bloc, est l'identité. Le but étant de mettre en évidence les principales dimensions de la variabilité sensorielle des vins, et de relier ces dimensions avec l'appellation.
L'application de STATIS-LDA pour les données, fournit comme coordonnées du vecteur ν : 1.40, 0.41, 1.21, 0.88 et 0.14. C'est-à-dire que dans la diagonalisation du compromis de STATIS-LDA, les variables Olfaction avant agitation et Olfaction après agitation sont les plus importantes, le groupe des variables jugement d'ensemble semble de peu d'intérêt. La Figure 1, donne la représentation globale de STATIS-LDA pour les variables et les observations, c'est-à-dire les individus compromis. On note bien, pour ces derniers, que les trois appellations semblent relativement bien séparées mais une validation croisée est nécessaire pour valider le pouvoir prédictif du modèle obtenu.

La Table 1, compare les pouvoirs discriminants (variance inter/variance totale) de STATIS-LDA par rapport à la LDA. On remarque le peu de perte pour STATIS-LDA par rapport à LDA usuel (sur le tableau déplié). Il faut bien noter que, en terme de discrimination simple, STATIS-LDA ne prétend pas mieux faire que la LDA sur le tableau déplié, mais son objectif est de permettre une interprétation plus poussée et surtout mieux adaptée à ce type de données (plusieurs tableaux de données). La validation croisée réalisée en formant aléatoirement 10 groupes d'observation, que l'on réalise 100 fois, fournit un moyen de comparer le pouvoir prédictif de STATIS-LDA et LDA (sur le tableau déplié). Les pourcentages de bien classés obtenus sont de 63.8% pour LDA, 52.4% pour STATIS-LDA. Cependant, si on considère les résultats partiels de STATIS-LDA pour chaque tableau, on obtient 88.6% pour le tableau X_1, 39.0% pour X_2, 46.7% pour X_3, 25.2% pour X_4 et 34.8% pour X_5. Ceci montre que les tableaux 2 à 5 ajoutent du bruit pour la discrimination et le tableau X_1 semble être le plus intéressant.

n°	STATIS-LDA	STATIS-LDA(X_1)	STATIS-LDA(X_5)	LDA	LDA(X_1)	LDA(X_5)
1	0.821	0.722	0.107	0.882	0.736	0.130
2	0.713	0.534	0.126	0.712	0.505	0.011

Table 1: Comparaisons numériques de STATIS-LDA avec LDA pour le pouvoir discriminant (variance inter/variance totale). Les chiffres donnés correspondent respectivement aux résultats obtenus pour le tableau complet, pour le premier tableau et pour le cinquième tableau.

5 Conclusion

Cette procédure, simple à réaliser et rapide, semble pertinente et répond à de nombreux problèmes pratiques, qui actuellement, sont sans solution opérationnelle. Cette méthode, outre le fait qu'elle peut fournir des résultats tableau par tableau, permet de mettre en évidence les tableaux importants ou inutiles pour la discrimination. Cette technique utilise les critères habituels de la discrimination (variance inter, variance intra), elle peut facilement se généraliser aux multi-tableaux à quatre entrées et peut également gérer la non-linéarité (B-splines).

Bibliographie

Carroll, J. D. (1968) A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th convention of the American Psychological Association*, n°3, 227-228.

Gower, J. C. (1975) Generalised procrutes analysis. *Psychometrika*, 40,33-51.

Guimet, F., Ferré, J. & Boqué R. (2005) Rapid detection of olive-pomace oil adulteration in extra virgin oils from the protected denomination of origin "'Siurana"' using excitation-emission fluorescence spectroscopy and three-way methods of analysis. *Analytica Chimica Acta*, 544, 143-152.

Harshmann, R. A. (1970) Foundation of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16,

1-84.

Lavit, Ch. (1988) *Analyses conjointe de tableaux quantitatifs.* Masson, Paris, 252p.

Louwerse, D.J., Tates, A.A., Smilde, A.K., Koot, G.L.M. & Berndt, H. (1999) PLS discriminant analysis with contribution plots to determine differences between parallel batch reactors in the process industry. *Chemometrics and Intelligent Laboratory System*, 46, 197-206.

Pagès, J. Analyse Factorielle Multiple, document Agro-campus, Rennes.

Saporta, G. (2006)*Probabilités Analyse des Données et Statistique.* Technip, Paris.

Tucker, L. R. (1963) Implications of factor analysis of three-way matrices for measurement of change. In *Problems in measuring change*, 122-137, Madison University of Winconsin Press.

Vivien, M. (2002) *Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications.* Thèse de l'Université Montpellier I.

Westerhuis, J. A., Kourti, T. & MacGregor, J.F. (1998) Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12, 301-321.

Utilisation de techniques multi-tableaux pour l'étude des relations entre les caractéristiques physico-chimiques et les caractéristiques sensorielles de vins rouges

Using multi-table analysis to study relationships between physico-chemical and sensory characteristics of red wines

Soline Caillé[1], Guillaume Dedieu, Alain Samson, Cécile Morel-Salmi, Pascale Williams, Thierry Doco, Véronique Cheynier & Gérard Mazerolles

[1] *UMR 1083 Sciences pour l'œnologie, Institut National de la Recherche Agronomique, Centre de Recherche de Montpellier, 2 place Viala, 3460 Montpellier cedex1, France*
E-mail : *soline.caille@supagro.inra.fr*

Résumé
L'objectif de cette étude est de mesurer l'impact de techniques de vinification sur la composition en polyphénols et en polysaccharides de vins rouges de Carignan et les modifications sensorielles et de couleur qui en découlent. Les vins ont été élaborés en comparant quatre techniques de vinification : vinification classique, vinification classique avec ajout d'enzymes, flash détente et flash détente avec ajout d'enzymes. Les variables mesurées sur les vins sont organisées en cinq tableaux : composition en polyphénols, en tanins et en polysaccharides et caractéristiques sensorielles et de couleur. Des analyses multi-variées (ANOVA - Simultaneous Component Analysis ou ASCA) sont réalisées pour étudier l'impact des facteurs (flash détente et enzymes). Puis, une Analyse de Co-inertie Multiple (ACoM) permet d'établir les liens entre les variables des tableaux. Ces analyses statistiques sont performantes pour ce type de problématique et permettent une interprétation aisée. L'ACoM a permis une synthèse de toute l'information interprétable.
Mots-clés : vin rouge, flash détente, enzymage, ANOVA-SCA, analyse de co-inertie multiple

Abstract
The impact of winemaking techniques on polyphenolic and polysaccharide composition and on sensory characteristics and color of Carignan red wines has been investigated. Four winemaking techniques were compared: control, control enzyme treated, flash détente, flash detente enzyme treated. Variables measured on wines were organized into five tables: polyphenols, tanins and polysaccharides composition, color and sensory characteristics. Multivariate analyses (ANOVA - Simultaneous Component Analysis ou ASCA) are performed to study impact factors (flash detente and enzymes) on wine composition and characteristics. Then, a multiple co-inertia analysis (MCoA) established the relationships between variables on the five tables. These statistical analysis are impressive for this problematic and are easy to interpret. MCoA permit a synthesis of all interpretable information.
Keywords : red wine, flash release, enzyme, ANOVA-ASCA, multiple co-inertia analysis

1. Introduction

Le programme de recherche Européen MAXFUN avait pour objectif le développement de nouvelles technologies de transformation pour les industries de transformation des fruits. Son objectif principal était de maximiser la qualité des produits, en se souciant particulièrement de leur impact sur la santé humaine en relation avec leur composition polyphénolique.
Pour la filière vitivinicole, ce programme a porté sur les effets de l'utilisation de la flash détente et/ou de l'addition d'enzymes, ces procédés de vinification permettant un enrichissement du vin en polyphénols. Des vins ont été élaborés en fonction de ces deux facteurs, puis des mesures de composition en polyphénols, en polysaccharides et des mesures de couleur ont été réalisées. Le traitement par analyse univariée des résultats obtenus pour chaque famille de composés a montré l'impact de la flash détente sur la composition en polyphénols (Morel-Salmi et al, 2006) et de l'effet de la flash détente et des enzymes sur la composition en polysaccharides (Doco et al, 2007). Parallèlement, une analyse sensorielle descriptive des vins a été réalisée.
L'objectif de cette étude est de réunir l'ensemble des données dans une même analyse statistique afin de pouvoir évaluer l'impact des procédés utilisés sur les modifications de la composition en polyphénols et en polysaccharides, et les caractéristiques sensorielles et de couleur des vins associées.
Deux analyses statistiques sont particulièrement adaptées à ce type d'étude : l'ANOVA-SCA permettant d'analyser un plan d'expérience en situation multivariée et l'Analyse de Co-inertie Multiple (ACoM) permettant de relier toutes les variables mesurées.

2. Matériel et Méthodes

2.1 Les vins

Les vins sont issus du cépage Carignan, vinifiés en 2004 à l'Unité Expérimentale de Pech Rouge (INRA).
Après foulage et égrappage, la vendange a été séparée en deux lots de 600 kg. Le premier a été distribué dans six cuves de 1hL et le second traité par flash détente puis répartie dans six autres cuves de 1hL. Trois cuves de chaque série ont été ensuite traitées par addition d'enzymes pectolytique.

L'espace produit est donc composé de 4 modalités, chacune ayant été répétée 3 fois (Tableau 1).

Code échantillon	Mode de vinification	Enzymage
CC1 - CC2 - CC3	Témoin	Non
CE1 - CE2 - CE3	Témoin	Oui
FC1 - FC2 - FC3	Flash Détente	Non
FE1 - FE2 - FE3	Flash Détente	Oui

Tableau 1: Plan d'expérience

2.2 Recueil des données

2.2.1 Données de composition en polyphénols et paramètres de couleur

L'analyse quantitative des polyphénols (anthocyanes, acides phénols, flavonols, catéchines et anthocyanes dérivées) a été réalisée par HPLC-DAD. (Morel-Salmi et al, 2006)

Après évaporation à sec du vin, les tanins ont été extraits par précipitation avec du méthanol (Preys et al, 2006). Une caractérisation qualitative de ces composés a ensuite été réalisée par thiolyse suivie par une analyse HPLC (Souquet et al, 1996).
Les mesures de couleur ont été effectuées par spectromètrie UV-visible (Atanasova et al, 2002). Les valeurs des absorbances à 420 (pur 420), à 520 (pur 520) et à 620 (pur 620) ont été mesurées dans une cellule de 1mm, et converties pour un trajet optique de 10 mm. La teinte (T) a été calculée par le rapport Pur420/Pur520, et l'intensité de la couleur (IC) par la somme Pur420, Pur520 et Pur620. L'indice des polyphénols totaux (IPT) a été défini par une absorbance de 280nm dans du vin dilué avec une solution d'HCL à 2%. Une absorbance a également été mesurée à 520 nm (Dilué HCL 520). La couleur, due aux dérivés résistants à la décoloration par les sulfites, a été déterminée à 520 nm après addition d'une solution de métabisulfite (CDR SO2).
L'ensemble de ces variables, regroupées par familles, est présenté dans le tableau 2.

Composés phénoliques (en mg/L) (27 variables)		Tanins (4 variables)	Couleur (10 variables)
Anthocyanes :	***Flavonols***	Tanins (mg/L)	Pur 420
Delphinidine 3 glucoside	Myricétine 3 glucoside	Degré moyen de polymérisation (mDP)	Pur 520
Pétunidine 3 glucoside	Quercétine 3 glucoside	% d'unités galloylées (%Gall)	Pur 620
Péonidine 3 glucoside	Myricétine	% d'unités épigallocatechine (%EGC)	IC
Malvidine 3 glucoside	Quercétine		T
Delphinidine 3 acétylglucoside	***Catéchines***		Dilué HCL 280 - IPT
Pétunidine 3 acétylglucoside	Catéchine		Dilué HCL 520
Péonidine 3 acétylglucoside	Epicatéchine		Couleur due aux dérivés SO2 (CDRSO2 ou SO2 520)
Malvidine 3 acétylglucoside	***Anthocyanes dérivées***		Anthocyanes libres totaux (At)
Delphinidine 3 p-coumaroylglucide	(epi)catechine-peonidine 3-glucoside (F-Péo)		
Péonidine 3 p-coumaroylglucide	(epi)catechine-malvidine 3-glucoside (F-Mv)		
Malvidine 3 p-coumaroylglucide	carboxypyranomalvidine 3-glucoside (Mv-pyruvique)		
Acides hydroxycinnamiques	phenylpyrano malvidine 3-glucoside (Mv-vinyl)		
Acide trans-caftarique	phenylpyrano malvidine 3-coumaroylglucoside (Mvcoum-vinyl)		
Acide cis-coutarique			
Acide trans-coutarique			
Acide caféique			
Acide para-coumarique			

Tableau 2: Variables de composition en polyphénols et paramètres de couleur analysées

2.2.2 Données de composition en polysaccharides

Les polysaccharides constituent un des principaux groupes de macromolécules des vins. Ils sont regroupés en trois grandes familles : Les PRAG ou polysaccharides Riches en Arabinose et en Galactose, les Rhamnogalacturonanes (majoritairement RGII) qui ont pour origine les parois de la baie de raisin et les Mannoproteines (MPs) qui sont libérées par les levures au cours de la fermentation alcoolique (Vidal et al., 2003). La quantité en monosaccharides constitutifs des polysaccharides, déterminés par chromatographie en phase gazeuse après hydrolyse, réduction puis acétylation (Doco 1999), permet de calculer les concentrations en MPs, en PRAGs et en RG-II présents dans chaque vin.

Mannoprotéines (MP)
Polysaccharides riches en arabinoses et galactoses (PRAG)
Rhamnogalacturonane de type II (RGII)

Tableau 3: Variables de composition en polysaccharides analysées en g/L (3 variables)

2.2.3 Données Sensorielles

Une analyse descriptive quantitative a été réalisée par un jury expert (20 juges), sélectionné sur ses aptitudes sensorielles et entraîné à la description des vins de l'étude (AFNOR ISO 8586-2). Les juges sont entraînés à identifier et évaluer l'intensité des descripteurs par l'utilisation de standards. L'intensité des descripteurs (liste tableau 4) est évaluée sur une échelle linéaire continue de « faible » à « fort ». Les vins sont présentés de façon monadique, selon un ordre basé sur un carré latin de Williams (Macfie 1989) et dans des verres standardisés INAO. Une répétition de l'analyse des vins a été réalisée. Les données sont recueillies avec un système d'acquisition de données informatisé (FIZZ software, Biosystemes, Couternon France).

Descripteurs olfactifs	Empyreumatique - Poivré - Bonbon - Animal - Moisi - Alcool - Cassis - Boisé
Descripteurs gustatifs	Acide - Amer - Chaud - Astringent - Piquant - Sucré - Empyreumatique

Tableau 4: Descripteurs sensoriels analysés (15 variables)

2.3 Traitement des résultats

Les analyses destinées à déterminer l'influence de chaque facteur de vinification sur chaque famille de variables sont réalisées par ANOVA-Simultaneous Component Analysis (ASCA) (Smilde 2005, Jansen 2005). Cette technique propose de réaliser, pour chaque facteur du plan d'expériences l'ACP des moyennes des réponses obtenues pour ses différentes modalités. Dans la forme mise à disposition par les auteurs auprès des utilisateurs, elle est très proche de l'ACP-VI (Sabatier 1989) qui pourrait aussi être utilisée dans ce cadre. Le traitement simultané des cinq tableaux de données (sensoriel, polysaccharides, couleur, polyphénols et tanins) a été réalisé par Analyse de Co-inertie Multiple (ACoM) (Chessel 1996). L'objectif de cette technique est de décrire simultanément plusieurs tableaux de données en restituant le maximum de la variabilité présenté par chacun d'entre eux. Elle suppose l'existence d'une structure commune aux différents tableaux étudiés. L'analyse de co-inertie est particulièrement bien adaptée aux situations où le nombre d'individus est faible par rapport aux nombres de variables.

3. Résultats et Discussion

3.1 ASCA sur les données de composition phénolique

L'axe 1 de l'ASCA sur le tableau de données de polyphénols explique 93,52% de la variance et permet de différencier les modalités FD des modalités Témoin. .

Les modalités Flash Détente ont des concentrations plus importantes en acide trans caftarique et en acide trans coutarique ; tandis que les modalités Témoin ont des valeurs plus élevées en acide para-coumarique et en acide caféique (figure 1). C'est donc la composition en acides hydroxycinnamiques qui est la plus modifiée en fonction du type de vinification.

On observe également des teneurs en anthocyanes (delphinidine 3 glucoside, pétunidine 3 glucoside et péeonidine 3 glucoside) supérieures pour les modalités de Flash Détente.

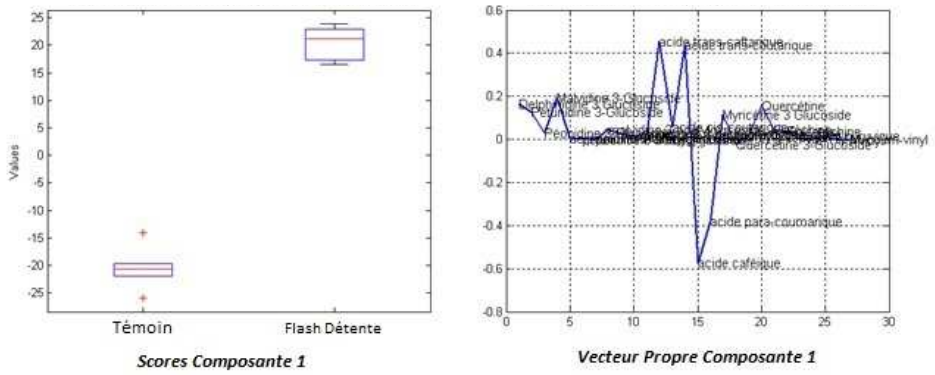


Figure 1: ASCA sur tableau de polyphénols – effet flash détente

L'ajout d'enzymes n'a pas un effet important sur les concentrations en composés phénoliques de ces vins (1,10% de la variance).

3.2 ASCA les données de Tanins

L'effet de la flash détente explique 62,33% de la variance du tableau de données de tanins. Ces derniers permettent de séparer très distinctement les modalités FD qui ont une concentration en tanins et un pourcentage d'unités galloylées plus élevés. Les modalités Témoin sont caractérisées par un pourcentage d'unité épigallocatechine plus important (figure 2).

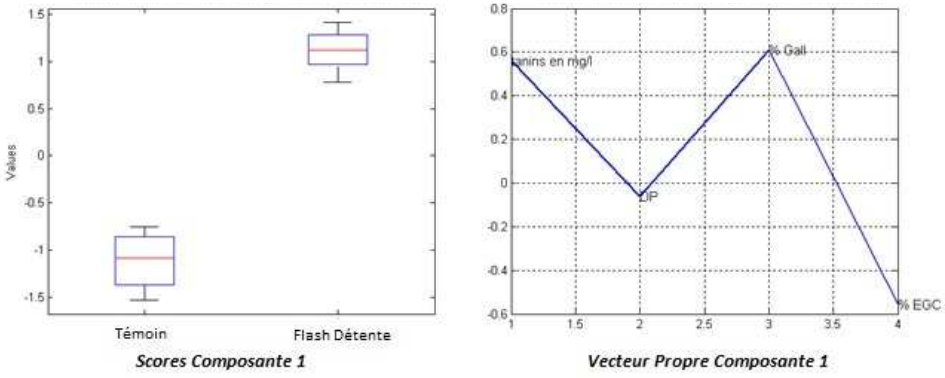


Figure 2: ASCA sur tableau de tannins – effet flash détente

On observe une absence d'effet de l'enzymage sur les variables de tanins.

3.3 ASCA sur les données de Couleur

Comme pour les données de polyphénols, les données de couleur permettent de séparer les modalités de Flash Détente des modalités Témoin, ces derniers présentant une teinte plus importante, donc une couleur plus orangée (79,05% de la variance totale) (figure 3).

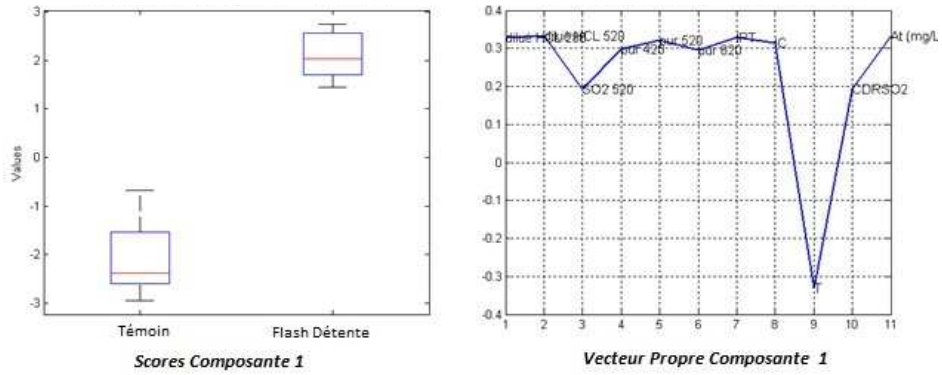


Figure 3: ASCA sur tableau de couleur – effet flash détente

L'ajout d'enzymes a peu d'effet sur les variables de couleur (6,78% de la variance) et serait lié essentiellement à un échantillon atypique.

3.4 ASCA sur les données de composition en polysaccharides

L'axe 1 de l'ASCA sur le tableau de données de polysaccharides permet de différencier les modalités enzymées des modalités non-enzymées (72,08% de la variance totale). L'ajout d'enzyme conduit à des concentrations en RGII plus importantes, et au contraire à une diminution de la concentration en PRAG (figure 4).

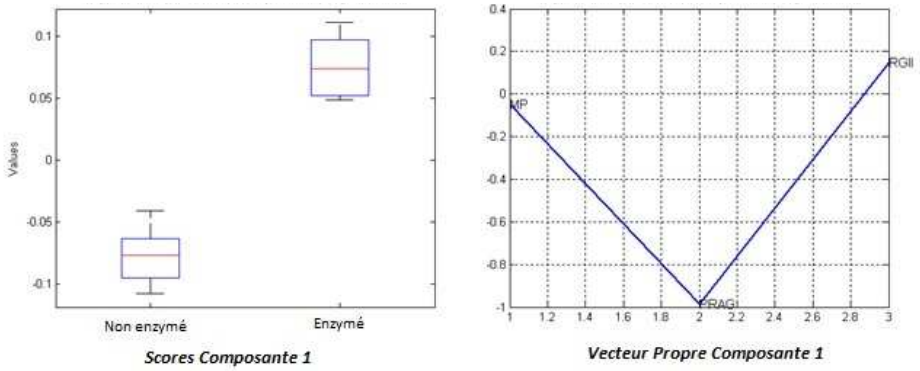


Figure 4: ASCA sur tableau de Polysaccharides – effet enzyme

L'effet de la flash détente est moins marqué que celui de l'ajout d'enzyme (13,51% de la variance totale). Les vins obtenus par flash détente sont toutefois caractérisés par une concentration en PRAG plus élevée et au contraire une teneur en mannoprotéines (MP) plus faible (figure 5).

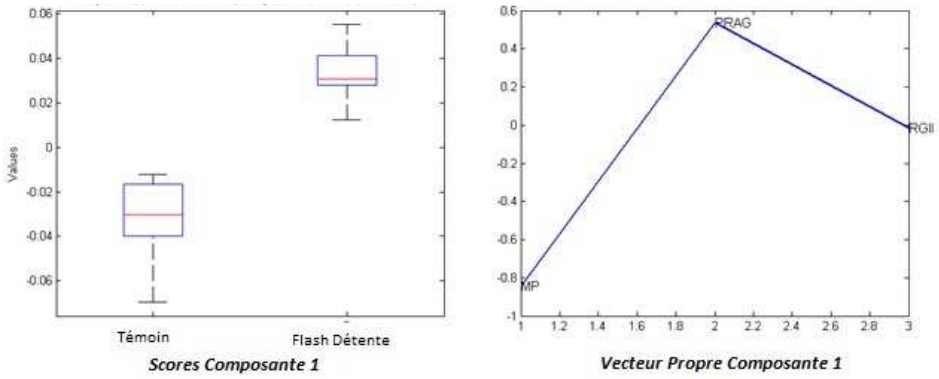


Figure 5: ASCA sur tableau de polysaccharides – effet flash détente

3.5 ASCA sur les données Sensorielles

On observe que les caractéristiques sensorielles permettent de différencier essentiellement les modalités selon leur vinification (42,82% de la variance), puis dans une moindre mesure selon l'enzymage (7,80% de la variance).

Les modalités de Flash Détente sont caractérisées par des odeurs de bonbon, de cassis et d'alcool plus importantes, ainsi qu'un peu plus d'acidité et d'astringence en gustatif.

Les modalités Témoin présentent des intensités supérieures pour les caractères « animal et moisi » en olfactif et « sucré et chaud » en gustatif.

On constate qu'un échantillon de Témoin (CE3) est très proche sensoriellement des modalités de Flash Détente (figure 6).

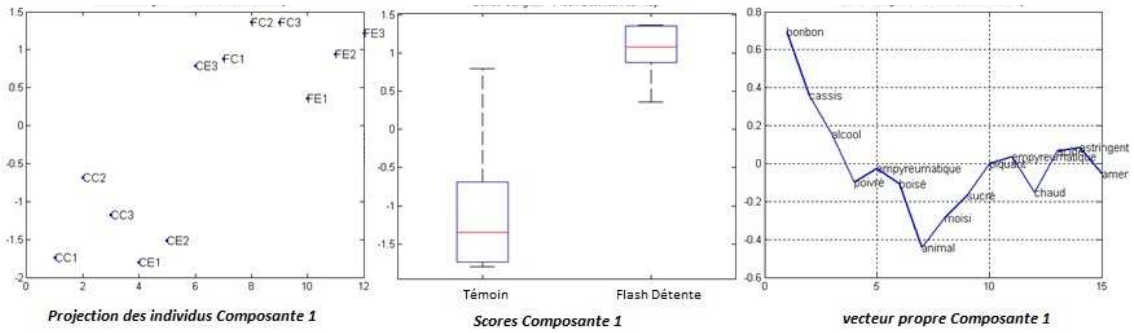


Figure 6: ASCA sur tableau sensoriel – effet flash détente

L'effet d'ajout d'enzymes est moins marqué ; cependant, les vins enzymés sont perçus avec des odeurs « animal et moisi » plus importantes, et un peu plus « amer et astringent ». (Figure 7)

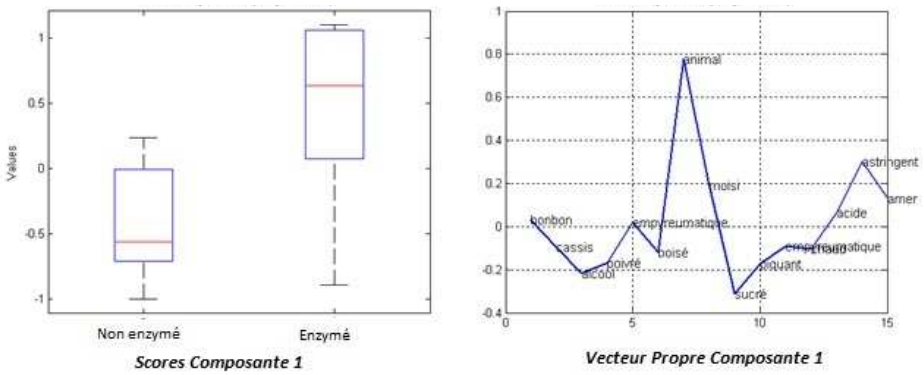


Figure 7: ASCA sur tableau sensoriel – effet enzyme

3.6 Analyse Multi-tableaux

L'objectif de l'Analyse de Co-inertie Multiple (ACoM) est d'établir des liens entre les 5 tableaux et de savoir comment se positionnent les individus par rapport à ces liens.

Tableau	CC1	CC2
Sensoriel	0.82	0.13
Couleur	0.90	0.07
Polysaccharide	0.18	0.55
Composés phénoliques	0.97	0.00
Tanin	0.80	0.08

Tableau 5: Pourcentage de variance expliqué pour les axes 1-2 du compromis de l'ACoM

La figure 8 présente la disposition des vins dans le plan 1-2, chaque point compromis est représenté par le barycentre de la contribution pour chaque tableau.
La dimension 1 permet de visualiser l'effet de la vinification, en opposant les vins issus de flash détente aux autres. Par les contributions, il apparaît que les tableaux de composés phénoliques, couleur, sensoriel et tanins sont fortement associés à la dimension 1 du compromis.
La dimension 2, quant à elle, représente l'effet des enzymes, associé au tableau de polysaccharides.

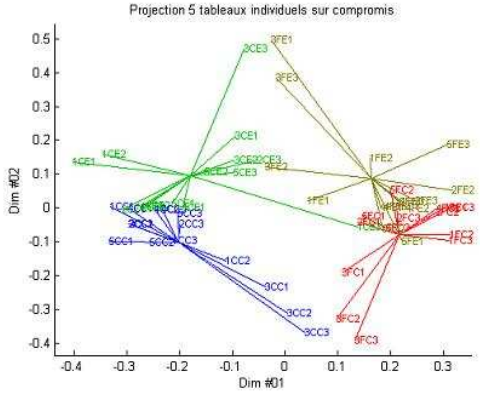


Figure 8: ACoM – représentation des individus sur le compromis 1-2

La visualisation des variables sur l'axe 1 pour les tableaux sensoriel, composés phénoliques, couleur et tanin permet d'établir un certain nombre de corrélations intra et inter tableaux entre les variables initiales. (Figure 9) Parmi ces corrélations, deux types de relations peuvent être interprétées sur la base de connaissances déjà acquises:

- la relation entre utilisation Flash Détente - concentration en anthocyanes natives et teinte plus faible, (couleur plus rouge des vins) d'une part et utilisation Flash Détente -concentration en tanins pourcentage d'unités galloylées - perception d'astringence un peu plus importante d'autre part. On s'adresse ici à un lien direct entre composés phénoliques et caractéristiques sensorielles.

- la relation entre les perceptions « animal et moisi » des modalités Témoin avec la présence d'acide caféique et d'acide paracoumarique, ces deux composés étant les précurseurs de molécules connues pour être les responsables de ces défauts. On peut envisager que l'opération de chauffage employé lors de la flash détente inhibe l'enzyme nécessaire à la synthèse de ces précurseurs, les vins obtenus étant alors exempts de ces défauts.

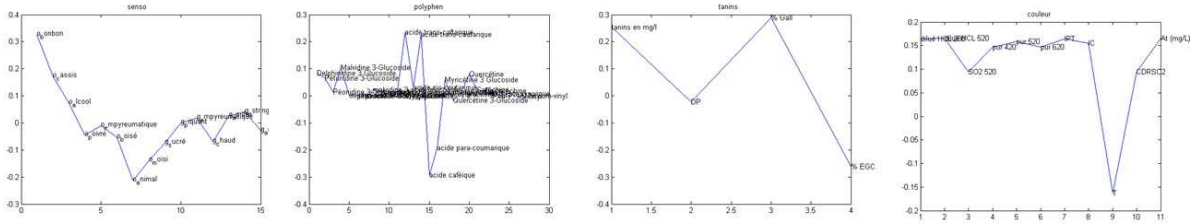


Figure 9: ACoM – vecteurs propres sur la composante 1 des tableaux sensoriel, couleur, polyphénols et tanins.

La dimension 2 du compromis concerne essentiellement les concentrations en polysaccharides qui sont déterminées par l'enzymage. Ce tableau semble toutefois présenter un lien ténu avec les données sensorielles qui devra être validé par des études ultérieures.

4. Conclusion

L'ASCA est une méthode statistique performante qui nous a permis d'analyser les deux facteurs flash détente et enzymage sur chacun des tableaux de variables, avec une interprétation aisée.

L'ACoM est, quant à elle, une analyse statistique adaptée à ce type de problématique, puisqu'en une seule analyse elle synthétise toute l'information interprétable. L'interprétation de l'ACoM, nous a également permis d'analyser les relations entre les variables : la concentration en anthocyanes induit une couleur plus rouge aux vins et une plus forte teneur en tanins accroît la perception d'astringence, ces deux conclusions étant liées au facteur flash détente. De plus, les données de polysaccharides permettent d'expliquer l'effet des enzymes qui semble légèrement corrélé à la perception de l'astringence et de l'amertume des vins.

Remerciements : Ce travail a été réalisé avec le soutien financier de la Commission de l'Union Européenne (proposition QLRT-2002-02364, acronyme MAXFUN) ; Il ne reflète pas ses vues, et en aucune manière n'anticipe la future politique de la Commission dans ce domaine.
Nous remercions Magali Bes (UE Pech Rouge) pour le suivi de la vinification des vins.

Bibliographie

Watt, D. K., Brasch, D. J., Larsen, D. S., & Melton, L. D. (1999). Isolation, characterisation, and NMR study of xyloglucan from enzymatically depectinised and non-depectinised apple pomace. *Carbohydrate Polymers*, 39(2), 165-180.

Morel-Salmi, C., Souquet, J.M., Bes, M., & Cheynier, V. (2006). Effet of Flash Release Treatment on Phenolic Extraction and Wine Composition. *J. Agric. Food Chem*. 54, 4270-4276.

Preys, S., Mazerolles, G., Courcoux, P., Samson, A., Fischer, U., Hanafi, M., Bertrand, D. & Cheynier, V. (2006) Sensory properties in red wines using multiway analysis. *Anal. Chim. Acta*, 563, 126-136.

Doco, T., Williams, P. & Cheynier, V. (2007). Effet of Flash Release and Pectinolytic Enzyme Treatments on Wine Polysaccharide Composition. *J. Agric. Food Chem*. 55, 6643-6649.

Vidal, S., Williams, P., Doco, T., Moutounet, M. & Pellerin, P. (2003) The polysaccharides of red wines: Total fractionation and characterization. *Carbohydr. Polymers*, 54, 439-447.

Doco, T., Quellec, N. & Moutounet, M. (1999) Polysaccharide patterns during the aging of Carignan noir red wines. *Am. J. Enol. Vitic*., 50, 25-32.

Souquet, J.M., Cheynier, V., Brossaud, F. & Moutounet M. (1996) Polymeric proanthocyanidins from grape skins. *Phytochemistry*. 43, 509-512.

Atanasova, V., Fulcrand, H., Cheynier, V. & Moutounet, M. (2002) Effect of oxygenation on polyphenol changes occurring in the course of wine-making. *Anal. Chim. Acta*, 458, 15-27.

Somers, T.C. (1971) The polymeric nature of wine pigments. *Phytochemistry*, 10, 2175-2186.

ISO 8586-2. (1994) Analyse Sensorielle – Guide général pour la sélection, l'entraînement et le contrôle des sujets. Partie 2 experts. *International Organisation for standardization*

MAcfie, H.J., & Bratchell, J. (1989) Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *J. Sensory studies*, 4, 129-148

Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamer,s R.J.A.N., Van der Greef, J. & Timmerman, M.E. (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analysing designed metabolomics data. *Bioinformatics*, 21, 3043.

Jansen, J.J., Hoefsloot, H.C.J., Van der Greef, J., Timmerman, M.E., Westerhuis, J.A. & Smilde, A.K. (2005) ASCA: analysis of multivariate data obtained from experimental design. *J. of Chemometrics*, 19, 469.

Sabatier, R., Lebreton, J.D., Chessel, D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data, in: Coppi and S. Bolasco (Eds.). Multiway data analysis; *Elsevier Science Publishers B.V.*, North Holland, 341.

Chessel, D. & Hanafi M. (1996) Analyses de la co-inertie de K nuages de points, *Rev. Statistique Appliquée*, XLIV-2, 35.

Validation des profils idéaux obtenus directement de consommateurs

Validation of the ideal profiles provided directly from consumers

Thierry Worch[1,2], Sébastien Lê[2], Pieter Punter[1] & Jérôme Pagès[2]

1 *OP&P Product Research, Utrecht, the Netherlands*
E-mail : *thierry@opp.nl*

2 *Laboratoire de Mathématiques Appliquées, Agrocampus-Ouest, Rennes, France*

Abstract:

The Ideal Profile Method is a sensory method in which, for each product tested, consumers are asked to rate both the perceived and ideal intensities of a list of attributes. In addition, they are also required to indicate how much they like each product. At the end of the task, three blocks of data are collected from each consumer: the product profiles, their ideal profile and the liking ratings.
The ideal profiles can be used to help improving the existing products. However, this information should be carefully managed since *(1)* it is obtained from consumers, and *(2)* it describes a virtual product. In order to use the full potential of the ideal profiles, and to avoid a possible misinterpretation of the data, one has to ensure that the information collected is valid.
The validation process proposed here is based on the liking ratings: an ideal product should achieve higher hedonic ratings than the tested products, if it would be tested. But since the liking scores of the ideal products are unknown, they are estimated. However, the comparison between liking scores (estimated for the ideals, measured for the tested products) would only make sense if the ideal intensities have not been randomly rated. For that matter, a hypothesis test checking for the significance of the ideal profiles is defined.

Keywords:

Consumer, Ideal Profile Method, liking, validation, permutation test

1. Introduction

The *Ideal Profile Method* (*IPM*) is a method which aims at acquiring sensory data from consumers. During this test, products are presented to consumers who are asked to rate both the perceived and ideal intensities on a set of pre-defined attributes. During the task, the same consumers are also asked to provide hedonic ratings of the products. In this sense, the *IPM* can be seen as a combination of *QDA*® (profiling products, ***Stone, Sidel, Oliver, Woosley & Singleton, 1974***) and *JAR* scaling (providing ideal profiles).

The application potential of the data provided from the *IPM* is large as three types of information are obtained from each consumer: the sensory profiles of the products (*i.e. how consumers perceive the products*), the hedonic scores (*i.e. how much consumers like the products*) as well as the ideal profiles (*i.e. what are the consumers' expectations*) (***Van Trijp, Punter, Mickartz & Kruithof***, ***2007***).
Gathering this diverse information, and more specifically the ideal profiles, is crucial as it can help manufacturers to improve existing products (***Worch, Dooley, Meullenet & Punter***, ***2010***). However, this information (the ideal descriptions) is delicate as *(1)* it comes directly from consumers and *(2)* it describes virtual products.

In order to use the full potential of the ideal descriptions, and to avoid any possible misinterpretation which could lead to an incorrect reformulation of the tested products, one has to be sure that the information collected is relevant. Hence, it seems important for the authors to validate the ideal data before use. The validation procedure proposed here is performed in two steps.

The first part consists in studying the relationships between the different types of data by checking for the ***consistency*** of the ideal. In this case, the consistency of the ideal is defined according to both the sensory and hedonic data: the sensory profile of an ideal product is considered consistent if it matches the same sensory characteristics as the profile of the most liked product. As the methodology checking for the consistency of the ideal has already been proposed by ***Worch, Lê, Punter & Pagès*** (***2012***), it is not developed here.

The second part consists in checking for the ***validity*** of the ideal ratings provided by the consumers. This validation process is based on liking: in theory, the ideal ratings should correspond to a product that is more liked than the tested products, if it happens to exist. As the liking of the ideal profile (also called *liking potential of the ideal product*) is unknown, it has to be estimated. For that reason, a model explaining the liking scores in function of the way products are perceived is constructed for each consumer. The model is then applied to the ideal ratings, and the liking potential of the ideal product is estimated and compared to the liking scores given to the products. The ideal is considered *valid* if the estimation of its liking potential is superior than the liking scores given to the products.

Although the principle of the validation procedure is rather simple, it is based on a strong hypothesis: consumers do not rate their ideal in a *random* way. For that reason, a hypothesis test checking for the significance of the estimated liking potential of the ideal products is also set up.

2. Material and methods

2.1 Material

In order to illustrate the methodology presented below, the dataset presented by ***Worch, Lê & Punter*** (***2010***) is used. It concerns 12 luxurious women perfumes among which two were duplicated (***Table 1***). Each product has been rated on 21 attributes (also listed ***Table 1***) by 103 Dutch consumers. For each perfume and each attribute, both the perceived and ideal intensities have been rated on 100mm line scales. After rating each product on perceived and ideal intensity, overall liking was rated on a structured 9-point category scale.

Products	Type	Attributes	
Angel	Eau de Parfum	Intensity	Spicy
Cinema	Eau de Parfum	Freshness	Woody
Pleasures	Eau de Parfum	Jasmine	Leather
Aromatics Elixir	Eau de Parfum	Rose	Nutty
Lolita Lempicka	Eau de Parfum	Chamomile	Musk
Chanel $N^0 5$	Eau de Parfum	Fresh lemon	Animal
L'Instant	Eau de Parfum	Vanilla	Earthy
J'Adore (EP)	Eau de Parfum	Citrus	Incense
J'Adore (ET)	Eau de Toilette	Anis	Green
Pure Poison	Eau de Parfum	Sweet fruit	
Shalimar	Eau de Toilette	Honey	
Coco Mademoiselle	Eau de Parfum	Caramel	

***Table 1*:** List of products and attributes.
Note: the products Pure Poison and Shalimar have been duplicated

The samples were presented in monadic sequence taking care of order and carry-over effects (***MacFie, Bratchell, Greenhoff & Vallis***, ***1989***) in two 1-hour sessions.

2.2 Method

In the *ideal* situation, when consumers rate their ideal profiles, they describe *fictive* products which they should appreciate more than the products tested, if they would physically exist. In other words, the ideal products provided by consumers should have a higher liking potential than the products themselves. This property has to be checked at the individual level, i.e. for each consumer separately. Unfortunately, the liking potentials of the individual ideal products are unknown, as they cannot be measured directly. So they have to be estimated.

To do so, individual models expressing the consumers' appreciations h_{jp} in function of their perception of the products y_{jpa} are constructed. These individual models are then applied to the averaged ideal ratings $\bar{z}_{j.a}$ the consumers provided, and the liking potentials of the ideal products $\hat{h}_{j.} | \bar{\mathbf{z}}_{j..}$ are estimated (summary ***Figure 1***).

Figure 1: Procedure used to estimate the liking potential of the averaged ideal profile for consumer *j*

In this procedure, the ideal profile described by each consumer is *valid* if the estimated liking potential of the ideal product is higher than the likings scores given to the actual products by the consumer considered. But this comparison (estimated liking potential of the ideal product *vs.* liking scores given to the products) is only meaningful if we are sure that the ideal description provided by a consumer is not given randomly. Hence the significance of the estimated liking potential of the ideal product has to be tested first.

2.2.1 Significance test of the liking potential of the ideal profiles

Testing the significance of the estimated liking potential of the ideal profile is less straightforward than the validation procedure of the ideal profiles presented above.

A *valid* ideal product is defined as being consistent with the other descriptions (sensory and hedonic) provided and is associated with a high liking potential. A *non-valid* ideal product can reflect a lack of consistency in the ideal description, which can be explained by a lack of agreement between the ideal profiles and the hedonic scores provided. In that case the estimation of the liking potential of the ideal product would be relatively low indicating that it could be obtained randomly.

In that sense, testing the significance of the ideal product is done by comparing the estimated liking potential ($\hat{h}_j|\bar{z}_{j..}$) obtained in the real situation (denoted as the *real* estimated liking potential) to estimated liking potentials (noted $\hat{h}^{'}_j|\bar{z}_{j..}$) obtained in random situations. As the difference between the real and the random situations is the consistency of the ideal data with the other information, one can expect that in the real situation, the ideal description fits the individual models constructed and hence is associated to a large estimated liking potential while in the random situations, this would not be the case.

The estimated liking potential is expected to be larger in the real than in the random situations: the test performed is a one-tailed test. The null and alternative hypotheses are defined by:

H_0: "the ideal profile is associated to a low liking potential": it is defined randomly (no structure)
H_1: "the ideal profile is associated to a high liking potential": it is not defined randomly (structure).

To perform the test, the distribution under H_0 of the liking potential related to the averaged ideal product is estimated for each consumer. The procedure used here to obtain this distribution under H_0 consists in simulating many random situations (in practice 500) and to compute $\hat{h}^{'}_j|\bar{z}_{j..}$ every time. The *real* estimated liking potential is then positioned on this distribution according to the usual approach of hypothesis testing in statistics. Specifically, we count (in percentage) how many times the liking potential obtained in random situation is higher than the *real* estimated liking potential ($\hat{h}^{'}_j|\bar{z}_{j..} > \hat{h}_j|\bar{z}_{j..}$), this percentage being used as a p-value.

In practice, the random situations are obtained by random permutation of the individual hedonic judgments. This permutation procedure aims at putting consumers in situations where they would score their liking randomly without generating new liking scores. A summary of the simulation procedure is given ***Figure 2***.

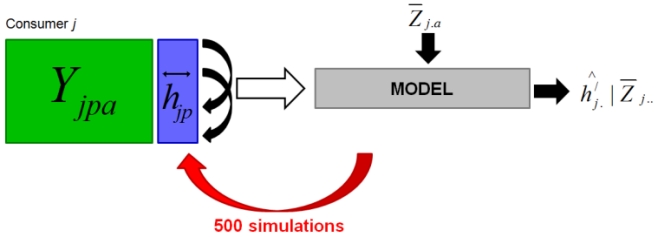


Figure 2: Procedure used to estimate the distribution of the liking potential under H_0.

A high observed liking potential $\widehat{h}_j|\bar{z}_{j\cdot}$ is simple to interpret as it validates the agreement between all types of data at once. In the opposite case, one can wonder whether the low estimation is caused by wrong hedonic judgments or incorrect ideal profiles. Obviously, this cannot be formally answered. However, it is hard to imagine a consumer providing a "good" ideal profile on one hand, and hedonic judgments which are not related to the sensory descriptions on the other hand. Thus, for abusive but pragmatic reasons, the inconsistency between ideal profiles and hedonic judgments is interpreted as a non-validation of the ideal profiles.

2.2.2 Model

In order to estimate the liking potential of the ideal products, various families of models can be considered. Here, only one type of models based on *PCR* is considered. A PCA is performed on the product profile obtained from each consumer. The liking scores are regressed on the first five principal components of that individual PCA. Only linear effects are used (***Equation 1***).

A selection of the best model by removing step by step the non-significant dimensions is performed.

$$\mathcal{M}_{PCR}^{(j)}:\ h_{jp} = \mu + \alpha_1^{(j)} Dim_1^{(j)} + \cdots + \alpha_i^{(j)} Dim_i^{(j)} + \cdots + \alpha_5^{(j)} Dim_5^{(j)} + \varepsilon_{jp} \quad (1)$$

With

μ : the constant

$\alpha_i^{(j)}$: the regression weight associated to the dimension *i* on overall liking for the consumer *j*

ε_{jp} : the residual

For each consumer, the liking potential $\widehat{h}_j|\bar{z}_{j\cdot}$ of each averaged ideal product is estimated. In that process, it is important to look at the quality of the individual models. It can be evaluated through the goodness of fit and through the significance of the individual models (*p-value*). Concerning the goodness of fit, the individual models don't have all the same number of degrees of freedom due to the selection of the best model. Hence, the adjusted R^2 is used.

The consumers for whom the model doesn't fit the data (i.e. no model can be found) are dismissed. This elimination is done based on an α-risk of 10% due to the low number of degrees of freedom.

3. Results

3.1 Quality of the individual models

3.1.1 Selection of the best individual models

For each consumer, a selection of the best model is performed. In some cases, this selection leads to the "elimination" of certain consumers as it is not always possible to explain the appreciation with the sensory dimensions selected, no dimension being significant. In the perfume study, it is the case for 18 consumers. Hence, 85 models are retained.

3.1.2 Goodness of fit of the individual models for the different families

The distribution of the adjusted R^2 associated with the individual models is represented ***Figure 3***. It seems here that the coefficients are globally high. For most consumers, the liking scores are well predicted, as the individual models fit the data. Hence, these individual models are reliable and can be used to estimate the liking potential of the different averaged ideal products.

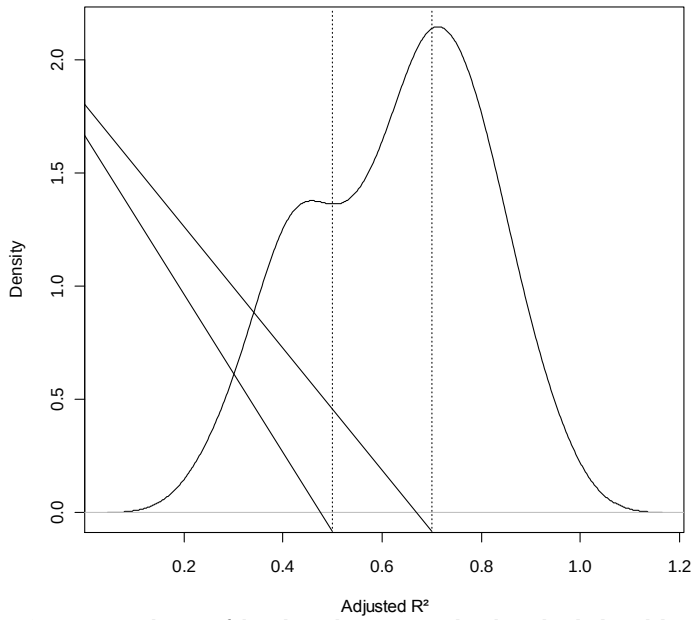


Figure 3: Distribution of the adjusted R^2 associated to the individual models.

3.1.3 Significance of the estimated liking potential of the individual averaged ideal products

The next step in the validation procedure consists in checking whether the ideal descriptions are obtained randomly or not. To do so, the significance of the liking potential of the averaged ideal products is tested. This test is done based on simulations.

The distribution of the p-values associated to the *real* estimated liking potential of the averaged ideal products is given ***Figure 4***.

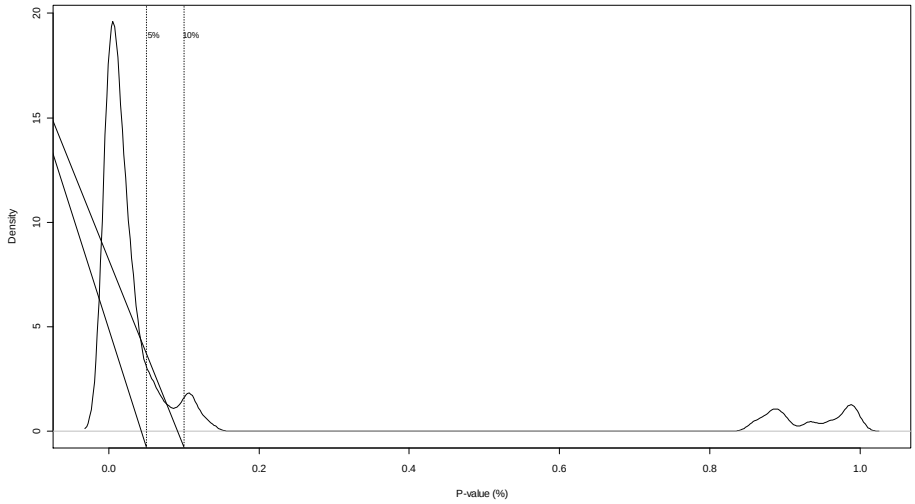


Figure 4: Distributions of the p-values of the observed liking potentials.

For most of the consumers, the observed liking potential is significant at the 5% threshold (***Figure 4***). It is the case for 75% of the consumers. This percentage increases at 84% when the 10% threshold is considered. In other words, the null hypothesis is rejected for most of the consumers. Hence, the ideal products are associated with a high estimated liking potential that couldn't be obtained randomly, showing that the ideal ratings are not given randomly and are consistent with the other descriptions (sensory and hedonic).

3.2 Liking potential of the averaged ideal profiles *vs.* liking scores of the products

Besides the significance of the estimated liking potential of the averaged ideal products, one still needs to check whether the ideal ratings provided by a consumer are *valid*. The liking potential of an averaged ideal profile has no value on its own, and only makes sense when it is compared to the liking scores given to the products. Therefore, the estimated liking potential of each averaged ideal product is *standardized* according to the liking scores given to the products by the consumer considered. For each consumer, the averaged liking score given to the products is subtracted from the estimated liking potential of his/her averaged ideal product. The difference is then standardized by dividing it by the standard deviation of the liking scores given to the products (***Equation 2***).

$$standardized\ liking\ potential_j = \frac{\hat{h}_j|\bar{\mathbf{z}}_{j\cdot} - \bar{h}_{j\cdot}}{\sigma_{\mathbf{h}_{j\cdot}}} \quad (2)$$

with

$\hat{h}_j|\bar{\mathbf{z}}_{j\cdot}$: the liking potential of the averaged ideal profile provided by the consumer *j*

$h_{j\cdot}$: the averaged liking score given to the products by the consumer *j*

$\sigma_{\mathbf{h}_{j\cdot}}$: the standard deviation of the liking scores given to the products by the consumer *j*

By definition, this standardized liking potential is high for consumers who described valid ideal profiles. As the quality of the individual models is of main interest for the good estimation of the liking potentials of the ideal products, these standardized liking potentials are represented in function of the adjusted R^2 of the corresponding individual models (***Figure 5***).

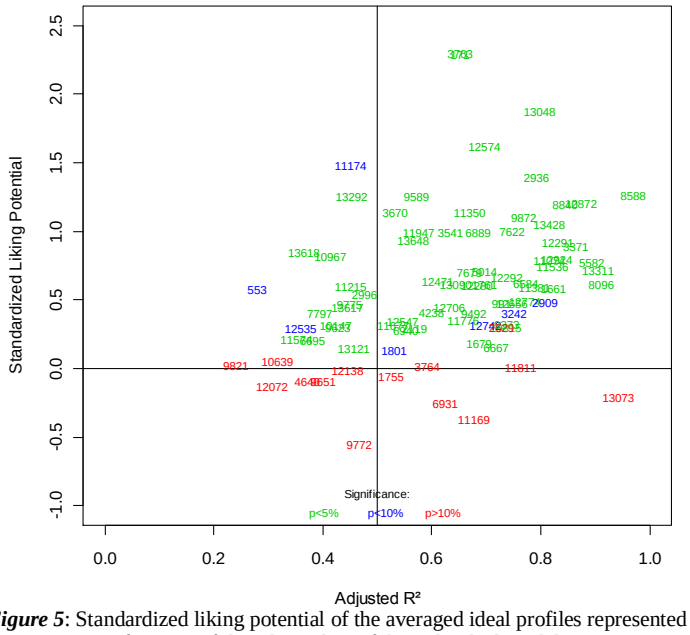


Figure 5: Standardized liking potential of the averaged ideal profiles represented in function of the adjusted R^2 of the individual models.

The results show that the standardized liking potentials are always positive, when the liking potentials are significant at 10% or less (consumers represented in *green* and *blue* ***Figure 5***). Hence, the estimated liking potential of the averaged ideal products are larger than the averaged liking scores. In other words, the ideal products would be more appreciated: they are *potential ideals* as they represent valid ideal descriptions.

Actually, for 51% of the consumers, the standardized liking potentials are higher or equal to 0.5. This percentage drops to 22% when a threshold of 1 is considered.
Nevertheless, 15% of the consumers have a negative (or close to 0) standardized liking potential. These consumers all have a non-significant liking potential, i.e. their ideal descriptions are not consistent. This might be explained by the fact that their corresponding individual models don't fit the data (adjusted $R^2 \leq 0.5$), or is due to the fact that the consumers described their ideal randomly.

4. Conclusion

Although they are often criticized, the sensory methodologies measuring (in)directly ideal intensities from consumers (such as the *IPM* or *JAR*) are widely used by practitioners, who seem to take benefits from them. The study presented here provides arguments for the use of such methodologies. It presents an idea on how to "validate" the ideal ratings directly provided from consumers by measuring their reliability according to other descriptions (sensory and hedonic).
To do so, the liking potential of the averaged ideal product from each consumer is estimated and compared to the liking scores they gave to the products. The significance of these estimated liking potentials has also been tested by comparing them to situations where consumers would rate their liking scores randomly.

In the *perfume* example, the results support the use of *IPM*. Indeed, for most of the consumers, the ideal descriptions are reliable: the ideal descriptions are not obtained randomly (i.e. the *real* estimated liking potentials are significant) and the estimated liking potentials of the averaged ideal products are higher than the liking scores provided for the products themselves. This is especially true when the relationship between the hedonic judgments and the sensory descriptions is strong (i.e. individual models with a high goodness of fit). However, when this relationship is weak, it is difficult to draw conclusions about the reliability of the ideal descriptions. Indeed, in that case, the estimated liking potentials are usually not significant, and it is unclear whether it is due to the lower quality of the predictive model, or to unreliable ideal ratings.

For the good understanding and the good use of ideal information provided by consumers, this validation step seems essential. Unfortunately, it can only be applied to data collected according to the *IPM*. Indeed, for JAR data, important information such as the sensory profiles of the products and the individual ideal profiles is missing.

It is also a good complement to the assessment of the consistency of the ideal descriptions proposed by ***Worch et al.*** (***2012***), which aims at measuring the relationship between the ideal descriptions and the sensory and hedonic ratings of the products.

References

MacFie, H.J., Bratchell, N., Greenhoff, K., & Vallis, L.V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *Journal of Sensory Studies*, 4, 129-148.

Stone, H., Sidel, J., Oliver, S., Woosley, A., & Singleton, R.C. (1974). Sensory evaluation by quantitative descriptive analysis. *Food Technology*, 28, 24-34.

Van Trijp, H.C.M., Punter, P.H., Mickartz, F., & Kruithof, L. (2007). The quest for the ideal product: Comparing different methods and approaches. *Food Quality and Preference*, 18, 729-740.

Worch, T., Dooley, L., Meullenet, J.F., & Punter, P.H. (2010). Comparison of PLS dummy variables and Fishbone method to determine optimal product characteristics from ideal profiles. *Food Quality and Preference*, 21, 1077-1087.

Worch, T., Lê, S., & Punter, P. (2010). How reliable are the consumers? Comparison of sensory profiles from consumers and experts. *Food Quality and Preference*, 21, 309-318.

Worch, T., Lê, S., Punter, P.H., & Pagès, J. (2012). Assessment of the consistency of ideal profiles according to non-ideal data for the Ideal Profile Method. *Food Quality and Preference*, 24, 99-110.

Session 8 : Chimiométrie II / *Chemometrics II*

Optimisation de B-splines par Algorithme Génétique pour une PLS non linéaire : Application en chimiométrie et en drug-design

B-spline optimization with genetic algorithms for a non-linear PLS : Application to chemometrics and drug design

Christelle Reynès[1], & Robert Sabatier[1]

[1] *Laboratoire de Physique Industrielle et Traitement de l'Information, EA 2415, Université Montpellier 1, FRANCE*
E-mail : *christelle.reynes@univ-montp1.fr*

Résumé
La méthode Partial Least Squares (PLS) est une méthode très puissante de régression linéaire. Son efficacité est reconnue dans de nombreux domaines, notamment en chimiométrie et dans la conception de molécules thérapeutiques. Cependant, cette méthode est limitée par son aspect *linéaire* qui ne peut convenir à tous les jeux de données. C'est pourquoi l'introduction de non linéarité à l'intérieur de la méthode PLS est un important sujet de recherche. Nous présentons ici une méthode très flexible permettant d'ajuster la non linéarité à l'aide de fonctions B-splines dont les paramètres sont ajustés par Algorithme Génétique. L'optimisation tient compte de la parcimonie et de la précision du modèle de régression obtenu. La méthode est appliquée à deux jeux de données.
Mots-clés : régression non linéaire, algorithme génétique, PLS, B-splines

Abstract
Partial Least Squares (PLS) is a really powerful method for linear regression. Its efficiency is well known in many application fields, especially in chemometrics and drug design. However, this method is limited by its linear way of modelling data, which cannot fit all datasets. Hence, introducing non linearity in PLS has led to several studies. The method introduced here is a very flexible one, which allows to fit any kind of non linearity thanks to B-spline functions whose parameters are optimized through a genetic algorithm. The optimization takes into account both parcimony and precision of the obtained regression model. The method is applied to two datasets.
Keywords : non linear regression, genetic algorithms, partial least squares, B-splines

1 Introduction

La méthode Partial Least Squares (PLS, Wold, 1966) est reconnue et utilisée dans une grande variété de domaines. Son utilisation est notamment importante dans le domaine de la chimiométrie

et pour la conception de molécules thérapeutiques. En particulier, les analyses ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) permettent d'analyser le comportement de composés pharmaceutiques dans l'organisme (Lagorce et al., 2011) et nécessitent souvent de prédire la valeur d'un indicateur continu (perméabilité, solubilité,...) d'où l'utilisation de méthodes telles que PLS (Obrezanova et al., 2008). Cependant, PLS est adaptée si la relation entre la ou les variable(s) à prédire et les variables prédictives est linéaire. Or, ce n'est évidemment pas toujours le cas. Plusieurs méthodes ont été proposées pour introduire la prise en compte de la non linéarité dans un contexte de régression PLS. La plus adaptative consiste à transformer indépendemment chaque variable. Dans ce cas, à moins d'un savoir *a priori*, il est plus intéressant d'utiliser des fonctions pouvant s'adapter à de nombreuses tendances, c'est le cas des B-splines (Durand and Sabatier, 1997). Cependant, l'optimisation des nombreux paramètres (nombre et position des nœuds, degré polynomial *pour chaque variable*) est réalisée soit manuellement, soit dans un cadre très simplifié (Lombardo et al., 2009).
La méthode proposée, appelée AG-PLSS, permet d'optimiser automatiquement les paramètres des transformations B-splines. Cette optimisation est réalisée à l'aide d'un Algorithme Génétique (AG). La méthode sera testée sur deux jeux de données.

2 Méthode

2.1 Rappels sur PLS

Soit $\mathbf{Y}$ la matrice $n \times p$ ($p \geq 1$) contenant la valeur de p variables réponses pour les n observations et $\mathbf{X}$ la matrice $n \times q$ ($q > 1$) des q variables prédictives. Ces deux matrices sont centrées et éventuellement réduites. La méthode PLS usuelle consiste à rechercher des combinaisons linéaires successives des variables prédictives, $\mathbf{t}_k$ ($k = 1, \ldots, A$), et des combinaisons linéaires successives des variables à prédire, $\mathbf{u}_k$, en maximisant la covariance entre $\mathbf{t}_k$ et $\mathbf{u}_k$. Cette optimisation est obtenue, en particulier, par l'algorithme NIPALS (Tenenhaus, 1998). Le nombre A de composantes à retenir est généralement choisi par validation croisée (CV). Une fois A choisi, $\{\mathbf{t}_1, \ldots, \mathbf{t}_A\}$ fournit une base orthogonale de $Im(\mathbf{X})$, ainsi, la projection orthogonale de $\mathbf{Y}$ sur cet espace engendre le modèle $\hat{\mathbf{Y}}_A$.

2.2 Rappels sur les transformations B-splines

Une fonction B-splines est un polynôme par morceaux de degré d comportant K nœuds. On peut construire un modèle spline additif pour modéliser chacune des variables de $\mathbf{Y}$:

$$\hat{\mathbf{y}}_A^j = \hat{f}_A^{j,1}(\mathbf{x}^1) + \ldots + \hat{f}_A^{j,q}(\mathbf{x}^q),$$

où $\hat{\mathbf{y}}_A^j$ est la reconstruction de rang A de la j-ème colonne de $\mathbf{Y}$ et $\hat{f}_A^{j,i}(\mathbf{x}^i)$ est une fonction B-spline appliquée à la i-ème colonne de $\mathbf{X}$, $\mathbf{x}^i$. On note d_i le degré du polynôme et K_i, le nombre de noeuds pour la i-ème variable, $\hat{a}_{k,A}^{j,i}$ les coefficients splines optimisés par PLS et $\mathbf{B}^i = \{\mathbf{B}_1^i, \ldots, \mathbf{B}_{r_i}^i\}$ l'ensemble de la base des fonctions splines pour la i-ème variable (De Boor, 2001). On peut alors définir :

$$\hat{f}_A^{j,i}(\mathbf{x}^i) = \sum_{k=1}^{r_i} \hat{a}_{k,A}^{j,i} \mathbf{B}_k^i(\mathbf{x}^i),$$

où $r_i = d_i + K_i + 1$ Alors, il est possible d'appliquer la méthode PLS usuelle sur $\mathbf{Y}$ et $\mathbf{B} = \{\mathbf{B}^1, \ldots, \mathbf{B}^q\}$. Cette méthode, sans optimisation des paramètres des transformations, sera notée

PLSS (Durand, 2001).

2.3 Rappels sur les Algorithmes Génétiques

Les Algorithmes Génétiques (AG) appartiennent à une famille de méthodes d'optimisation heuristique utilisée pour résoudre des problèmes complexes, notamment combinatoires (Reeves and Rowe, 2002). Ils sont inspirés de la sélection naturelle. En effet, l'algorithme opère sur une population de solutions potentielles qui évolue le long de génération grâce à trois opérateurs : mutation, croisement et sélection. Cette sélection est basée sur la qualité des individus mesurée à l'aide d'une fonction appelée *fitness*.

2.4 Description de la méthode AG-PLSS

Cette partie décrit la construction d'un AG pour l'optimisation des paramètres B-splines dans le contexte de la régression PLS : degré, nombre et position des nœuds pour chaque variable prédictive, ainsi que la dimension A du modèle PLS.

2.4.1 Construction du fitness

Cette étape va permettre de quantifier la valeur de chaque solution potentielle. Ici, il s'agit d'obtenir un modèle précis mais généralisable à de nouvelles observations. Cette dernière partie est particulièrement importante car les B-splines peuvent s'adapter tellement précisément qu'elles peuvent créer un sur-ajustement.
Pour le cas d'un seul $\mathbf{Y}$ à prédire la précision du modèle peut être quantifiée par le R^2, part de la variance de $\mathbf{Y}$ expliquée par $\hat{\mathbf{Y}}_A$ pour le modèle choisi. Quant à la généralisabilité, elle est quantifiée par l'écart entre R^2_{ap}, le R^2 obtenu sur le jeu d'apprentissage et R^2_{CV}, le R^2 obtenu par validation croisée. Pour le cas de plusieurs variables à prédire, on prendra le coefficient RV (Escoufier, 1973) pour la précision et le nombre de paramètres des transformations pour la généralisabilité. Nos applications ne concernent que des modèles de PLS1, c'est donc le premier critère que nous allons détailler.
On obtient, pour la solution potentielle s, la fonction de fitness suivante :

$$fit(s) = \frac{R^2_{ap}(s) + R^2_{CV}(s)}{2} - \left(R^2_{ap}(s) - R^2_{CV}(s)\right).$$

2.4.2 Déroulement de AG-PLSS

On génère une population initiale de T_{pop} solutions potentielles. On cherche à obtenir des solutions très hétérogènes afin de réaliser l'exploration la plus large possible de l'espace des solutions. Pour chaque solution potentielle s, on choisit aléatoirement et de manière uniforme les différents paramètres : $A \in \{1, \ldots, q\}$, $K_i \in \{1, \ldots, 3\}$ pour $i \in \{1, \ldots, p\}$ et la localisation des K_i nœuds est choisie dans l'étendue de la i-ème variable.
Le croisement (appliqué à une proportion π_c de la population) consiste à échanger certaines caractéristiques de deux solutions potentielles pour en obtenir deux nouvelles. La mutation (appliquée à une proportion π_m de la population), permet de modifier les solutions potentielles par diminution/augmentation du degré, du nombre de nœuds ou par changement de la localisation des nœuds. Enfin, pour réaliser l'étape de sélection, une probabilité proportionnelle à la qualité de chaque solution est calculée. Les différentes solutions sont alors introduites dans la génération suivante à hauteur de cette probabilité, excepté pour la meilleure solution qui

	OLS	PLS	PLSS	AG-PLSS	SVM (gaussien)	Réseaux Neurones	Forêt Aléatoire
R^2	0.9140	0.9044	0.9139	0.9953	0.9253	0.7622	0.9562
R^2_{CV}	0.5898	0.6047	0.7189	0.8175	0.6156	0.6989	0.7135

Table 1: Comparaison des qualités de prédiction pour les *Jus d'oranges* après optimisation par validation croisée des différents paramètres (6 composantes ont été obtenues pour PLS et AG-PLSS et 2 composantes pour PLSS).

est automatiquement sélectionnée pour la génération suivante (élitisme). La population évolue durant N_{gene} générations à l'issue desquelles on choisit comme transformation finale celle correspondant à la meilleure solution de la dernière génération.
Les paramètres par défaut sont $N_{gene} = 500$, $T_{pop} = 200$, $\pi_c = 0.5$ et $\pi_m = 0.9$.

3 Résultats

3.1 Exemple issu de la chimiométrie

Afin de pouvoir comparer AG-PLSS à d'autres méthodes, nous avons utilisé un jeu de données de la littérature (Durand, 2001) décrivant 24 jus d'orange à l'aide de 10 descripteurs minéralogiques. L'objectif est de prédire la valeur d'un descripteur sensoriel. Les résultats de différentes méthodes sont donnés dans la Table 1. OLS est la régression linéaire par moindres carrés classique et PLSS est utilisée comme indiqué dans Durand (2001), c'est-à-dire que tous les paramètres sont fixés aux valeurs indiquées par cet article puisqu'il n'est pas prévu d'optimisation.
Il est très intéressant de constater que les variables initialement corrélées à la variable à prédire ne sont que peu transformées. Pour les variables qui n'étaient que très peu (ou pas) corrélée à la variable à prédire, on a deux situations : soit elles ne sont pas transformées, soit elles subissent une transformation importante. Celles qui ne sont pas transformées ne jouent quasiment aucun rôle dans le modèle alors que celles qui ont été transformées se trouvent être les plus explicatives du modèle. L'algorithme a donc su laisser de côté une information dont il n'avait pas besoin et au contraire accentuer la transformation des variables qui le nécessitaient pour améliorer le modèle.

3.2 Exemple issu de la conception de molécules thérapeutiques

Cette application porte sur un paramètre très important dans la conception de molécules thérapeutiques, le volume de distribution à l'état d'équilibre, noté Vss. Sa prédiction est très importante puisqu'il relie la quantité de médicament présente dans un compartiment à sa concentration plasmatique. Cependant, cette prédiction est très difficile.
Pour prédire cette propriété on dispose, pour 177 molécules, de leur valeur de Vss et de 1666 descripteurs de leur structure 1D, 2D et 3D fournis par le logiciel e-Dragon (http://www.vcclab.org/). Cependant, pour améliorer le modèle, on réalise une sélection des variables en deux étapes: une étape dite *généraliste* qui permet d'éliminer les variables constantes et redondantes et une étape dite *spécialiste* qui sélectionne les variables les plus intéressantes pour PLS et pour lesquelles la transformation spline est la plus bénéfique. On obtient finalement 25 variables.
Les résultats obtenus pour ce jeu de données sont présentés dans la Table 2. On constate qu'AG-PLSS est bien moins sujet au sur-ajustement que PLS puisqu'il perd moins de 4% de

	PLS	AG-PLSS	SVM	NN	RF
R^2	0.6384	0.6059	0.7956	0.8879	0.9078
R^2_{CV}	0.5571	0.5765	0.6212	0.5391	0.6066
R^2_{test}	0.4710	0.5695	0.2097	0.2821	0.3166

Table 2: Comparaison des qualités de prédiction pour *Vss* (paramètres optimisés par validation croisée).

variabilité expliquée entre jeu d'apprentissage et jeu test alors que PLS perd plus de 16%. Par ailleurs, sur le jeu test, AG-PLSS permet d'expliquer 10% de plus que PLS. On a donc un réel gain tant en précision qu'en généralisabilité.

4 Discussion - Conclusion

L'introduction de non-linéarité dans la méthode PLS est un enjeu très important pour traiter les jeux de données pour lesquels une modélisation linéaire n'est pas adaptée. Mais une telle généralisation induit nécessairement l'augmentation importante du nombre de paramètres avec une réelle complexité pour fixer les valeurs. AG-PLSS propose une méthode entièrement automatisée qui permet, par des transformations B-splines individuelles de chaque variable, de s'ajuster au mieux aux données tout en maintenant ses performances sur un jeu de données indépendant. Ce type de méthode est particulièrement intéressant dans les domaines où les problèmes de régression sont très présents comme la chimiométrie ou le *drug design*. Les jeux de données proposés ont permis de montrer les performances de la méthode, sa capacité à obtenir des transformations efficaces et à ne transformer que les variables qui le nécessitent.

References

De Boor, C. (2001). *A practical guide to splines*. Springer Verlag.

Durand, J. (2001). Local polynomial additive regression through PLS and splines: PLSS. *Chemometrics and Intelligent Laboratory Systems*, 58(2):235–246.

Durand, J. and Sabatier, R. (1997). Additive Splines for Partial Least Squares Regression. *Journal of the American Statistical Association*, 92(440).

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

Lagorce, D., Reynes, C., Camprouc, A.-C., Miteva, M. A., Sperandio, O., and Villoutreix, B. O. (2011). In silico adme/tox predictions. In K. Tsaioun, S. K., editor, *ADMET for Medicinal Chemists: A practical guide*. Wiley.

Lombardo, R., Durand, J., and De Veaux, R. (2009). Model building in multivariate additive partial least squares splines via the GCV criterion. *Journal of Chemometrics*, 23(12):605–617.

Obrezanova, O., Gola, J., Champness, E., and Segall, M. (2008). Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility. *Journal of Computer-Aided Molecular Design*, 22(6):431–440.

Reeves, C. and Rowe, J. (2002). *Genetic algorithms: principles and perspectives: a guide to GA theory*. Kluwer Academic Pub.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions Technip.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In Krishnaiah, P., editor, *Multivariate Analysis*, pages 391–420. Academic Press.

Analyse des redondances multibloc. Application aux usages médicamenteux en élevages

Multiblock redundancy analysis. Application to drug use on farms

Stéphanie Bougeard[1], Fatima Laanaya-Tazani[1,2], Sophie Le Bouquin[1] & Claire Chauvin[1]

[1] *Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses), unité d'épidémiologie, Zoopôle, 22440 Ploufragan, France*
E-mail : *stephanie.bougeard@anses.fr ; sophie.lebouquin-leneveu@anses.fr ; claire.chauvin@anses.fr*
[2] *Université de Rennes 2, Place du recteur Henri Le Moal, 35043 Rennes*
E-mail : *atazani@yahoo.fr*

Résumé
Une analyse factorielle optimisée permettant l'étude conjointe de $(K+1)$ tableaux de données est proposée pour le cas où un tableau Y est expliqué par K tableaux $(X_1, \ldots, X_K)$. La méthode proposée est basée sur une extension du critère de l'analyse des redondances. Les méthodes multiblocs fournissent à l'utilisateur un grand nombre d'indices d'aide à l'interprétation des liens entre variables et blocs de variables. Nous proposons dans le cadre de l'analyse des redondances multibloc des indices cohérents, directement issus du critère à maximiser. La démarche sera illustrée sur la base d'une étude de cas en épidémiologie vétérinaire.
Mots-clés : Régression multibloc, analyse des redondances multibloc, épidémiologie vétérinaire

Abstract
We discuss factor analytic methods to study a set of $(K+1)$ datasets where we wish to explain a dataset Y from K other datasets $(X_1, \ldots, X_K)$. The method of analysis is based on an extension of Redundancy Analysis. Multiblock modeling methods provide to the user a large spectrum of interpretation indices for the investigation of the relationships among variables and among datasets. The interest of multiblock Redundancy Analysis and the associated interpretation tools are illustrated using a dataset in the field of veterinary epidemiology.
Keywords : Multiblock modelling, multiblock redundancy analysis, veterinary epidemiology

1 Introduction

L'explication d'une variable composite par plusieurs variables explicatives est une problématique statistique généralement solutionnée par les méthodes analysant deux tableaux, *i.e.*, régression PLS2, analyse des redondances ou analyse canonique pour les plus classiques. Lorsque les variables explicatives sont nombreuses et présentent une structure en bloc ayant une réelle signification pour l'interprétation, les méthodes multiblocs pour $(K+1)$ tableaux sont alors utilisées. La plus classique est la régression PLS multibloc (Wold, 1984), mais il est démontré que son application à l'explication d'un seul tableau Y revient à une régression PLS2 classique (Westerhuis

et al., 1998; Qin *et al.*, 2001; Vivien, 2002). Nous proposons une méthode alternative basée sur l'extension de l'analyse des redondances à plusieurs tableaux explicatifs $(X_1, \ldots, X_K)$, appelée analyse des redondances multibloc. Cette méthode est basée sur un critère à maximiser clair et cohérent avec ses objectifs. La méthode prend en compte la structure en bloc des variables explicatives et fournit des solutions directes. Elle est plus orientée vers l'explication du tableau Y que la régression PLS multibloc, mais peut être plus sensible à la multicolinéarité au sein des blocs de variables X_k (Bougeard *et al.*, 2011a). Une application de cette méthode est proposée sur un jeu de données d'épidémiologie vétérinaire. Des indices d'aide à l'interprétation, spécifiques au cadre multibloc, sont proposés et illustrés.

2 Analyse des redondances multibloc

L'objectif est de réaliser une description ainsi qu'une explication du tableau Y contenant Q variables à partir d'un tableau X contenant P variables, partitionné en K blocs, $X = [X_1|\ldots|X_K]$. Chaque tableau X_k est de format $(N \times P_k)$, tel que $\sum_{k=1}^{K} P_k = P$. Toutes ces variables quantitatives sont mesurées sur les mêmes N individus et sont supposées être centrées.

Le principe de la méthode est que chaque tableau est résumé, pour chaque dimension $h = (1, \ldots, H)$, par une variable latente. Les variables latentes partielles $t_k^{(h)} = X_k w_k^{(h)}$ constituent des résumés des tableaux X_k ; $u^{(h)} = Y v^{(h)}$ résume le tableau à expliquer Y ; $t^{(h)} = X w^{(h)}$ constitue un résumé du tableau concaténé $X = [X_1|\ldots|X_K]$. Pour la dimension $(h = 1)$, la détermination des variables latentes est issue de la maximisation du critère (1), ou de celle, équivalente, du critère (2).

$$\sum_{k=1}^{K} cov^2(u^{(1)}, t_k^{(1)}) \text{ avec } ||t_k^{(1)}|| = ||v^{(1)}|| = 1 \tag{1}$$

$$cov^2(u^{(1)}, t^{(1)}) \text{ avec } t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)},\ \sum_{k=1}^{K} a_k^{(1)^2} = 1,\ ||t_k^{(1)}|| = ||v^{(1)}|| = 1 \tag{2}$$

La solution de cette maximisation est donnée par $v^{(1)}$ premier vecteur propre de la matrice $\sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$ associée à la valeur propre $\lambda^{(1)}$. L'ensemble des variables latentes est ensuite déduit de cette solution. Les solutions d'ordre suivant, *i.e.*, $h = (2, \ldots, H)$, sont issues de la maximisation du même critère, en remplaçant les tableaux $(X_1, \ldots, X_K)$ par leurs résidus de régression sur les sous-espaces engendrés par les variables latentes globales t. Afin de faciliter l'interprétation, il convient de déterminer le sous-espace optimal composé de $(t^{(1)}, \ldots, t^{(h_{opt})})$ variables latentes globales. Pour cela, une validation croisée, basée sur un échantillon de calibration et de validation, permet d'évaluer la qualité d'explication et de prédiction du modèle (Stone, 1974). Les détails de la résolution et les propriétés associées sont donnés dans (Bougeard *et al.*, 2007). Une comparaison à d'autres méthodes multiblocs, notamment à la régression PLS multibloc (Wold, 1984) est détaillée dans (Bougeard *et al.*, 2011a).

Afin de rendre interprétables les informations issues de la régression multibloc, plusieurs indices sont proposés (Bougeard *et al.*, 2011b). Seuls ceux permettant une interprétation du modèle optimal sont détaillés ici. L'interprétation la plus classique, mais aussi la moins synthétique, se situe au niveau des variables, généralement nombreuses en régression multibloc. Les coefficients de régression de l'ensemble des variables X pour expliquer les variables Y sont

donnés par la matrice de coefficients, selon l'Eq. (3).

$$\beta_{p,q}^{(1\to h_{opt})} = \sum_{h=1}^{h_{opt}} w^{(h)*} c^{(h)'} \tag{3}$$

avec $w^{(h)*} = \prod_{l=1}^{h-1}[I - w^{(l)}(t^{(l)'}t^{(l)})^{-1}t^{(l)'}X]w^{(h)}$ et $c^{(h)} = (t^{(h)'}t^{(h)})^{-1}Y't^{(h)}$. Lorsque plusieurs variables à expliquer Y sont en jeu, l'utilisateur a besoin de hiérarchiser l'importance des variables X pour l'explication du tableau Y. Pour cela, l'indice VIP issu de la régression PLS classique a été adapté au cadre multibloc selon l'Eq. (4), où les coefficients w sont pondérés par le poids du bloc a_k^2 dont ils sont issus et par le poids de la dimension λ.

$$VarImp_p^{(1\to h_{opt})} = \frac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} \frac{a_k^{(h)^2} w_{[p]}^{(h)^2}}{\sum_{p=1}^{P} a_k^{(h)^2} w_{[p]}^{(h)^2}}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}} \tag{4}$$

Il est aussi intéressant de hiérarchiser l'importance de chaque bloc dans l'explication du tableau Y. Pour cela, l'indice BIP proposé par Vivien (2002) dans le cadre des méthodes multiblocs a été adapté selon l'Eq. (5), où les poids des blocs a_k^2 sont pondérés par le poids de la dimension λ associée.

$$BlockImp_k^{(1\to h_{opt})} = \frac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} a_k^{(h)^2}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}} \tag{5}$$

3 Application en épidémiologie vétérinaire

3.1 Données et objectifs

Les données proviennent d'une enquête épidémiologique analytique menée en 2010 sur un échantillon représentatif de 112 élevages français. L'objectif est de caractériser les intrants médicamenteux et d'analyser leurs déterminants. Un questionnaire administré à chaque éleveur a permis de recueillir 177 variables relatives aux caractéristiques de l'élevage ainsi qu'aux consommations médicamenteuses. Parmi ces variables, 27 ont été sélectionnées pour leurs liens avec ces consommations. Le tableau Y à expliquer est composé de deux variables mesurant la consommation de médicaments administrés soit par voie médicamenteuse, soit par voie alimentaire. Les variables explicatives X sont organisées en quatre blocs, *i.e.*, X_1 relatif à la conduite d'élevage et aux mesures d'hygiène (8 variables), X_2 reflétant les problèmes sanitaires (7 variables), X_3 relatif à la structure de l'élevage (5 variables) et X_4 lié aux pratiques thérapeutiques (7 variables). Les variables, ayant des unités de mesure différentes, sont centrées et réduites.

Les objectifs du traitement statistique sont à la fois descriptifs et explicatifs. L'épidémiologiste est tout d'abord intéressé par la description des liens entre les variables et entre les blocs de variables. Le second objectif est de déterminer, parmi les variables explicatives, celles qui sont facteurs de risque d'une variable composite, *i.e.*, les usages médicamenteux. Le troisième objectif est de mesurer l'influence des blocs de variables explicatives dans l'explication de ces usages. La finalité pour l'épidémiologiste est d'avoir une vision globale des actions à mener en élevage pour comprendre et aider à la réduction des usages médicamenteux.

3.2 Interprétation au niveau des variables

L'ensemble des variables explicatives et à expliquer peut être représenté sur la base des composantes globales t orthogonales par construction. La Figure 1 illustre cette représentation pour les deux premières dimensions $t^{(1)}$ et $t^{(2)}$.

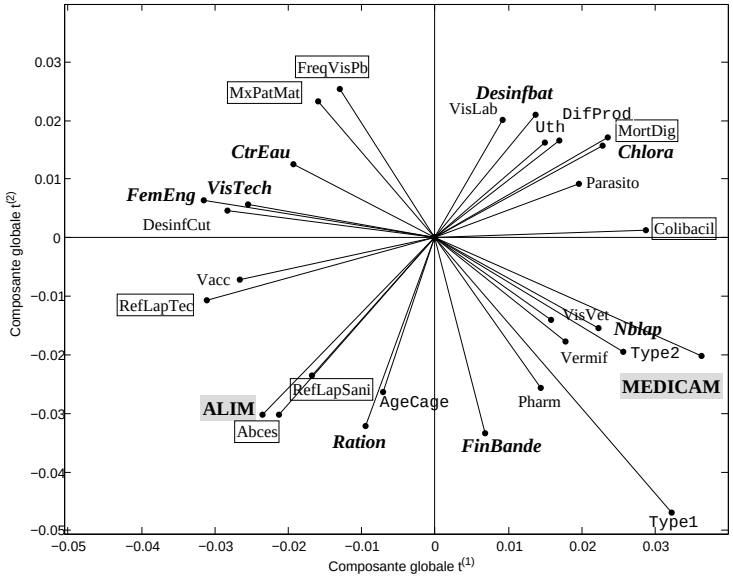


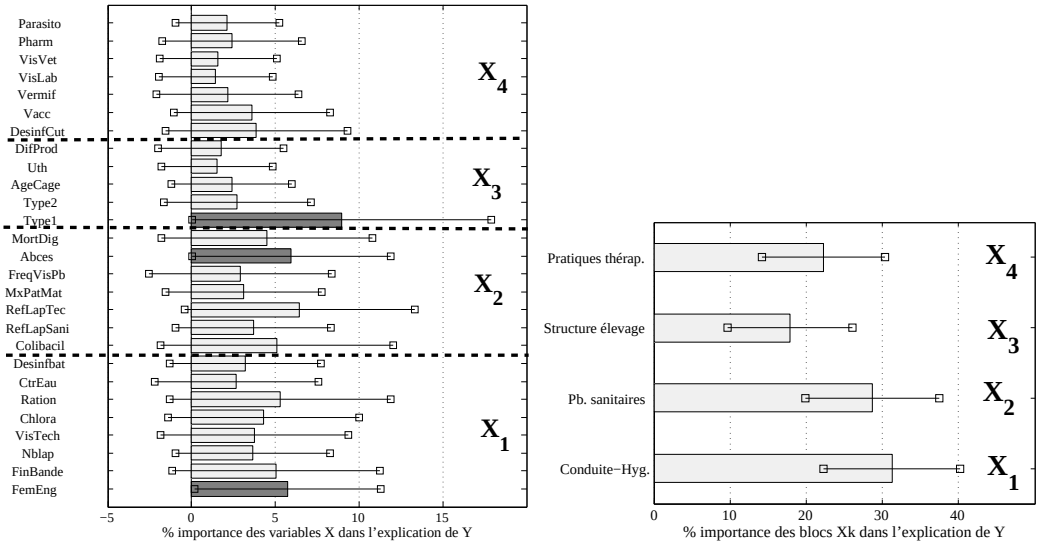
Figure 1: Représentation des variables par leurs coefficients sur le plan $(t^{(1)}, t^{(2)})$. Les variables Y sont en majuscule et grisé, les variables du bloc X_1 en italique gras, celles du bloc X_2 encadrées, celles du bloc X_3 en police `courrier` et celles du bloc X_4 normales.

Les variables Y relatives aux médicaments administrés par voie médicamenteuse (MEDICAM) ou alimentaire (ALIM) ne semblent pas liées entre elles. Chacune est liée à des facteurs de risque spécifiques appartenant aux différents blocs de variables explicatives (interprétation non détaillée dans ce résumé). Afin d'affiner cette interprétation, les coefficients de régression de chaque variable à expliquer par l'ensemble des variables explicatives peuvent aussi être donnés pour le modèle optimal ($h_{opt} = 2$ dimensions), associés à des intervalles de confiance boostrapés pour évaluer leur significativité (non donnés dans ce résumé). Ces interprétations au niveau des variables étant riches, l'épidémiologiste a aussi besoin d'indicateurs synthétiques pour dégager des pistes d'action claires.

3.3 Interprétation au niveau des blocs de variables

Dans l'objectif de hiérarchiser le poids des variables explicatives dans la médication, qu'elle soit médicamenteuse ou alimentaire, les indices $VarImpCum$ sont donnés par la Figure 2(a), pour le modèle optimal ($h_{opt} = 2$ dimensions), associés à des intervalles de confiance boostrapés pour évaluer leur significativité. Afin de hiérarchiser le poids des blocs de variables explicatives, les indices $BlocImpCum$ sont donnés par la Figure 2(b).

Il apparaît que les blocs explicatifs X_1, relatif à la conduite d'élevage et aux mesures d'hygiène ($BlocImpCum_{X_1} = 31,2\%\ [22,3; 40,2]_{95\%}$), et X_2, associé aux problèmes sanitaires



(a) Poids des variables (b) Poids des blocs

Figure 2: Représentation du poids des variables X et du poids des tableaux (X_1, X_2, X_3, X_4) dans l'explication du tableau Y (modèle optimal en $h_{opt} = 2$ dimensions).

($BlocImpCum_{X_2} = 28,7\%\ [20,0; 37,4]_{95\%}$) sont ceux qui ont le plus d'influence sur la médication Y. Parmi les variables explicatives, trois d'entre elles ont une forte influence sur l'usage médicamenteux Y : la présence de femelles lors de l'engraissement des petits ($VarImpCum_{FemEng} = 5,7\%\ [1,6; 11,3]_{95\%}$), la présence d'abcès ($VarImpCum_{Abces} = 6,0\%\ [0,0; 11,9]_{95\%}$) et le type d'animal ($VarImpCum_{Type1} = 8,9\%\ [0,0; 17,9]_{95\%}$).

4 Conclusion et perspectives

L'application de l'analyse des redondances multibloc à un jeu de données d'épidémiologie vétérinaire a permis, en comparaison aux résultats issus des méthodes plus classiques, d'expliquer une variable Y composite par des variables explicatives nombreuses et structurées en blocs $(X_1, \ldots, X_K)$. Les outils d'aide à l'interprétation associés permettent de valoriser les résultats à plusieurs niveaux. L'interprétation au niveau des variables, *i.e.*, représentations factorielles et coefficients de régression, permet une interprétation fine des facteurs de risque mais est souvent trop riche pour permettre à l'épidémiologiste de prendre des décisions. L'interprétation associée au niveau des blocs de variables permet ce recul.

L'analyse des redondances multibloc, ainsi que ses outils d'aide à l'interprétation, sont développés sur le logiciel Matlab et sont en cours de développement dans un package R, intégré dans le logiciel d'analyse de données ade4 (http://pbil.univ-lyon1.fr/ade4/). Des avancées méthodologiques sont en cours pour intégrer la structure hiérarchisée des observations aux méthodes multiblocs. Ce type de structure est fréquemment rencontré en pratique, notamment en épidémiologie vétérinaire. Dans le jeu de données présenté par exemple, ces méthodes permettront de prendre en compte un effet du groupement d'appartenance des élevages, très structurant dans l'analyse.

Bibliographie

Bougeard, S., Hanafi, M. & Qannari E. M. (2007). ACPVI multibloc. Application à des données d'épidémiologie animale. *Journal de la Société Française de Statistique*, 148(4) : 77-94.

Bougeard, S., Qannari, E.M., Lupo, C. & Hanafi, M. (2011a). From multiblock partial least squares to multiblock redundancy analysis. A continuum approach. *Informatica*, 22(1), 1-16.

Bougeard, S., Qannari, E.M. & Rose, N. (2011b). Multiblock Redundancy Analysis: interpretation tools and application in epidemiology. *Journal of Chemometrics*, 23, 1-9.

Qin, S.J., Valle, S. & Piovoso, M.J. (2001). On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15: 715-742.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36: 111-147.

Vivien, M. (2002). *Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications*. Thèse de doctorat, Université de Montpellier 1.

Westerhuis, J.A., Kourti, T. & MacGregor, J.F. (1998). Analysis of multiblock and hierarchical PCA and PLS model. *Journal of Chemometrics*, 12: 301-321.

Wold, S. (1984). Three PLS algorithms according to SW. *Proceeding of the Symposium MULDAST*, Umea University, Sweden.

Vision globale pour l'analyse de données multi-groupes. Application à la composition chimique d'huiles d'olives

Overview of methods of analysis of multi-group datasets. Application to the chemical composition of olive oils

Aida Eslami[1], El Mostafa Qannari[2], Achim Kohler[3] & Stéphanie Bougeard[1]

[1] *Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses), unité d'épidémiologie, Ploufragan, France*
E-mail : *aida.eslami@anses.fr*; *stephanie.bougeard@anses.fr*
[2] *École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes-Atlantique (Oniris), unité de Sensométrie et de Chimiométrie, Nantes, France*
E-mail : *elmostafa.qannari@oniris-nantes.fr*
[3] *Centre de Génétique Intégrative (CIGENE), Département de mathématiques Sciences et Techniques, Université norvégienne des Sciences de la Vie, Norvège*
E-mail : *achim.kohler@nofima.no*

Résumé
Il arrive fréquemment que des individus évalués par les mêmes variables présentent une structure multi-groupe; ces groupes pouvant comprendre un nombre différent d'individus. Plusieurs méthodes ont été développées pour étudier ce type de données. L'objectif de ces méthodes est de décrire les groupes d'individus sur la base de caractéristiques communes, *i.e.* composantes et/ou vecteurs de poids communs. Dans cette étude, nous proposons une vision globale et originale de quelques méthodes de type analyse en composantes principales multi-groupes. Ces méthodes sont comparées sur la base d'un ensemble de 572 huiles d'olives italiennes, originaires de neuf régions, décrites par huit variables relatives à leur composition en acide gras.
Mots-clés : Analyse en composantes principales communes, STATIS dual, analyse en composantes principales multi-groupes, analyse en composantes principales, comparaisons entre les groupes, données multi-groupes.

Abstract
In some studies, the data are presented in multi-group structure where the same variables are measured on a set of individuals partitioned into groups. Several methods have been developed to study multi-group datasets. The objective of these methods is to describe the groups by common characteristics, *i.e.* components and/or common vector of loadings. In this study, we consider a global overview of some multi-group principal component analysis methods. The comparison of these methods is done on a set of 572 Italian olive oils, sampled from nine regions of Italy, on which eight fatty acid variables were measured.
Keywords : Common principal components, dual STATIS, multi-group principal component analysis, principal component analysis, between group comparison, multi-group data.

1 Introduction

Il arrive fréquemment que les mêmes P variables soient mesurées sur un ensemble de N individus, divisés en M groupes (appelées par la suite données multi-groupes). Afin d'étudier la structure des données dans les différents groupes, l'analyse en composantes principales (ACP), outil largement utilisé pour la réduction de la dimensionnalité en analyse multivariée, peut être effectuée sur chaque groupe séparément. De toute évidence, cette stratégie fournit un grand nombre de paramètres, pouvant conduire à un problème de stabilité du fait du manque de degrés de liberté pour estimer ceux-ci. Cette stratégie entraîne de plus des difficultés dans l'interprétation pour la comparaison des résultats entre groupes. Il est également possible d'appliquer l'ACP sur l'ensemble des données concaténées, les lignes se référant aux individus des différents groupes. Cependant, dans ce cas, les variances des composantes principales mélangent à la fois la variance intra-groupe et inter-groupe. Dans cette étude, nous allons évaluer plusieurs méthodes dont l'objectif est de décrire des données structurées en plusieurs groupes. Pour chacune de ces méthodes, l'objectif est de décrire les P variables dans un espace commun, les individus étant représentés dans des espaces partiels relatifs à chaque groupe.

Nous proposons de classer ces méthodes selon différentes stratégies : (i) méthode de Flury (1984) basée sur le modèle dit 'analyse en composantes principales communes' dont les paramètres sont déterminés par maximum de vraisemblance, (ii) méthodes basées sur la détermination d'une matrice de variance-covariance commune à tous les groupes (analyse en composantes principales multi-groupes (Krzanowski, 1984), STATIS dual (Lavit et al., 1994)) et (iii) méthode consistant à déterminer de manière séquentielle les vecteurs de poids (loadings) communs (Krzanowski, 1979). Une comparaison des méthodes est effectuée sur un ensemble de 572 huiles d'olives, provenant de 9 régions d'Italie, sur la base de leur composition en huit acides gras (Forina, 1983).

2 Méthodes

Soit X une matrice de données relatives à la mesure de P variables sur N individus qui sont partitionnés en M groupes connus a priori. Chaque groupe comprend n_m individus ($\sum_{m=1}^{M} n_m = N$). Les matrices X_m pour $m = (1, \ldots, M)$ sont centrées séparément. La matrice de variance-covariance est définie par $V_m = \frac{1}{n_m} X_m^T X_m$ pour $(m = 1, \ldots, M)$. Nous cherchons à déterminer des composantes principales associées aux différents groupes mais nous imposons que ces composantes aient les mêmes vecteurs de poids. Le vecteur $a^{(h)}$ est le vecteur commun à l'ensemble des groupes, associé à la dimension $h = (1, \ldots, H)$, avec H=rang(X). La matrice A est la matrice commune de poids, telle que $A = [a^{(1)}, \ldots, a^{(H)}]$. Le vecteur $a_m^{(h)}$ est le vecteur de poids partiel associé à la dimension h et groupe m. La matrice commune de poids ,A, va servir à représenter les P variables dans un espace commun à tous les groupes. Les composantes partielles $t_m^{(h)} = X_m a^{(h)}$ vont servir à représenter les individus dans des espaces partiels relatifs à chacun des groupes.

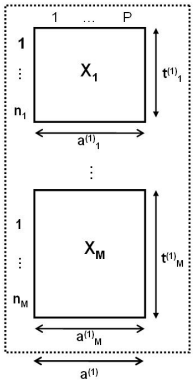


Figure 1: Illustration des vecteurs communs de poids $a^{(1)}$ et partiels $(a_1^{(1)}, \ldots, a_M^{(1)})$ et des composantes partielles $(t_1^{(1)}, \ldots, t_M^{(1)})$ donnés pour la dimension (h=1).

2.1 Analyse en composantes principales communes (Flury, 1984)

Flury (1984) a introduit une méthode, appelée analyse en composantes principales communes (ACPC), comme une généralisation de l'ACP pour le cas d'individus présentant une structure multi-groupes. Cette méthode est basée sur le calcul des matrices de variance-covariance $(V_1, \ldots, V_M)$ associées aux M groupes et la recherche des vecteurs orthogonaux communs $(a_1, \ldots, a_M)$ associés aux composantes t_m dans chaque groupe. Cela signifie que toutes les matrices V_m sont diagonalisées par une matrice commune de poids A, donnée par l'Eq. (1).

$$V_m = A\Lambda_m A^T, \qquad \text{avec } A^T A = I \text{ (où I est la matrice identité)} \tag{1}$$

avec Λ_m est une matrice diagonale des variances du groupe m. Pour calculer les paramètres de l'ACPC, sous les hypothèses de multinormalité (rarement vérifiées), Flury utilise l'estimation par maximum de vraisemblance. Le calcul des vecteurs communs de poids A est issu d'un algorithme itératif, appelé algorithme F-G (Flury, 1988). Cet algorithme est généralement long en temps de calcul ; de plus sa convergence n'est pas démontrée.

2.2 Méthodes basées sur la détermination d'une matrice de variance-covariance commune à tous les groupes

L'idée derrière ces méthodes est que, même si les groupes disposent de variance-covariance $(V_1, \ldots, V_M)$, celles-ci ont une structure commune qu'il convient de déterminer. Cette structure est la matrice A, grâce à laquelle les P variables sont représentées dans un même sous-espace.

2.2.1 Analyse en composantes principales multi-groupes

Dans l'objectif de déterminer une matrice commune de poids, A, à partir de matrices de variance-covariance $(V_1, \ldots, V_M)$, Krzanowski (1984) propose une stratégie plus simple et plus directe que celle de Flury, appelée analyse en composantes principales multi-groupes (ACPMG). En effet, l'auteur démontre que si la decomposition $V_m = A\Lambda_m A^T$ est vérifiée pour tout m, il en va

de même pour toutes les combinaisons linéaires des matrices V_m. En particulier :

$$\sum_{m=1}^{M} \frac{n_m}{N} A^T V_m A = A^T (\sum_{m=1}^{M} \frac{n_m}{N} V_m) A \qquad \text{avec } A^T A = I \tag{2}$$

Il s'ensuit que, la matrice commune de poids A est la matrice des vecteurs propres de la matrice de variance intra-groupe $V_W = \sum_{m=1}^{M} \frac{n_m}{N} V_m$. La méthode ACPMG a ainsi une résolution simple et directe. Les variances des composantes de chaque groupe sont données par $\Lambda_m = diag(A^T V_m A)$. Il est clair que V_W peut être considérée comme une matrice de variance-covariance commune issue de l'optimisation du critère (3).

$$min \sum_{m=1}^{M} n_m \parallel V_m - V_W \parallel^2 \tag{3}$$

Ainsi, la méthode ACPMG cherche à déterminer une matrice de variance-covariance commune aux matrices de variance-covariance $(V_1, \ldots, V_M)$. De ce point de vue, il est possible de déterminer d'autres stratégies d'analyse en adoptant d'autres critères.

2.2.2 STATIS dual

Comme un critère alternatif au problème posé par l'Eq. (2), nous proposons de chercher une matrice compromis V_c, solution du problème (4).

$$min \sum_{m=1}^{M} n_m \parallel V_m - \alpha_m V_c \parallel^2, \qquad \text{avec } \sum_{m=1}^{M} \alpha_m^2 = 1 \tag{4}$$

Ce problème est connu sous le nom de STATIS dual (Lavit et al., 1994). En effet, STATIS est une méthode populaire en analyse de données multibloc. Elle cherche une configuration commune à plusieurs blocs de variables mesurées sur les mêmes individus. Mais, quand les groupes se rapportent aux mêmes variables au lieu des mêmes individus, la méthode est dénommée STATIS dual. La solution du problème (4) est donnée par la matrice de variance-covariance compromis $V_c = \sum_{m=1}^{M} \alpha_m V_m$, où $\alpha = (\alpha_1, \ldots, \alpha_M)^T$ est le vecteur propre de la matrice R associé à la plus grande valeur propre, avec R mesurant la similitude entre les matrices de variance-covariance V_m dont les éléments sont $(r_{jk} = trace(V_j V_k))$ pour $(j, k = (1, \ldots, M))$. Une fois la matrice V_c déterminée, nous considérons sa décomposition spectrale sous la forme $V_c = A \Lambda A^T$. La matrice A ainsi déterminée constitue la matrice commune des vecteurs de poids. Les variances spécifiques au groupe m sont données par $\Lambda_m = diag(A^T V_m A)$.

L'intérêt de STATIS dual par rapport à la méthode précédente ACPMG est de prendre en compte les similitudes entre les matrices de variance-covariance dans des différents groupes au travers de la matrice compromis pondérée.

2.3 Méthodes consistant à déterminer de manière séquentielle les vecteurs de poids (loadings) communs

Nous proposons de chercher un vecteur de poids commun $a^{(1)}$ qui soit le plus proche possible des vecteurs $a_m^{(1)} = X_m^T t_m^{(1)}$ où $t_m^{(1)}$ est une composante associée à X_m. De manière plus précise nous cherchons à maximiser le critère (5).

$$\sum_{m=1}^{M} n_m < a_m^{(1)}, a^{(1)} >^2 = \sum_{m=1}^{M} n_m ((a^{(1)})^T X_m^T t_m^{(1)})^2, \qquad \text{avec } \parallel a_m^{(1)} \parallel = 1 \text{ et } \parallel a^{(1)} \parallel = 1 \tag{5}$$

Nous montrons que la solution conduit à une analyse très proche de la méthode introduite par Karzanowski (1979) sous l'appellation "comparaisons entre les groupes".

2.4 Critères de comparaison des méthodes

La comparaison des méthodes est basée sur deux critères. Le premier reflète la similitude entre les vecteurs communs de poids sur $(1, \ldots, h)$ dimensions.

$$S^{(h)} = \frac{1}{h} \sum_{r=1}^{h} |(a^{(r)})^T a^{*(r)}| \tag{6}$$

où a et a^* sont les vecteurs communs de poids issus des deux méthodes comparées. Le deuxième critère est basé sur l'inertie de chaque groupe restituée par une dimension h donnée.

$$Inertie_m^{(h)} = \frac{\lambda_m^{(h)}}{trace(V_m)} \tag{7}$$

3 Application

L'application est réalisée sur un ensemble d'échantillons de 572 huiles d'olive provenant de neuf régions différentes d'Italie (Forina, 1983). Pour chaque échantillon, les mêmes huit variables correspondant aux teneurs en acides gras sont mesurées. La comparaison des méthodes est basée sur les deux critères présentés dans le paragraphe précédent. Les résultats de similitude des vecteurs communs de poids A des méthodes sont donnés par la Table 1.

Les inerties restituées du groupe 1 (région des Pouilles du nord) et du groupe 2 (région de Calabre) sont représentées respectivement sur les Figures 2(a) et 2(b).

Nous constatons que toutes les méthodes conduisent à des résultats très proches.

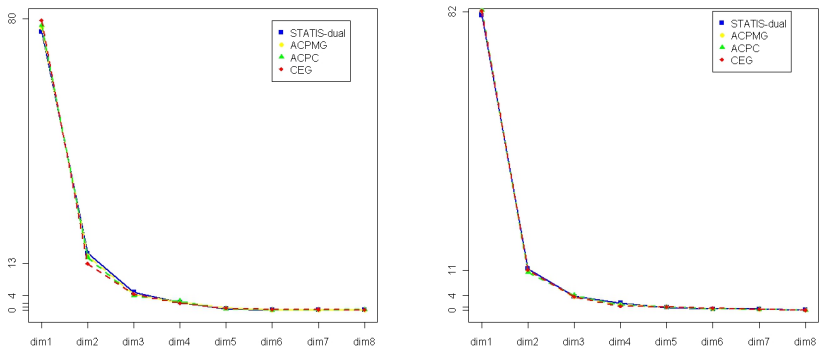
	STATIS dual	ACPMG	ACPC	CEG
STATIS dual	1.000			
ACPMG	0.999	1.000		
ACPC	0.997	1.000	1.000	
CEG	0.967	0.976	0.980	1.000

	STATIS dual	ACPMG	ACPC	CEG
STATIS dual	1.000			
ACPMG	0.999	1.000		
ACPC	0.996	0.999	1.000	
CEG	0.960	0.965	0.966	1.000

Table 1: Matrices des similarités $S^{(h)}$ pour les dimensions $h = 1$ (à gauche) et $h = 2$ (à droite). Abréviations : analyse en composantes principales multi-groupes (ACPMG), analyse en composantes principales communes de Flury (ACPC) et comparaisons entre les groupes (CEG)

4 Conclusion et perspectives

L'étude présentée se situe dans le cadre des données multivariées présentant une structure hiérarchisée, appelées aussi structure multi-groupes. Nous proposons une vision globale d'une sélection de méthodes et les comparons selon les critères de similitude des vecteurs communs de poids et selon les pourcentages d'inertie de chaque groupe. Malgré des différences évidentes de calculs des différents paramètres, il existe une similitude entres les résultats des différentes méthodes.



(a) Groupe 1 (région des Pouilles du nord). (b) Groupe 2 (région de Calabre).

Figure 2: Pourcentage d'inertie des vecteurs communs de poids A, selon les dimension $h = (1, \ldots, H)$ pour les groupes 1 et 2.

Une version plus complète de ces comparaisons est actuellement développée sur d'autres méthodes (analyse Procustéenne généralisée duale, analyse factorielle multiple duale et analyse en composantes communes et poids spécifiques duale) et sur des critères supplémentaires.

La vision globale des méthodes multi-groupes $[X_1|\ldots|X_M]^T$, comme des extensions des méthodes multibloc $[X_1|\ldots|X_K]$ conduit à une unification des méthodes d'analyse de données. Cette vision globale va être utilisée pour l'analyse de la relation entre deux tableaux structurés en groupes $[X_1|\ldots|X_M]$ et $[Y_1|\ldots|Y_M]$.

Bibliographie

Flury, B. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79, 892-898.

Forina, M., Armanino, C., Lanteri, S. & Tiscornia, E. (1983). *Classification of Olive Oils from their Fatty Acid Composition*. In H. Martens, H.Jr Russwurm Eds. Food Research and Data Analysis (pp. 189-214). Applied Science Pub., London.

Krzanowski, J. (1984). Principal Component Analysis in the Presence of Group Structure. *Applied Statistics*, 33(2), 164-168.

Krzanowski, J. (1979). Between-groups comparison of principal components. *Journal of the American statistical Association*, 74, 703-707.

Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics and Data Analysis*, 18, 97-117.

Session 9 : Maîtrise des Procédés II/ *Process Control II*

Une approche particulaire de l'identification et de l'inférence statistique de modèle en microbiologie prévisionnelle

A particle approach of model identification and inference in predictive microbiology

Jean-Pierre Gauchi[1], Jean-Pierre Vila[2], Caroline Bidot[1], Jean-Christophe Augustin[3], Louis Coroller[4] & Pierre Del Moral[5]

[1] *INRA-MIA (UR341), Jouy-en-Josas*
E-mail : *jean-pierre.gauchi@jouy.inra.fr, caroline.bidot@jouy.inra.fr*

[2] *INRA-MIA (UMR729), Montpellier,*
E-mail : *jean-pierre.vila@supagro.inra.fr*

[3] *ENV-Alfort Unité MASQ, Maison-Alfort*
E-mail : *jcaugustin@vet-alfort.fr*

[4] *Univ. Bretagne Occ. LUBEM, Quimper*
E-mail : *louis.coroller@univ-brest.fr*

[5] *INRIA-Alea, Bordeaux*
E-mail : *pierre.del-moral@inria.fr*

Résumé

Les systèmes dynamiques microbiens de type alimentaire se caractérisent par leur complexité structurelle (écosystèmes) mais aussi par une complexité d'ordre opérationnelle quand on cherche à les identifier au travers de modèles stochastiques non linéaires à espace d'état. L'observation indirecte de leur évolution par séries de prélèvements, dilutions et dénombrements en milieu de culture, rend statistiquement problématique l'estimation de leurs paramètres par les méthodes classiques (maximum de vraisemblance) ainsi que toute opération d'inférence sur les paramètres ou sur les modèles eux-mêmes (sélection).

Dans ce projet, une récente approche bayésienne par filtrage particulaire, adaptée aux systèmes à observation indirecte, a donc été mise en oeuvre : elle conduit à des estimations convergentes des densités de probabilités a posteriori des effectifs microbiens au long de l'évolution du système et des densités des paramètres du modèle postulé. Elle permet également la restauration des méthodes d'inférence et de comparaison de modèles basées sur les fonctions de vraisemblance. Sa combinaison avec l'optimisation de certains indices de sensibilité paramétrique, conduit naturellement à la détermination des protocoles d'échantillonnage temporel les plus favorables à l'estimation des paramètres du modèle.

L'ensemble de ces approches particulaires a été introduit au sein de la plate-forme logicielle FILTREX pour des systèmes essentiellement mono-spécifiques, à destination de la communauté microbiologiste.

Mots-clés : systèmes microbiologiques ; analyse de risque ; système à espace d'état ; filtrage particulaire à convolution ; méthodes SMC ; Facteur de Bayes ; coefficients de sensibilité ; échantillonnage optimal.

Abstract
Food microbial dynamic systems can be characterized by their structural complexity as ecosystems, but also by the complexity of their statistical study when modelized by stochastic nonlinear state-space models. As the evolution of these systems can only be observed by repeated series of samplings, dilutions and counts, parameter estimation and statistical inference can hardly be performed by the usual approaches (e.g. maximum likelihood). To this aim, a Bayesian approach by particle filtering recently developed, has been brought into play: it leads to convergent estimates of the posterior probability density functions of the bacteria numbers all along the system evolution and to convergent estimates of the posterior of the model parameters themselves. Moreover this approach allows restoring inference and model comparison likelihood-based methods, and when combined with the maximization of some parameter sensitivity indices it allows sampling designs for improved parameter estimation to be built. All these facilities have been included in the software FILTREX for mono-species systems, intended for the microbiologist communauty.
Keywords: microbiological systems, risk analysis, state-space system; convolution particle filtering; SMC methods; Bayes Factor; sensitivity coefficients, optimal sampling.

1 Introduction

L'ambition de ce projet qui rassemble des équipes de statisticiens et de microbiologistes, est d'apporter des solutions à des problèmes cruciaux rencontrés en microbiologie prévisionnelle, par le développement ou l'adaptation d'approches stochastiques, statistiques et numériques innovantes, pour l'identification, l'inférence et le contrôle de systèmes dynamiques microbiens de type agro-alimentaire (*Listeria, Clostridium,...*). En effet dans ce domaine, les microbiologistes font face à des difficultés importantes d'accès à des informations nécessaires à l'étude et à la maîtrise de l'évolution des systèmes microbiens pathogènes : effectifs réels des bactéries dans les substrats considérés (aliments, milieux de culture) d'une part, valeurs de paramètres cruciaux (vitesses maximales de croissances, temps de latences, températures et pH optima, etc) caractérisant les sensibilités bactériennes vis-à-vis des conditions environnementales (croissance, non-croissance, inactivation), d'autre part. La complexité de ces systèmes bactériens est d'abord de nature structurelle : il s'agit en fait d'éco-systèmes dans lesquels peuvent coexister un grand nombre d'espèces bactériennes en interaction. Mais la complexité de ces systèmes est aussi de nature fonctionnelle, conséquence de la distribution de leur dynamique selon au moins trois niveaux hiérarchiques que l'on peut approximativement résumer par :

- un niveau externe, seul accessible aux mesures (dénombrements en milieu de culture après prélèvements et séries de dilutions,...), lesquelles sont actuellement réalisées avec de grandes incertitudes.
- un niveau intermédiaire, correspondant aux dynamiques de croissances ou décroissances bactériennes proprement dites, non mesurables dans le substrat considéré, mais a priori modélisables sous forme de modèles stochastiques dits *primaires*. S'il est vrai que l'on dispose de modèles relativement efficaces pour des systèmes mono-spécifiques (e.g. modèles de Baranyi-Roberts (1995), de Rosso (1995)), la prise en compte simultanée de plusieurs espèces reste aujourd'hui encore principalement qualitative.

- un niveau plus profond, caractérisé par des variables cinétiques qui conditionnent les dynamiques précédentes, et elles-mêmes résultats d'interactions de différents facteurs biotiques et abiotiques (conditions de milieu). Ces éléments cinétiques peuvent parfois être modélisables sous forme de modèles stochastiques dits secondaires portant sur les paramètres du modèle primaire.

Cette double complexité rend toute modélisation prédictive de ces dynamiques bactériennes par les approches classiques (e.g. moindres carrés non linéaires), particulièrement difficile sinon impossible, et donc illusoire toute opération fiable de détection, de prévision d'évolution critique et a fortiori de contrôle. Ces systèmes ne pouvant être observés qu'indirectement (systèmes à espace d'état non linéaires), nous avons développé ou adapté de nouvelles approches de type Monte-Carlo séquentiel (SMC) pour les opérations d'estimation d'effectifs et de paramètres (filtrage particulaire à convolution), de prédiction d'évolutions (densités de probabilités prédictives), d'inférence paramétrique et fonctionnelle (bandes de confiance, estimation de facteurs de Bayes) et d'échantillonnage optimal (cf. Gauchi & Vila 2011, Vila 2011a). Ces approches ont été ou sont en cours d'être intégrées dans la plate-forme FILTREX (Gauchi et al. 2011). D'autres sont en prévision : détection de ruptures de modèle (dysfonctionnements) par généralisation particulaire de tests séquentiels CUSUM, contrôle prédictif d'évolution, caractérisation et analyse d'évènements rares (Del Moral 2004).

2 Un cadre de modélisation adapté

S'agissant de systèmes dynamiques observés indirectement, la représentation de l'évolution des systèmes microbiens par des chaînes de Markov cachées à temps discret de type modèles à espace d'état non linéaires, est particulièrement bien adaptée

$$\begin{cases} x_t &= f_t(x_{t-1}, \theta, \varepsilon_t) \\ y_t &\sim g_t(.|x_t, \theta) \end{cases} \tag{1}$$

où x_t et y_t représentent respectivement les vecteurs des différents effectifs bactériens dans le milieu d'origine à l'instant t et les vecteurs des comptages correspondants effectués en milieu de culture après série de prélèvements-dilutions. $\theta \in \Theta \subset I\!R^p$ est le vecteur des p paramètres du modèle. ε_t est un vecteur de variables aléatoires indépendantes (éventuellement des bruits). g_t est une distribution de probabilité, loi des comptages conditionnellement aux effectifs dans le milieu d'origine. Le modèle auto-régressif f_t est supposé connu. La loi $\mathcal{L}_{\varepsilon_t}$ de ε_t et la distribution g_t ne sont pas forcément connues mais supposées simulables.
Cette modélisation générale recouvre tous les modèles classiques d'évolution bactérienne. Dans ce qui suit les approches proposées seront illustrées à partir d'un modèle de croissance bien connu.

2.1 Un modèle de croissance bactérienne

Considérons le modèle de Baranyi-Roberts-2 (1995), très utilisé pour la formalisation de la croissance de bactéries pathogènes mono-spécifiques. Sa forme discrète peut s'écrire

$$x_{t+1} \;=\; \delta x_0 \exp(\mu_{max} A_t) \frac{1}{B_t} (\mu_{max} \frac{dA_t}{dt} - \frac{dB_t}{dt} \frac{1}{B_t}) \;+\; x_t \;+\; \varepsilon_t \tag{2}$$

avec $A_t = t + \frac{1}{\mu_{max}} \ln(\exp(-\mu_{max} t) + \exp(-\mu_{max}\lambda) - exp(-\mu_{max} t - \mu_{max}\lambda))$
et $B_t = 1 + \frac{\exp(\mu_{max} A_t) - 1}{\frac{x_{max}}{x_0}}$
où :

- μ_{max} (taux de croissance maximum), λ (temps de latence moyen), x_0 (nombre de bactéries en début de croissance), x_{max} (nombre maximum de bactéries en fin de croissance) : soit 4 paramètres à estimer.
- ε_t est une variable de Poisson centrée et δ le pas de discrétisation du modèle différentiel initial de Baranyi-Roberts.

$y_t \sim g_t(.|x_t, \theta)$ est la variable d'observation : comptage de bactéries formant colonies (UFC) en boîte de Petri, à partir d'un prélèvement effectué au temps t. La distribution de probabilité g_t est le résultat de l'interaction de plusieurs phénomènes aléatoires : l'échantillonnage spatial dans le milieu primaire au temps t, les dilutions et échantillonnages successifs (sous répartition de Poisson dans le cas le plus simple), les erreurs successives de prélèvements volumiques (qu'on peut supposer gaussiennes) et enfin les erreurs de comptage des UFC en boîtes de Petri (qu'on pourra supposer lognormales). La complexité de la distribution g_t empêche sa caractérisation analytique mais autorise malgré tout sa simulation.

3 Identification du modèle par filtrage particulaire à convolution

Cette non accessibilité analytique de la distribution g_t ainsi que la structuration du modèle (1)-(2) empêchent l'utilisation des méthodes de moindres carrés ou de maximum de vraisemblance pour l'estimation des paramètres du modèle. Le filtrage particulaire à convolution permet de contourner ces problèmes. Rappelons que le but du filtrage statistique est l'estimation des densités a posteriori $p_t(x_t|y_1, \ldots, y_t)$ et $p_t(\theta|y_1, \ldots, y_t)$ à chaque instant t. Le filtrage à convolution (Rossi et Vila 2006, Campillo et Rossi 2009, Vila 2011a, 2011b) est une approche non paramétrique récente du filtrage particulaire, qui repose sur des hypothèses plus faibles que celles des méthodes particulaires SMC plus classiques (Doucet et al. 2001, Del Moral 2004), notamment la non exigence de l'accessibilité des fonctions de vraisemblance. Nous limiterons ici la présentation aux principes généraux d'un des filtres à convolution de particules développés, le filtre R-CF (cf. Rossi et Vila 2006).

3.1 Algorithme

A chaque instant t, n réalisations du système à espace d'état (1) appelées particules, sont simulées : $(x_t^i, \theta_t^i, y_t^i),\ i = 1, \ldots, n$, selon les deux étapes classiques du filtrage :

- $t = 0$: pour $i = 1, \ldots, n$, soit $\bar{x}_0^i \sim p_0^x$, $\bar{\theta}_0^i \sim p_0^\theta$, $\varepsilon_1^i \sim \mathcal{L}_{\varepsilon_1}$, $t = t + 1$.
- $t > 0$:
  - étape de prédiction : simulation de n particules, pour $i = 1, \ldots, n$
    si $t = 1$: soit $x_1^i = f_1(\bar{x}_0^i, \bar{\theta}_0^i, \varepsilon_1^i)$, $\theta_1^i = \bar{\theta}_0^i$, $y_1^i \sim g_1(.|x_1^i, \theta_1^i)$.
    si $t > 1$: soit $(\bar{x}_{t-1}^i, \bar{\theta}_{t-1}^i) \sim p_{t-1}^n(x, \theta|y_{1:t-1})$, $\varepsilon_t^i \sim \mathcal{L}_{\varepsilon_t}$,
    et $x_t^i = f_t(\bar{x}_{t-1}^i, \bar{\theta}_{t-1}^i, \varepsilon_t^i)$, $\theta_t^i = \bar{\theta}_{t-1}^i$, $y_t^i \sim g_t(.|x_t^i, \theta_t^i)$.
  - étape de correction (approximation de la formule de Bayes) :
    estimation des densitées conditionnelles et des espérances a posteriori de x_t et θ_t

$$
\begin{aligned}
p_t^n(x,\theta|y_{1:t}) &= \frac{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)\times K_{h_n}^\theta(\theta-\theta_t^i)\times K_{h_n}^x(x-x_t^i)}{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)}\\
p_t^n(\theta|y_{1:t}) &= \frac{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)\times K_{h_n}^\theta(\theta-\theta_t^i)}{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)}\\
p_t^n(x|y_{1:t}) &= \frac{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)\times K_{h_n}^x(x-x_t^i)}{\sum_{i=1}^n K_{h_n}^y(y_t-y_t^i)}
\end{aligned}
$$

$$
\hat{x}_t^n = \frac{1}{n}\sum_{i=1}^n \bar{x}_t^i \quad \text{et} \quad \hat{\theta}_t^n = \frac{1}{n}\sum_{i=1}^n \bar{\theta}_t^i \tag{3}
$$

- $t = t+1$

- Retour à l'étape t.

$K_{h_n}^y(.), K_{h_n}^x(.)$ et $K_{h_n}^\theta$ sont des fonctions noyaux (Parzen 1962) de dimensions égales aux dimensions respectives de y, x et θ, (*e.g.* noyaux gaussiens) et h_n le paramètre de fenêtre correspondant (que l'on peut supposer commun).

Notation : $y_{1:t} := y_1, y_2, \ldots, y_t$.

3.2 Propriétés de convergence

Sous des conditions très générales de bornitudes pour les densités $p_t(y|x_t,\theta)$, $p_t(x|x_{t-1},\theta)$, $p_t(\theta|y_{1:t})$ et pour les variances $\mathrm{Var}[x_t,\theta_t|y_{1:t}]$, et de décroissance pour le paramètre de fenêtre h_n quand n tend vers l'infini, ont été démontrées (Rossi et Vila 2006, Vila 2011b) :

- la convergence L_1 presque sûre et la convergence ponctuelle presque sûre des estimateurs de densité a posteriori $p_t^n(x,\theta|y_{1:t})$, $p_t^n(x|y_{1:t})$ et $p_t^n(\theta|y_{1:t})$, vers les vraies densités $p_t(x,\theta|y_{1:t})$, $p_t(x|y_{1:t})$ et $p_t(\theta|y_{1:t})$, pour tout t fini lorsque le nombre de particules n tend vers l'infini,
- la convergence dans les mêmes conditions de $\hat{x}_t^n$ vers $\mathrm{E}[x_t|y_{1:t}]$,
- la convergence de $\hat{\theta}_t^n$ vers θ^* vraies valeurs des paramètres, lorque n et t tendent vers l'infini.

3.3 Application au filtrage du modèle de Baranyi-Roberts

Des UFC de *Listeria-monocytogenes* ont été obtenues et dénombrées en milieu de culture, par l'Unité MASQ de l'ENV-Alfort, selon le protocole suivant :

- 10 instants de prélèvement (heures) : 0, 72, 120, 168, 240, 264, 288, 336, 408, 504.
- 3 prélèvements à chacun de ces instants, conduisant à 3 comptages en boîtes de Petri associés à chacun de ces instants.
- 5 facteurs de dilution différents successifs, selon l'instant de prélèvement (pour garantir un nombre raisonnable d'UFC à dénombrer après mise en culture pour chaque prélèvement).
- lois a priori uniformes pour 4 les paramètres, de supports respectifs :
  $\mu_{max} \in [0.01, 2]$, $\lambda \in [10, 70]$, $x_0 \in [100, 400]$, $x_{max} \in [10^8, 10^9]$.
- pas de discrétisation $\delta = 2$heures.
- nombre de particules pour le filtrage : $n = 10^5$.

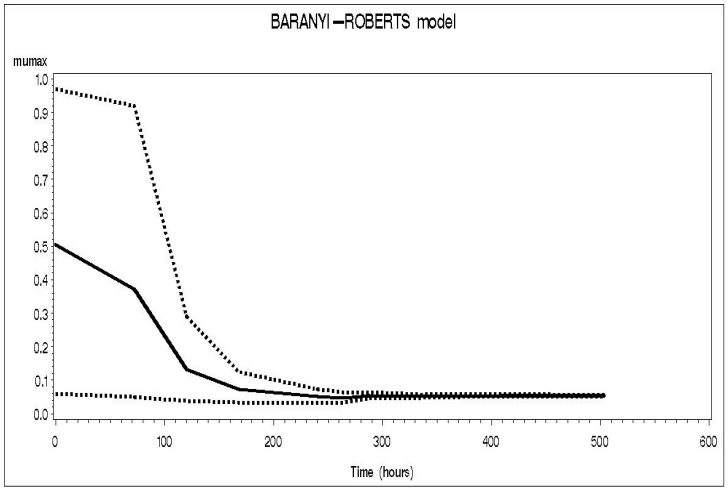


Figure 1: Estimations successives du paramètre μ_{max} par le filtre R-CF, en chacun des 10 instants de prélèvement et bornes de confiance inférieure et supérieure à 95%.

Les estimations successives (moyennes de particules selon (3)) du paramètre μ_{max} par le filtre R-CF, pour chacun des 10 instants de prélèvement, sont représentées dans la Figure 1, avec les bornes de confiance à 95% correspondantes. La Figure 2 représente en échelle logarithmique la courbe de Baranyi après estimation de ses quatre paramètres par filtrage R-CF jusqu'au dernier instant de prélèvement ($t = 504$). Les (*) correspondent aux extrapolations au volume primaire des comptages d'UFC en boîte de Petri (3 répétitions pour chaque instant de prélèvement).

4 Détermination d'instants de prélèvements optimaux

Deux approches ont été développées, qui toutes deux reposent sur la maximisation séquentielle de coefficients de sensibilité paramétrique, de manière à favoriser la concentration des densités conditionnelles $p_t(\theta|y_0, \ldots, y_t)$ autour de leurs modes (Gauchi et Vila 2011).

1. *Approche par maximisation d'estimations particulaires de coefficients de sensibilité globaux:*

   Soit H le nombre de prélèvements autorisé sur un intervalle d'observation $[0, T]$.
   Soit $J(t)$ une fonction de sensibilité définie dans la suite. Principe :

   A partir de l'instant $t = 0$

   - Etape 1 :
     Soit $(t_1^*, t_2^*, \ldots, t_H^*) = \arg\max_{\{t<t_1<t_2<\ldots<t_{H-1}<t_H<T\}} C(t_1, t_2, \ldots, t_{H-1}, t_H)$
     avec $C(t_1, t_2, \ldots, t_{H-1}, t_H) = \sum_{j=1}^{j=H} J(t_j)$.
     $t \leftarrow t_1^*$
     $H \leftarrow H - 1$
   - Etape 2
     Effectuer le nouveau comptage en $t = t_1^*$: y_t
     Tant que $H > 0$ retour à l'Etape 1.

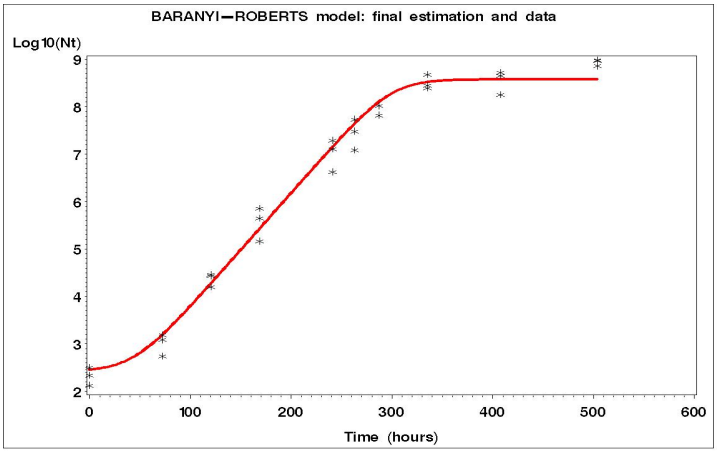


Figure 2: Courbe de Baranyi-Roberts calculée après les estimations de ses paramètres par filtrage R-CF jusqu'au dernier instant de prélèvement.

Choix de J : selon l'approche d'Analyse de Sensibilité Globale (Sobol' 2001, Rodriguez-Fernandez et al. 2007), $J(s) = Q(s)^T Q(s)$ avec $Q(s) = \Big(S_1(s), S_2(s), \ldots, S_j(s), \ldots, S_p(s)\Big)$ et $S_j(s) = \mathrm{Var}(\mathrm{E}[y_s|\theta_{j,s}])/\mathrm{Var}(y_s)$. Pour $s = t+k$, $1 < k \leq T-t$, les deux termes du rapport $S_j(s)$ peuvent être estimés de façon consistente à partir des n particules $\bar{\theta}_t^i$ et de séries de simulations de $x_{t+k}^i = f_{t+k}(x_{t+k-1}^i, \bar{\theta}_t^i, \varepsilon_{t+k}^i)$ et $y_{t+k}^i \sim g_{t+k}(.|x_{t+k}^i, \bar{\theta}_t^i)$.

2. *Approche par maximisation d'estimation de coefficients de sensibilité individuels :*

   Cette approche numériquement moins coûteuse que la précédente, repose sur les coefficients de sensibilité T*d*SI-VIP (SIV en abrégé) récemment développés (Ellouze et al. 2010, Gauchi et al. 2010). Dès l'enregistrement d'une observation, les SIV, exprimés en pourcentage, sont systématiquement calculés pour chacun des p paramètres, pour tous les temps de mesures possibles futurs, conduisant à un faisceau de p courbes de sensibilité. Le prochain temps de prélèvement est le premier temps en lequel une des p courbes de sensibilité passe par un maximum. Après prélèvement en cet instant et dénombrement bactérien correspondant, un nouveau faisceau de courbes SIV est calculé et l'opération précédente est répétée. Cette démarche est illustrée dans le paragraphe suivant.

4.1 Application à l'optimisation du filtrage du modèle de Baranyi-Roberts

Une croissance bactérienne a été simulée selon le modèle (2) avec les valeurs de paramètres :
$x_0 = 200,\ x_{max} = 5 \times 10^8,\ \mu_{max} = 0.05,\ \lambda = 50,$
et avec des lois a priori uniformes de supports :
$x_0 \in [140, 260],\ x_{max} \in [0.5 \times 10^8, 10 \times 10^8],\ \mu_{max} \in [0.01, 0.09],\ \lambda \in [10, 110].$
Des faisceaux de courbes SIV ont été successivement calculés selon l'approche précédente. La Figure 3 représente le faisceau ainsi obtenu en $t = 0$. Pour un nombre de prélèvements fixé à 10, les temps optimaux suivants ont été obtenus : $t^* = 2, 69, 97, 125, 140, 160, 176, 263, 384, 504$. La Figure 4 représente la courbe de Baranyi-Roberts ajustée par filtrage à partir de cet échantillonnage optimal simulé, avec 3 prélèvements en chaque temps optimal (*).

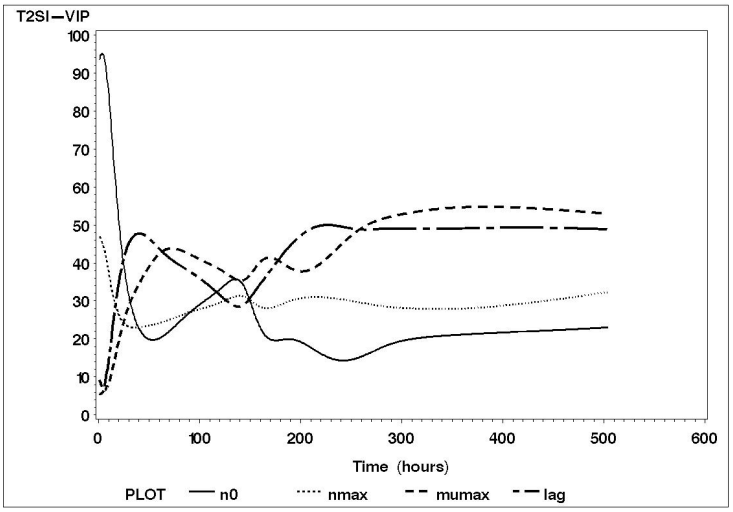


Figure 3: Courbes SIV pour chacun des quatre paramètres du modèle de Baranyi-Roberts, calculées en t=0, pour les temps futurs de t=1 à t=504.

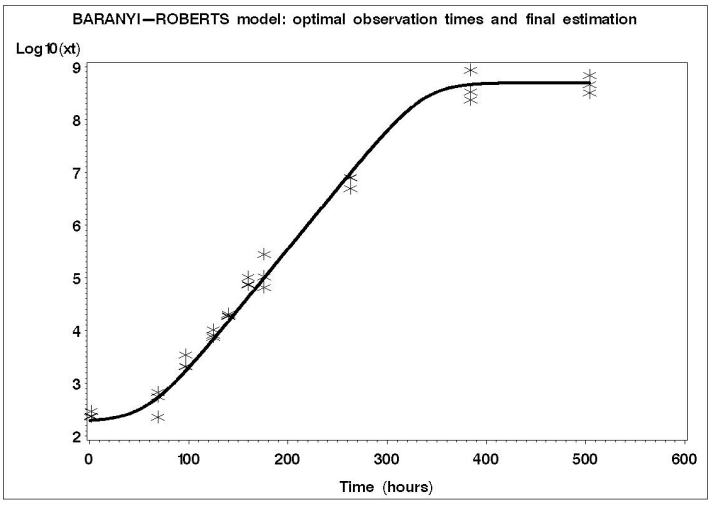


Figure 4: Courbe de Baranyi-Roberts obtenue à partir du filtrage sur les 10 instants d'échantillonnage optimal.

On pourra remarquer l'accumulation prioritaire des instants optimaux ainsi déterminés dans la première partie de la cinétique microbienne, ce qui a pour effet de favoriser l'estimation des paramètres x_0, λ et μ_{max}. On notera également la disposition judicieuse des instants les plus lointains, qui favorisent l'estimation du quatrième paramètre d'asymptote x_{max}.

A partir de ces dix prélèvements optimaux, un filtrage à convolution avec $n = 10^5$ particules a fourni les estimations des paramètres et de leurs intervalles de confiance à 95% présentées en Table 1.

paramètre	x_0	x_{max}	μ_{max}	λ
estimation	198	5.6×10^8	0.051	51.6
borne inf.	193	5.3×10^8	0.049	48.6
borne sup.	204	5.8×10^8	0.053	54.6

Table 1

Ces estimations sont très proches des valeurs correspondantes de simulation. Elles sont globalement meilleures que celles obtenues à partir de l'échantillonnage introduit au § 3.3, données en Table 2, et meilleures également que les estimations obtenues à partir d'un échantillonnage naïf équidistribué entre $t = 0$ et $t = 504$, données en Table 3. De plus on pourra remarquer que dans les Tables 2 et 3 les intervalles de confiance estimés pour μ_{max} et λ ne recouvrent même pas les vraies valeurs (0.05 et 50).

paramètre	x_0	x_{max}	μ_{max}	λ
estimation	196	5.85×10^8	0.054	65.1
borne inf.	191	5.7×10^8	0.053	61.7
borne sup.	200	6.0×10^8	0.055	68.5

Table 2

paramètre	x_0	x_{max}	μ_{max}	λ
estimation	195	5.29×10^8	0.055	64.0
borne inf.	191	5.11×10^8	0.053	59.6
borne sup.	199	5.47×10^8	0.056	68.4

Table 3

5 Sélection de modèle par estimation particulaire de Facteur de Bayes

Le facteur de Bayes (Jeffrey, 1961) est un des outils les plus efficaces pour discriminer entre deux modèles concurrents. Son estimation, souvent délicate, est réalisée par des procédures de type Chaînes de Markov de Monte-Carlo (MCMC). Celles-ci nécessitent la disponibilité des densités de probabilités des réponses des modèles. Une application du filtre à convolution permet de contourner cet inconvénient dans le cas des modèles à espace d'état, comme les modèles de dynamiques microbiennes considérés ici. De plus cette approche particulaire fournit un estimateur convergent du facteur de Bayes (Vila et Saley 2009).

5.1 Construction d'un facteur de Bayes

Pour deux modèles M_1 et M_2, d'espaces de paramètres respectifs Θ_1 et Θ_2, pour une série d'observations $Y = y_1, \ldots, y_T$, le facteur de Bayes s'écrit $B_{12} = p_1(Y)/p_2(Y)$ avec $p_i(Y) = \int_{\Theta_i} p_i(Y|\theta)p_i(\theta)d\theta,\ i = 1, 2$, où $p_i(Y)$ est la vraisemblance marginale du modèle M_i où encore la constante de normalisation de la loi a posteriori $p_i(\theta|Y)$. M_1 sera préféré à M_2 si $B_{12} >> 1$.

5.2 Estimation particulaire d'un facteur de Bayes

Soit $p_j^n(y_{t+1}|y_{1:t}) = \frac{1}{n}\sum_{i=1}^n K_{h_n}^y(y_{t+1} - y_{t+1}^{i,j})$ où $\{y_{t+1}^{i,j}\}$ sont les n particules générées par l'algorithme de filtrage à convolution du § 3.1 pour le modèle $M_j, j = 1, 2$.
Soit $p_j^n(Y) = p_j^n(y_1)\Pi_{t=1}^{t=T-1} p_j^n(y_{t+1}|y_{1:t})$ l'estimation particulaire de la vraisemblance marginale du modèle M_j.

Alors, sous les conditions de fenêtre et de bornitudes évoquées au §3.2, quand n tend vers l'infini $B_{12}^n = p_1^n(Y)/p_2^n(Y)$ est un estimateur convergent du facteur de Bayes entre les modèles M_1 et M_2 (Vila et Saley 2009).

5.3 Application : comparaison entre les modèles de Baranyi et de Rosso

Le modèle de Baranyi est souvent confronté à un modèle concurrent, celui de Rosso (Rosso, 1995). Pour les données de comptages déjà présentées au §3.3, la Table 4 présente les estimations successives du facteur de Bayes entre ces deux modèles, obtenues au fur et à mesure de l'évolution de la croissance microbienne. La stabilité relative de ces estimations, légèrement inférieures à 1, confirme l'équivalence des deux modèles, plus qu'une supériorité même légère du modèle de Rosso sur le modèle de Baranyi, pour cette expérimentation.

t	0	72	120	168	240	264	288	336	408	504
$B^n_{Baranyi/Rosso}(t)$	0.934	0.938	0.941	0.944	0.944	0.925	0.889	0.922	0.591	0.817

Table 4

6 Conclusion

Pour permettre une identification efficace des modèles de dynamiques bactériennes observées indirectement, une approche par filtrage particulaire à convolution est proposée. Les estimations de paramètres ainsi réalisées, peuvent être optimisées par la détermination de plans d'échantillonnage optimaux calculés par maximisation de coefficients de sensibilité. Cette approche permet de plus de restaurer les méthodes d'inférence basées sur les vraisemblances, inaccessibles ici, comme l'estimation de facteurs de Bayes pour la comparaison entre deux modèles, ou les tests séquentiels de type CUSUM pour la détection de rupture de modèle. La plate-forme logicielle FILTREX, orientée vers la communauté microbiologiste, a été conçue pour offrir toutes ces fonctionnalités et s'enrichit régulièrement de fonctionnalités nouvelles.

Bibliographie

Baranyi, J., Roberts, T. A. (1995). Mathematics of predictive food microbiology. *Int. Journal of Food Microbiology*, **26**, 199-218.

Campillo, F., Rossi, V. (2009). Convolution particle filter for parameter estimation in general state-space models. *IEEE Trans. Aero. Elec. Syst.* **45**, 1063-1072.

Del Moral, P. (2004). Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications, Springer, New York.

Doucet, A., de Freitas, N., Gordon, N. (2001). Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science, Springer, New York.

Ellouze, M., Gauchi, J.P., Augustin, J.C. (2010). Global sensitivity analysis applied to a contamination assessment model of Listeria monocytogenes in cold smoked salmon at consumption. *Risk Analysis*, **30**, 5, 841-852.

Gauchi, J.P., Lehuta, S., Mahévas, S., (2010). Constrained global sensitivity analysis: partial least squares regression metamodelling and D-optimal computer experiment design. RESS, *en soumission.*

Gauchi, J.P., Vila, J.P. (2011). Optimal sequential sampling design for improving parametric identification of complex microbiological dynamic systems by nonlinear filtering. *7^{th} International Conference on Predictive Modelling in Foods*, Dublin, Ireland.

Gauchi, J.P., Vila, J.P., Bidot, C., Atlijani, E., Coroller, L., Augustin, J.C., Del Moral, P. (2011). FILTREX: a new software for identification and optimal sampling of experiments for complex microbiological dynamic systems by nonlinear filtering. *7^{th} International Conference on Predictive Modelling in Foods*, Dublin, Ireland.

Jeffreys, H. (1961). Theory of Probability, Oxford University Press, Oxford.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065-1076.

Rodriguez-Fernandez, M., Kucherenko, S., Pantelides, C., Shah, N. (2007). Optimal experimental design based on global sensitivity analysis. *17^{th} European Symposium on Computer Aided Process Engineering.*

Rossi, V. and Vila, J.P. (2006). Nonlinear filtering in discrete time: A particle convolution approach. *Pub. Inst. Stat. Univ. Paris*, L, **3**, 71-102.

Rosso, L. (1995). Modélisation et microbiologie prévisionnelle: élaboration d'un nouvel outil pour l'agro-alimentaire. Thèse de Doctorat en Science, Université Claude Bernard - Lyon I.

Sobol', I. (2001). Global sensitivy indices for nonlinear mathematical models and their Monte-Carlo estimates. *Mathematical Computers in Simulation*, **55**, 271-280.

Vila, J.P. (2011a). Nonparametric multi-step prediction in nonlinear state space dynamic systems. *Statistics and Probability Letters*, **81**, 71-76.

Vila, J.P. (2011b). Enhanced consistency of the Resampled Convolution Filter. *en révision.*

Vila, J.P., Saley, I. (2009). Estimation de facteurs de Bayes entre modèles dynamiques non linéaires à espace d'état. *C.R. Acad. Sci. Paris, Ser. I,* **347**, 429-434.

Comment accepter l'équivalence entre deux méthodes de mesure? Comparaison et améliorations de l'approche de Bland et Altman et des régressions avec erreurs sur les variables

How to accept the equivalence of two measurement methods? Comparison and improvements of the Bland and Altman's approach and errors-in-variables regressions

Bernard Francq[1] & Bernadette Govaerts

[1] *Institut de Statistique, Biostatistique et sciences Actuarielles*
ISBA – IMMAQ – UCL (Université Catholique de Louvain)
Voie du Roman Pays, 20
1348 Louvain-la-Neuve (Belgium)
E-mail : *bernard.g.francq@uclouvain.be*

Résumé
Les besoins des industries d'évaluer rapidement la qualité de leurs produits ou leur performance conduit au développement et améliorations de méthodes alternatives de mesure. Ces méthodes doivent idéalement donner des résultats comparables ou équivalents à une méthode standard.
Premièrement, les mesures données par deux méthodes peuvent être modélisées par une régression avec erreur sur les variables. Les paramètres estimés sont utiles afin de vérifier l'équivalence en testant la présence d'un biais entre les méthodes. Deuxièmement, les différences entre paires de mesures peuvent être analysées pour évaluer la concordance entre les méthodes à l'aide des limites d'agrément données par l'approche très connue et très utilisée de Bland et Altman.
Nous comparerons ces deux méthodologies et proposerons des améliorations telles que les intervalles de tolérance et régressions avec erreurs corrélées sur les variables.
Mots-clés : régressions avec erreurs sur variables, comparaison de méthodes de mesure, graphique de Bland et Altman, intervalle de tolérance

Abstract
The needs of the industries to quickly assess the quality of products and the performance of the manufacturing methods lead to the development and improvement of alternative analytical methods. These methods should ideally lead to results comparable or equivalent to those obtained by a standard method.
Firstly, the measures given by both methods can be modeled by an errors-in-variables regression. The estimated parameters are useful to check the equivalence by testing the presence of a bias between the methods. Secondly, the differences between paired measures can be analyzed to assess the agreement between the methods with the limits of agreement given by the well-known and widely used Bland and Altman's approach.
We'll compare the two methodologies and propose improvements like the tolerance interval and the use of correlated-errors-in-variables.

Keywords : errors-in-variable regressions, measurement method comparison, Bland and Altman's plot, tolerance interval

1. Introduction

There are different approaches to deal with the method comparison studies. Firstly, some authors use an approach based on a regression analysis [1]. Secondly, the most known and widely used is the approach proposed by Bland and Altman [2]. These approaches have their own advantages and disadvantages. In the errors-in-variable regressions approach, the pairs of measures taken by the reference method and the alternative one can be modeled by a linear regression (a straight line). The estimated parameters and their confidence intervals are very useful to test the equivalence. To achieve this correctly, it is essential to take into account the errors and heteroscedasticity in both axes if necessary. Different types of regression exist to handle these cases. We review the equations for estimating the regression by these different techniques in the case of homoscedasticity and standardize the notations. Secondly, the differences between paired measures can be plotted against their averages and analyzed to assess the degree of agreement between the two measurement methods by computing the limits of agreement. This is the very well-known and widely used Bland and Altman's approach. This approach can be improved by using tolerance intervals.
We'll conclude by explaining whether it's more suitable to regress in a X-Y plot or to regress the differences with their averages like in a Bland and Altman plot to assess the equivalence of measurement methods.

2. The model and the goal of testing the equivalence

2.1 What is 'equivalence' ?

When two devices are available in an industry or a laboratory to get a measure, we can wonder whether these two devices are equivalent or not. But what is the meaning of such question? What do we want to compare exactly between these two devices? Is it 'Are the measures taken by these two devices equivalent notwithstanding the errors of measurement or is there a bias between the two devices?' or is it about these two devices 'Do they have the same accuracy?' or is it 'can we substitute a device by another one without affecting the decision about the result of the measure?' We'll focus mainly on the question to know whether there is a bias between the two devices or if they give equivalent measures notwithstanding the errors of measurement.
The easiest design in method comparison studies consists probably to measure each sample one time by both devices. Unfortunately, like this, it's not possible to estimate the variances of the errors of measurement if necessary, because there are no repeated measures. That's the reason why, we recommend measuring each sample at least two times by both devices and ideally for a given device, a constant number of repetitions for all the samples.

2.2 The general model

To compare two measurement methods, N samples (i=1, 2,…, N) are measured by both methods [3]:

$$X_{ij} = \xi_i + \tau_{ij} \text{ and } Y_{ij} = \eta_i + \nu_{ij}$$

$X_{ij}\left(j = 1,2, \dots, n_{X_i}\right)$ and $Y_{ij}\left(j = 1,2, \dots, n_{Y_i}\right)$ are the repeated measures for the sample i by both methods, n_{X_i} and n_{Y_i} are the number of repeated measures of the sample i by respectively the methods

X and Y. ξ_i and η_i are the true but unobservable measures, which we'll assume that they can be modeled by a linear regression:

$$\eta_i = \alpha + \beta\xi_i$$

X_i and Y_i are the means of the repeated measures for a given sample:

$$X_i = \frac{1}{n_{X_i}}\sum_{j=1}^{n_{X_i}} X_{ij} \text{ and } Y_i = \frac{1}{n_{Y_i}}\sum_{j=1}^{n_{Y_i}} Y_{ij}$$

τ_{ij} and ν_{ij} are the measurement errors assumed to be independent and normally distributed (with constant variances in the case of homoscedasticity):

$$\begin{pmatrix}\tau_{ij}\\ \nu_{ij}\end{pmatrix} \sim iN\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma^2_{\tau_i} & 0\\ 0 & \sigma^2_{\nu_i}\end{pmatrix}\right)$$

So, the means of the repeated measures are also normally distributed around ξ_i or η_i:

$$\begin{pmatrix}X_i\\ Y_i\end{pmatrix} \sim iN\left(\begin{pmatrix}\xi_i\\ \eta_i\end{pmatrix},\begin{pmatrix}\frac{\sigma^2_{\tau_i}}{n_{X_i}} & 0\\ 0 & \frac{\sigma^2_{\nu_i}}{n_{Y_i}}\end{pmatrix}\right)$$

If the variances $\sigma^2_{\tau_i}$ and $\sigma^2_{\nu_i}$ are unknown, they can be estimated with repeated measures. Otherwise, these variances are unknown and inestimable. The estimates of $\sigma^2_{\tau_i}$ and $\sigma^2_{\nu_i}$ are given by $S^2_{\tau_i}$ and $S^2_{\nu_i}$:

$$S^2_{\tau_i} = \frac{1}{n_{X_i}-1}\sum_{j=1}^{n_{X_i}}\left(X_{ij} - X_i\right)^2 \text{ and } S^2_{\nu_i} = \frac{1}{n_{Y_i}-1}\sum_{j=1}^{n_{Y_i}}\left(Y_{ij} - Y_i\right)^2$$

2.3 The homoscedastic model

In this paper, we'll focus on the homoscedastic model. Moreover, as we'll regress the mean measures X_i and Y_i, we'll consider constant repeated measures to prevent the model to be heteroscedastic. In the case of homoscedasticity, the variances of error terms can be assumed constant and we can consider that the variances $S^2_{\tau_i}$ (and $S^2_{\nu_i}$) are the estimates of the one and the same variance σ^2_τ (and σ^2_ν) in such way that we can consider a global estimate of σ^2_τ (and σ^2_ν) given by S^2_τ (and S^2_ν):

$$S^2_\tau = \frac{\sum_{i=1}^{N}\left(n_{X_i}-1\right)s^2_{\tau_i}}{\left(\sum_{i=1}^{N} n_{X_i}\right)-N} \text{ and } S^2_\nu = \frac{\sum_{i=1}^{N}\left(n_{Y_i}-1\right)s^2_{\nu_i}}{\left(\sum_{i=1}^{N} n_{Y_i}\right)-N}$$

or with constant repeated measures, we have $n_{X_i} = n_X$ and $n_{Y_i} = n_Y$ $\forall i$:

$$S^2_\tau = \frac{(n_X-1)\sum_{i=1}^{N} s^2_{\tau_i}}{N(n_X-1)} \text{ and } S^2_\nu = \frac{(n_Y-1)\sum_{i=1}^{N} s^2_{\nu_i}}{N(n_Y-1)}$$

Finally, we'll use these notations: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$;

$$S_{xx} = \sum_{i=1}^{N}(X_i - \bar{X})^2;\ S_{yy} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \text{ and } S_{xy} = \sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}).$$

2.4 How to test the equivalence?

2.4.1 The Bland & Altman's plot vs the classical scatterplot (X,Y)

If the two measurement methods are equivalent, we can expect either that they provide (on average) the same measure for a given sample notwithstanding the errors of measurement or either that the differences between the observed measures are not higher than a practical threshold of equivalence. These two points of view are not exactly the same. Actually, the first point of view focuses directly on a 'strict' equivalence between the devices while the second point of view popularized by Bland and

Altman (BA) focuses directly to the differences between the measures. With the first point of view, errors-in-variables regressions are mainly applied with a classical scatterplot (X,Y) while the BA's plot focuses directly on the differences where intervals are mainly applied such agreement intervals or tolerance intervals.

2.4.2 Equivalence in a classical scatterplot (X,Y)

The goal of testing the equivalence is to check that $\xi_i = \eta_i \; \forall i$. In practice, these terms are unobservable because of the errors of measurement and the test of equivalence is based on the following regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \; ; \; \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \text{ and } \sigma_\varepsilon^2 = \frac{\sigma_v^2}{n_Y} + \beta^2 \frac{\sigma_\tau^2}{n_X} \; ;$$

where α, the intercept and β, the slope are estimated respectively by $\hat{\alpha}$ and $\hat{\beta}$. This regression model is applied on the averages of repeated measures assuming homoskedastic errors and constant repetitions for a given measurement method. A lot of practitioners wonder which measurement method do we have to put on the X-axis or on the Y-axis, and this is one of the critics made by Bland and Altman about the regression's approach. We'll give a rule in a further section to tackle this problem.

The estimated parameters $\hat{\alpha}$ and $\hat{\beta}$, are very useful to test the equivalence [4]. Indeed, an intercept significantly different from 0 means that there is a constant bias between the two measurement methods and a slope significantly different from 1 means that there is a proportional bias between the two measurement methods. So, we'll use the following two-sided hypothesis:

$$H_o: \alpha = 0 \; ; H_1: \alpha \neq 0 \text{ and } H_o: \beta = 1 \; ; H_1: \beta \neq 1$$

We'll consider only the confidence intervals (CI): the hypothesis that there is no constant bias will be rejected if 0 isn't included inside the CI around $\hat{\alpha}$ and the hypothesis that there is no proportional bias will be rejected if 1 isn't included inside the CI around $\hat{\beta}$. A joint-CI can also be applied.

2.4.3 Equivalence in a Bland and Altman's plot

In the previous section, the goal of testing the equivalence was to check that $\xi_i = \eta_i \; \forall i$ which is the same than $\xi_i - \eta_i = 0 \; \forall i$. As Bland and Altman focuses directly to the observed differences between the measurements methods [5-6], the goal of testing the equivalence is to check that $|Y_i - X_i| < k \; \forall i$ where k is a 'threshold' of equivalence. For instance, if the differences between measures provided by two devices are not meaningful below 10, then the threshold k is equal to 10. In a BA's plot, the differences are plotted on the Y-axis and the averages on the X-axis:

$$D_i = Y_i - X_i \; ; \; M_i = (Y_i - X_i)/2.$$

The choice to compute the differences $Y_i - X_i$ or $X_i - Y_i$ doesn't really matter. The mean and the variance of the differences can also be computed:

$$\overline{D} = \frac{1}{N}\sum_{i=1}^{N} D_i \; ; \; S_D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(D_i - \overline{D})^2$$

Horizontal agreement interval or tolerance intervals (formula given in a further section) can be computed around $\overline{D}$ and ideally such intervals are included inside the 'acceptance' interval $[-k, k]$. Otherwise, a regression line must be estimated to compute such intervals which are not necessary horizontal because computed and displayed around a regression line. To estimate a regression line in a BA's plot, the formulas given in a (X,Y) scatterplot can be easily adapted into a BA's plot. Then by analogy to the previous section, the estimated parameters can also be used to test a 'strict' equivalence with the following hypothesis:

$$H_o: \alpha = 0 \; ; H_1: \alpha \neq 0 \text{ and } H_o: \beta = 0 \; ; H_1: \beta \neq 0$$

3. The different regressions

In this paper, we'll consider only the usual parametric regressions with a clear criterion of minimization as shown in Figure 1 and describe these criteria in the next sections. We'll therefore focus on the estimation of the slope and intercept of our model $Y_i = \alpha + \beta X_i + \varepsilon_i$ using six regression techniques, study their relationships and the confidence intervals for their parameters.

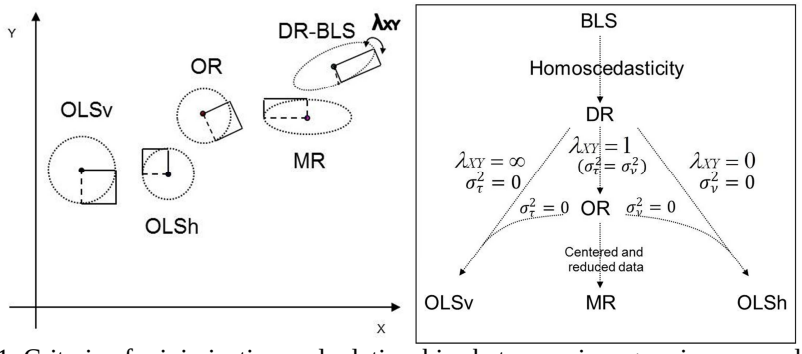


Figure 1: Criteria of minimization and relationships between six regressions according λ_{XY}

3.1 Presentation of six different regressions

3.1.1 The Ordinary Least Square regression - OLSv

The easiest way to estimate the parameters α and β is to use the OLS technique. It minimizes the sum of the squared of the vertical distances (residual) between each point and the line as shown in Figure 1 with the dashed line. To distinguish this technique to another one described in the next section, we decide to call this technique the OLSv. The OLSv minimizes the following criterion:

$$C_{OLSv} = \sum_{i=1}^{N} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2.$$

The parameters α and β can be estimated with the following formula:

$$\hat{\beta}_{OLSv} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\alpha}_{OLSv} = \bar{Y} - \hat{\beta}_{OLSv}\bar{X}.$$

Unfortunately, the OLSv doesn't take into account the errors in the X-axis. In other words, the OLSv supposes that the τ_{ij} are equal to zero. These estimates are therefore biased.

3.1.2 The Ordinary Least Square regression - OLSh

Instead of minimizing the sum of the squares of the distances between each point and the line in a vertical direction like the OLSv, we can minimize these distances in a horizontal direction (Figure 1). We'll call this technique the OLSh by analogy to the classical OLSv. The OLSh can be applied by computing the OLSv in an 'inverse' model like this:

$$X_i = \alpha^* + \beta^* Y_i + \varepsilon_i^* \text{ with } \alpha^* = -\alpha/\beta \text{ and } \beta^* = 1/\beta;$$

we find with the OLSv technique:

$$\hat{\beta}^*_{OLSv} = \frac{S_{xy}}{S_{yy}} \text{ and } \hat{\alpha}^*_{OLSv} = \bar{X} - \hat{\beta}^*_{OLSv}\bar{Y};$$

and finally the estimates given by the OLSh are:

$$\hat{\beta}_{OLSh} = \frac{1}{\hat{\beta}^*_{OLSv}} = \frac{S_{yy}}{S_{xy}} \text{ and } \hat{\alpha}_{OLSh} = -\frac{\hat{\alpha}^*_{OLSv}}{\hat{\beta}^*_{OLSv}} = \bar{Y} - \hat{\beta}_{OLSh}\bar{X}.$$

Unfortunately, by analogy to the OLSv, the OLSh doesn't take into account the errors in the Y-axis. In other words, the OLSh supposes that the v_{ij} are equal to zero. These estimates are therefore biased.

3.1.3 The geometric Mean Regression - MR

As the OLS regressions cannot tackle the problem of the errors in both axes, another regression consists to minimize the sum of area of rectangles (or ellipses) built with the projections (half-axes of ellipses) of the points to the line in parallel to the axes as shown in Figure 1. We'll call this regression the Mean Regression MR because it can be proved that its slope is the geometric mean of the two slopes given by the OLS regressions:

$$C_{MR} = \sum_{i=1}^{N} |X_i - X(Y_i)||Y_i - Y(X_i)| = \sum_{i=1}^{N} \left| X_i - \frac{Y_i}{\hat{\beta}} + \frac{\hat{\alpha}}{\hat{\beta}} \right| \left| Y_i - \hat{\alpha} - \hat{\beta} X_i \right|$$

$$\hat{\beta}_{MR} = \sqrt{\frac{S_{yy}}{S_{xx}}} \text{ and } \hat{\alpha}_{MR} = \bar{Y} - \hat{\beta}_{MR} \bar{X} \text{ or } \hat{\beta}_{MR} = \sqrt{\hat{\beta}_{OLSv} \hat{\beta}_{OLSh}}.$$

The MR was first introduced to tackle the problem of estimating a regression line between two variables but without a clear dependent variable and an independent. The MR can be used when both variables are 'interdependent' which is the case in our context because we regress the results of two measurement methods, so neither of these two methods can be considered as the response of the other.

3.1.4 The Orthogonal Regression - OR

Instead of minimising the sum of the square of the distances between each point and the line in a direction parallel to an axis like the OLS, we can choose an orthogonal direction, we'll call this regression OR for Orthogonal regression. The OR minimizes the sum of the square of the orthogonal distances between each point and the line (Figure 1):

$$C_{OR} = \sum_{i=1}^{N} \left(\left(X_i - \frac{Y_i + X_i/\hat{\beta} - \hat{\alpha}}{\hat{\beta} + 1/\hat{\beta}} \right)^2 + \left(Y_i - \hat{\alpha} - \frac{\hat{\beta} Y_i + X_i - \hat{\alpha}\hat{\beta}}{\hat{\beta} + 1/\hat{\beta}} \right)^2 \right)$$

$$\hat{\beta}_{OR} = \frac{S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}} \text{ and } \hat{\alpha}_{OR} = \bar{Y} - \hat{\beta}_{OR} \bar{X}.$$

The OR is probably the most used regression in the context of (linear-) errors-in-variables regressions because it is a good compromise when the OLS fail because of the errors in both axes. Unfortunately, the OR is only valid when variances of errors in both axes are equal.

3.1.5 The Deming Regression - DR

The GMR and OR take into account the errors in both axes but not the ratio of their variances. Indeed, we can compute the ratio between the two variances of errors:

$$\lambda_{XY} = \frac{\sigma_v^2 / n_Y}{\sigma_\tau^2 / n_X}.$$

The Deming regression that we'll call DR is the maximum likelihood solution of our model when λ is known [7-8]. In practice, if λ is unknown, it can be estimated with replicated data. The DR minimizes the sum of the square of oblique distances between each point to the line (Figure 1):

$$C_{DR} = \sum_{i=1}^{N} \left(\left(X_i - \frac{Y_i + \lambda X_i/\hat{\beta} - \hat{\alpha}}{\hat{\beta} + \lambda/\hat{\beta}} \right)^2 + \left(Y_i - \hat{\alpha} - \frac{\hat{\beta} Y_i + \lambda X_i - \hat{\alpha}\hat{\beta}}{\hat{\beta} + \lambda/\hat{\beta}} \right)^2 \right)$$

$$\hat{\beta}_{DR} = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \text{ and } \hat{\alpha}_{DR} = \bar{Y} - \hat{\beta}_{DR} \bar{X}.$$

The assumption of the DR is that λ is constant. It's obvious that the DR is equivalent to the OR when λ_{XY}=1 for the estimation of α and β. It can also be proved that the DR is equivalent to the OLSv when λ_{XY}=∞ ($\sigma_\tau^2 = 0$) and to the OLSh when λ_{XY}=0 ($\sigma_v^2 = 0$).

3.1.6 The Bivariate Least Square regression - BLS

The BLS can take into account error and heteroscedasticity in all axes (Figure 1) and is written usually in matrix notation [9-10]. We present here its formula in the case of homoscedasticity and with replicated data. The BLS minimizes the sum of weighted residuals:

$$C_{BLS} = \frac{1}{W}\sum_{i=1}^{N}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2 = (N-2)s_{BLS}^2 \text{ with } W = \sigma_\varepsilon^2 = \frac{\sigma_v^2}{n_Y} + \beta^2\frac{\sigma_\tau^2}{n_X}$$

The estimations of the parameters (the vector b) are computed by iterations with the following formula:

$$Rb = g \text{ or } b = R^{-1}g$$

$$\begin{pmatrix} N & \sum_{i=1}^{N} X_i \\ \sum_{i=1}^{N} X_i & \sum_{i=1}^{N} X_i^2 \end{pmatrix}\begin{pmatrix} \hat{\alpha}_{BLS} \\ \hat{\beta}_{BLS} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} X_i Y_i + \hat{\beta}_{BLS} C_{BLS}\, \sigma_\tau^2/n_X \end{pmatrix}$$

3.2 The relationships between the six regressions

The BLS is the most general regression and includes the five others regressions (Figure 1, right). Actually, in the case of homoscedasticity, the weights W_i in the formula of the BLS are constant and equal to W and the BLS is exactly equivalent to the DR for the estimation of the parameters. The DR is exactly equivalent to the OR when $\lambda_{XY} = 1$, to the OLSv when $\sigma_\tau^2 = 0$ (no errors in the X-axis), to the OLSh when $\sigma_v^2 = 0$ (no errors in the Y-axis). Finally, the OR and MR are exactly identical when the data are standardized. Like the OR and the MR, the DR and the BLS regressions can also take into account the reversibility of axes by adapting the value of λ.

3.3 Comparison of the six regressions

In order to compare the 6 regressions, we simulated 100000 samples (N=50), unreplicated data (λ=λ_{XY}) under the equivalence ($\eta_i = \xi_i$) for different values of λ. For each simulation, $\hat{\beta}$ were computed according the formulas given for the 6 regressions and its 95% CI (formulas not given in this paper). Thereafter, the mean of the 10000 values of $\hat{\beta}$ were computed per value of λ, by each regression's technique, in a (X,Y) plot or a Bland and Altman plot and displayed at Figure 2 on the Y-axis with smoothed lines and λ in the X-axis.

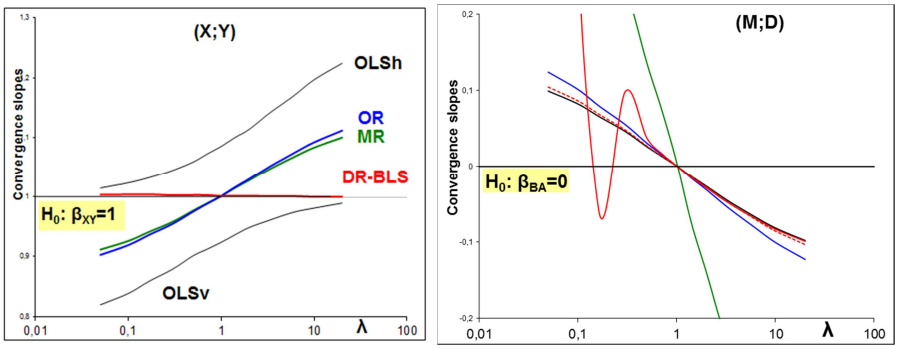


Figure 2 Convergence of slopes according λ, in a (X,Y) plot (left) or BA's plot (right)

In a (X,Y) plot: the bias of the OLS decrease when λ moves closer to their assumption (Figure 2, left). The OR and MR regressions are no-biased for λ=1. The DR and BLS regressions are equivalent to estimate the parameters and asymptotically no-biased whatever λ but the biases are lower for λ>1. That's the reason why me recommend to choose to put the two devices on the X-axis or Y-axis in such way that λ_{XY} is higher than 1. MR is more suitable in the case of unknown and inestimable λ_{XY}. Otherwise, the DR or the BLS regressions are the most suitable regressions. In a Bland and Altman plot, all the regression perform 'equally' at λ=1 (Figure 2, right) otherwise the bias increase when λ moves away from 1.

For the coverage probabilities of the CI around the slope, in a (X,Y) plot, the exact CI for the slope given by the DR provides excellent coverage whatever λ (Figure 3, left), the BLS provides good coverage especially for λ>1 while the coverage of other regressions is better when λ moves closer to their assumption. In a Bland and Altman plot, all the coverage probabilities for all kind of studied regressions collapse drastically when λ moves away from 1 (Figure 3, right).

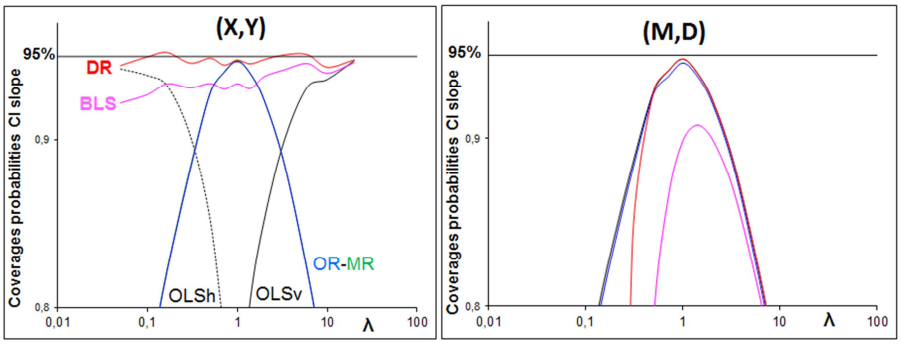


Figure 3 Coverage probabilities of CI for the slope, in a (X,Y) plot (left) or BA's plot (right)

4. The agreement and tolerance intervals

4.1 Horizontal intervals in a Bland and Altman plot

Horizontal intervals are computed under H_o and around $\overline{D}$. We give the formulas for unreplicated data. If most of the differences are not 'big' in other words lower than k (threshold of equivalence), we could conclude that the differences are not meaningful in a practical point of view even if there is not a 'strict' equivalence as defined in section 2.4.2. The formula of an agreement interval (AL) (and its CI) is (are) given by Bland and Altman such that $100(1-\alpha)\%$ of differences lie into this interval:

$$\overline{D} \pm z_{1-\alpha/2} S_D \text{ and its CI (AL CI): } \left(\overline{D} \pm z_{1-\alpha/2} S_D\right) \pm t_{N-1;1-\alpha/2} S_D \sqrt{\frac{1}{N} + \frac{z_{1-\alpha/2}^2}{2(N-1)}}$$

where z is the quantile of a standardized normal Z-distribution. Unfortunately, this interval doesn't take into account the uncertainty on the estimated variance S_D^2. This can be done by using tolerance intervals (TI) [11-12]. The beta TI (b TI) is given by the following formula:

$$\overline{D} \pm t_{N-1;1-\alpha/2} S_D \sqrt{1 + \frac{1}{N}};$$

where t is the quantile of the student t-distribution with N-1 degrees of freedom. With the beta-gamma TI (bg TI), we can add a $100(1-\gamma)\%$ confidence level with a χ^2 distribution meaning that in $100(1-\gamma)\%$ of cases, the TI will contain at least $100(1-\alpha)\%$ of observations (differences):

$$\overline{D} \pm z_{1-\alpha/2} S_D \sqrt{1 + \frac{1}{N}} \sqrt{\frac{N-1}{\chi^2_{N-1;1-\gamma}}}.$$

4.2 Comparison of the horizontal intervals

We simulated 10000 samples per value of N for a given λ, unreplicated data, under equivalence ($\xi_i - \eta_i = 0\ \forall i$) with a known distribution for the differences D_i. For each simulated sample, we compute the proportion of 'new' differences inside each interval (all levels at 95%). In other words, the proportion of the true distribution of the differences that lies into each interval. We can observe that the AL is too narrow for small sample sizes (Figure 4), the b TI is more appropriate and excellent whatever N, the bg TI is larger than the b TI but its confidence level is good whatever N, the use of the CI on AL gives an interval too large and a confidence level too high (Figure 4, sub-chart).

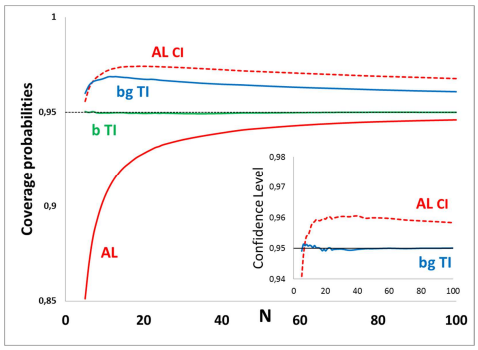


Figure 4 Coverage probabilities and Confidence level of Intervals in a BA's plot

4.3 Other intervals in a Bland and Altman plot

To build more complicated intervals and especially not-horizontal intervals, a 'good' regression line (to take into account all the specificities of a Bland and Altman plot) must be first estimated. After, TI can be applied around such intervals. Such intervals are not given in this paper, we're still working on this topic.

4. Conclusions

The Bland and Altman plot is probably the most applied method, widely used and very well-known but the tolerance intervals are more suitable than the agreement interval. In a (X,Y) plot to test the equivalence with an errors-in-variables regression, the BLS is the most suitable regression because it takes into account the errors on both axis and heteroscdasticity if needed. Unfortunately, we need to estimate a regression line in a Bland and Altman plot to get more suitable TI which are not necessary horizontal, but all the regressions here presented fail to estimate a line in a Bland and Altman plot. Actually, this is because the errors-terms are correlated in a Bland and Altman plot. So, we should apply a correlated-errors-in-variables regression to regress adequately in a Bland and Altman plot. This is further work.

Bibliographie

[1] Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *Clin Chem*, 39: 424-432.

[2] Bland, J.M., Altman, D.G. (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res*; 8; 135-160.

[3] Fuller, Wayne A. (1987), Measurement error models, *John Wiley & Sons, Inc.*

[4] Martinez, A., Del Rio, F. J., Riu, J. & Rius, F. X. (1999). Detecting proportional and constant bias in method comparison studies by using linear regression with errors in both axes. *Chemometrics and Intelligent Laboratory Systems*, 49: 181-195.

[5] Ludbrook, J. (2010). Confidence in Altman–Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology* 37, 143–149.

[6] Ludbrook, J. (2002). Statistical techniques for comparing measures and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology* 29, 527–536.

[7] Martin, R. F. (2000). General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies. *Clin Chem*, 46: 100-104.

[8] Tan, C. Y., Iglewicz, B. (1999). Measurement-Methods Comparisons and Linear Statistical Relationship. *Technometrics*; Augustus VOL. 41, NO. 3.

[9] Lisy, J. M., Cholvadova, A. & Kutej, J. (1990). Multiple straight-line least-squares analysis with uncertainties il all variables. *Computers Chem*, Vol. 14, No. 3, 189-192.

[10] Riu, J., Rius, F. X. (1996). Assessing the accuracy of analytical methods using linear regression with errors in both axes. *Anal. Chem.* 68, 1851-1857.

[11] Wald, A., Wolfowitz, J. (1946) Tolerance Limits for a Normal Distribution. *Ann. Math. Statist*. Volume 17, Number 2, 208-215

[12]Howe, W. G. (1969). Two-Sided Tolerance Limits for Normal Populations - Some Improvements, *Journal of the American Statistical Association*, 64, 610–620.

Mise en Evidence de l'Encrassement des échangeurs de chaleur à plaques lors de la pasteurisation du lait à l'aide des réseaux neurones

Demonstration of Fouling in a plate heat exchanger using artificial neural network models during milk heat treatment

Youcef Mahdi[1, 2]

[1] *Laboratoire des phénomènes de transfert, faculté de génie mécanique et de génie des procédés, université des sciences et de la technologie Houari Boumediene USTHB, bab-ezzouar, Alger 16111.*

[2]*Université de Médéa, faculté des sciences et de la technologie, pôle universitaire Médéa 26000.*
E-mail: mahdiyoucef@yahoo.fr

Résumé
L'encrassement par le lait est un phénomène qui a lieu lors de sa pasteurisation. Par formation de dépôt, il engendre une réduction de la surface de l'échangeur thermique à plaques (PHE), principal composant d'une unité de pasteurisation. Ce dépôt génère des problèmes à l'écoulement et réduit les échanges thermiques entre les deux fluides, qui circulent de part et d'autres des plaques. Dans le but de développer et mieux maîtriser le processus industriel de pasteurisation du lait, on se propose dans ce travail de prédire la masse de dépôt, ainsi que le coefficient d'échange global, à l'aide d'un modèle approprié de type réseaux de neurones préalablement adapté. Le modèle permet une recherche des conditions de fonctionnement critiques qui imposent un nettoyage des équipements une fois le seuil de tolérance dépassé. Globalement, les résultats montrent que l'encrassement est intimement lié aux conditions de fonctionnement du système. Des résultats antérieurs obtenus dans des conditions de fonctionnement analogues sont utilisés comme données de référence pour la simulation.
***Mots clés**: Lait– Encrassement – Echangeur à plaques – Réseaux de neurones*
Abstract
Milk Fouling is a phenomenon that occurs during pasteurization. By deposit formation, it results in a reduction of thermal surface heat exchanger (PHE), the main component of a pasteurization unit. This deposit generates both flow problems and reduces heat transfer between two fluids flowing from the two sides of the plates. Otherwise, in order to develop and better control the industrial process of pasteurization of milk, the main purpose of this work is to predict the mass of deposit as well as the global heat transfer coefficient, using an appropriate model for such system, a previously adapted neural network model. The model allows the determination of the critical operating conditions which required equipment cleaning once the threshold has been exceeded. Globally, the results show that fouling is closely related to the operating conditions of the system. Previous results obtained under similar operating conditions are used as set data for the simulation.
***Keywords:** Milk – Fouling – Surface heat exchangers – Neural networks*

1. Introduction et positionnement du problème

Les produits agroalimentaires subissent des transformations de plus en plus élaborées afin de présenter des qualités organoleptiques et une texture désirable. Les équipementiers ainsi que les transformateurs

cherchent à définir les lignes de processus fournissant le produit recherché dans des conditions sanitaires et économiques optimales en prenant en compte les phénomènes d'encrassement [1]. Ces derniers ont la particularité de limiter les performances des équipements thermiques augmentant ainsi le coût de production [2].

Une variété de traitements thermiques est utilisée dans les industries alimentaires. Un exemple typique est la pasteurisation du lait qui consiste en une opération de chauffage de ce dernier durant une certaine période [3]. L'augmentation de la température du lait modifie certaines de ses propriétés et le rend instable. Par la suite, un dépôt solide se forme sur la surface de l'échangeur de chaleur. Cependant, ce dépôt génère des perturbations hydrauliques et thermiques qui affectent la production et qui créent ainsi le besoin de nettoyer les équipements concernés [4]. Cette opération est nécessaire afin de reconstituer la surface initiale propre et assurer ainsi la qualité et l'hygiène des productions ultérieures. Dans l'industrie laitière, il est de pratique courante de procéder au nettoyage des échangeurs du pasteurisateur toutes les 5 à 10 heures.

Ces dernières années, différents auteurs se sont intéressés à l'emploi des techniques des réseaux de neurones pour prévoir l'encrassement des équipements. Dornier et al. [5], Teodosiu et al. [6] et plus récemment Shetty et Chellam [7] ont présenté un travail pour prédire l'encrassement par la technique des réseaux neuronaux. Zhang et al. [8] ont utilisé la technique pour estimer l'encrassement dans un réacteur lors d'un procédé de polymérisation. D'autres auteurs ont considéré l'encrassement dans des installations industrielles, mais l'utilisation de la technique dans le domaine du nettoyage de ces équipements reste presque inexistante. Dans le cas de la pasteurisation du lait, après une période de fonctionnement normale, les conditions de fonctionnement deviennent critiques. Ceci se produit suite à la formation d'un dépôt, qui n'a pas lieu dès le début du processus. Cette période est nécessaire pour atteindre l'activation du processus d'encrassement suite à l'opération de chauffage. Le modèle type du bilan de matière utilisé est composé par des relations fortement non linéaires, multivariables et soumises à des influences complexes d'où tout l'intérêt des réseaux de neurones.

Les réseaux de neurones sont un nouveau outil utilisé pour résoudre le problème de l'encrassement qui se base sur une stratégie de type boîte noire. Cependant, la finesse dans la présentation du processus peut aider l'architecte du dimensionnement à trouver rapidement des corrélations adéquates pour le modèle neuronal. Un ordre de l'évolution du système est également inclut dans l'architecture du modèle. Le choix du nombre des entrées est dicté par l'existence possible de contraintes et de leurs nombres.

L'objectif de ce travail est d'appliquer la technique des réseaux de neurones artificiels dans le cas d'une installation de pasteurisation du lait. Ceci afin de prédire la masse de dépôt due à l'encrassement des protéines et à l'entartrage des sels sur la surface du pasteurisateur afin d'estimer le coefficient d'échange thermique global, ainsi que et le seuil qui impose la nécessité de nettoyer.

2. Formulation du problème

Généralement, le modèle mathématique est utilisé pour réaliser les objectifs suivants :

- ✓ Permettre l'interprétation qualitative et quantitative des résultats,
- ✓ analyser les hypothèses relatives au fonctionnement du procédé,
- ✓ optimiser les opérations et le contrôle du procédé.

Dans le présent travail, une formulation du problème de l'encrassement des échangeurs lors de la pasteurisation du lait est proposée par la technique des réseaux neuronaux afin de le minimiser, sans pour autant déstabiliser le système et établir éventuellement une loi de commande. Mais avant d'y parvenir, nous avons besoin d'un modèle précis et simple pour prédire toute évolution des variables.

Dans un premier temps, une présentation des conditions optimales de fonctionnement est effectuée, afin de figurer comme référence pour adapter le modèle neuronal choisi. Par la suite, il est nécessaire de considérer une surface d'échange minimale qui permet un fonctionnement optimal de l'échangeur.

Le nettoyage devient nécessaire quand la surface de l'échange thermique devient inférieure à la surface de l'échange minimale préconisée.

Les hypothèses suivantes sont considérées:

- ✓ L'échange thermique a lieu entre une vapeur d'eau qui passe une fois par canal pour un total de cinq passes dans l'échangeur et le lait qui passe également dans cinq canaux avec un seul passage par canal en fonctionnement continu.
- ✓ Les profils d'encrassement des équipements considérés utilisés comme références sont estimés en utilisant des simulations antérieures basées sur une modélisation type génie chimique évoluant avec des conditions opératoires similaires.

La surface de l'échange thermique est calculée par:

$$S = \frac{Q}{U\Delta T_{log}} \tag{1}$$

$$\Delta T_{log} = \frac{\Delta T_1 - \Delta T_2}{\log\left(\frac{\Delta T_1}{\Delta T_2}\right)} \qquad \text{et} \quad \Delta T_1 = T_c^e - T_f^s \,,\ \Delta T_2 = T_c^s - T_f^e \tag{2}$$

Le flux thermique est calculé par:

$$Q = m_c\left(T_c^e - T_c^s\right) = m_f\left(T_f^s - T_f^e\right) \tag{3}$$

Durant la pasteurisation du lait, l'encrassement engendre une réduction de la section de passage du fluide et réduit les échanges entre les deux fluides circulant dans l'échangeur car le dépôt est de nature isolante. Une fois la valeur de la masse critique de dépôt M_{cr} atteinte, les réseaux de neurones opèrent pour déclencher le nettoyage. D'ou la contrainte suivante:

$$M < M_{cr} \tag{4}$$

La masse critique de dépôt est calculée en considérant une limite de chute de 15% du coefficient d'échange global, ce qui donne lieu à une mauvaise pasteurisation.

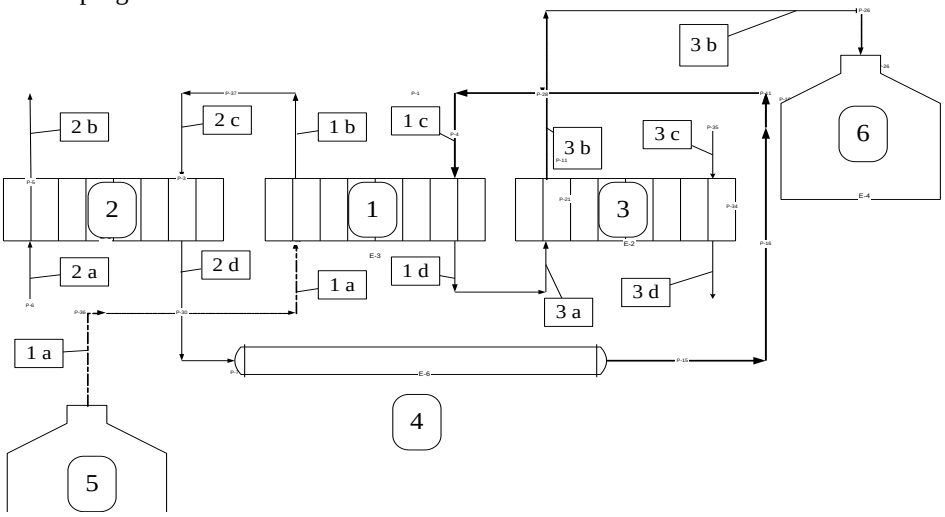
3. Description du système et méthodes

3.1 Description du système

Le procédé de pasteurisation a lieu dans un ensemble d'échangeurs de chaleur à plaques représentés sur la figure 1. Ils sont recommandés pour ce type d'opérations, car ils offrent un coefficient d'échange global important. La pasteurisation consiste en un chauffage du lait à une température donnée pendant un certain temps afin d'éliminer l'action pathologique des bactéries. C'est un traitement obligatoire que doit subir le lait avant toute utilisation. Le lait subit principalement un échange thermique avec une vapeur d'eau circulant dans le pasteurisateur.

L'échangeur de chaleur à plaques est un ensemble de plaques en acier maintenu dans une ossature métallique. Les plaques sont corrugées afin d'augmenter la turbulence de l'écoulement et améliorer ainsi l'échange thermique entre les deux fluides. Le pasteurisateur est divisé en trois sections. La section 1 est une section de récupération d'énergie du lait pasteurisé sortant, la section 2 est une section de chauffage et enfin la section 3 est une section de refroidissement. En premier lieu, le lait entre dans la section 1 à une température de 4°C pour récupérer la chaleur du lait sortant, qui est à une température est de 70°C. Il aborde ensuite la section 2 dans laquelle il subit l'opération de chauffage proprement dite grâce à l'échange thermique avec de la vapeur d'eau. En sortant, sa température atteint une valeur de 90 °C, température suffisante pour une pasteurisation. La durée de la pasteurisation est de 15 secondes. Pour ce faire, le lait traverse un tube totalement isolé. Enfin, il traverse successivement les sections 1 et 3 pour céder sa chaleur au lait entrant puis à l'eau de refroidissement. L'estimation de l'encrassement et de la température est réalisée à l'aide de simulations antérieures pour la section principale de chauffage (la section 2). La perte de pression est

considérée comme étant la différence de pression entre l'entrée et la sortie de la section 2. Le temps de l'échantillonnage est de 20 secondes et les valeurs sont enregistrées sur Excel. Les réseaux de neurones sont programmés sur Visual Basic V 5.0.



		a	b	c	d
1	**Section 1** Récupération	Entrée lait 4 °C	Sortie lait 70 °C	Lait 90 °C vers la récupération	Lait vers refroidissement
2	**Section 2** Chauffage	Entrée vapeur	Sortie vapeur	Entrée lait 70 °C	Lait 90 °C vers chambrage
3	**Section 3** Refroidissement	Lait vers refroidissement	Lait vers stockage	Entrée eau froide	Sortie eau froide
4	Opération de chambrage à 90°C (passage dans le tube)				
5	Stockage lait avant pasteurisation				
6	Stockage lait après pasteurisation				

Figure 1. Description du système

3.2 Caractérisation des réseaux neurones

L'élément de base d'un réseau est le nœud (i.e., le neurone). Il est nécessaire d'associer à chaque neurone un modèle. L'expression suivante est largement utilisée dans la littérature :

$$f(Z) = \frac{1}{1+\exp(-\lambda_j Z)} \tag{5}$$

Plusieurs possibilités d'interconnexion de ces neurones peuvent être définies selon le but recherché. Ces dernières années, l'utilisation de la technique des réseaux neuronaux est devenue fréquente. Hagglund [9] est le premier qui a formulé l'algorithme de base. Depuis, des améliorations ont été apportées. L'architecture des réseaux de neurones Adaline (élément linéaire adaptative) par exemple est composée de deux neurones en entrée et un seul en sortie. Ce dernier calcule la somme linéaire i des entrées X_i pour n'importe quel neurone de l'entrée connecté au neurone j de la sortie, multiplié par les poids correspondants Γ_{ij} et en considérant le seuil ϑ :

$$i = \sum_{i=1}^{n} \Gamma_i X_i + \vartheta \tag{6}$$

L'objectif de l'algorithme Adaline est de minimiser l'erreur entre la sortie recherchée et celle obtenue par les réseaux de neurones pour tout le système. Elle est définie comme étant une valeur normalisée du coefficient de transfert thermique. La fonction erreur est donnée par :

$$E = Min[f(a,b)] = \sum\left(\frac{U}{U_0}\right) \tag{7}$$

Avec a la valeur recherchée et b la valeur actuelle. Le but recherchée et de trouver une erreur minimale en se basant sur l'ajustement de chaque poids par un paramètre proportionnel au carré de la dérivée de l'erreur pour chaque entrée, selon :

$$\Delta\Gamma_{ij} = -\xi\left(\frac{dE}{d\Gamma_{ij}}\right)^2 \tag{8}$$

3.3 Mécanisme d'encrassement

Une fois le lait chauffé à une température supérieure à 65 °C, la protéine βétalactoglobuline devient instable et se transforme en précurseur de dépôt selon deux mécanismes possibles [10,11] :

- ✓ La protéine βétalactoglobuline naturelle subit une dénaturation (altération de sa structure) en dévoilant ses groupements (-SH) et devient ainsi réactive.
- ✓ Une réaction de polymérisation irréversible donne naissance à des particules insolubles sous forme d'agrégats (amas de matière).

Les cinétiques sont connues [12]. Un transfert de masse des trois formes de la protéine a lieu entre le fluide et la couche limite. Le dépôt est formé par la protéine agrégée. Il est primordial de connaître les différents paramètres physico-chimiques des différents phénomènes ayant lieu pour pouvoir quantifier le dépôt et ainsi connaître la résistance due à l'encrassement [13-15].

Le deuxième type d'encrassement supposé avoir lieu dans les pasteurisateurs Algériens en même temps que celui cité précédemment est celui du phosphate de calcium qui possède une solubilité inverse par rapport à la température, c'est à dire un gradient négatif de solubilité, due probablement à la dureté trop élevée des eaux de forages employées pour la préparation du lait. Lors du chauffage du lait, le produit ionique dépasse la concentration limite de solubilité. Les sels sédimentent sous forme cristalline et se déposent sur la paroi [9].

3.4 Application de la technique des réseaux neurones

La procédure utilisée pour développer et appliquer le réseau de neurones est la suivante:

- ✓ Détermination des mesures disponibles pour le modèle et définition du nombre d'entrées et de sorties du réseau,
- ✓ définition d'un ensemble d'apprentissage de paires de vecteurs d'entrées/sorties,
- ✓ sélection d'une configuration du réseau,
- ✓ apprentissage du réseau en utilisant les données de simulation,
- ✓ test du réseau par présentation d'un ensemble d'entrées non apprises et en observant les réponses sur les sorties obtenus,
- ✓ calcul de l'indice de performance du réseau à partir de l'erreur moyenne entre les sorties du réseau et les mesures des simulations
- ✓ si l'erreur est acceptable, poursuite de l'algorithme. Sinon, utilisation d'une autre configuration et retour au troisième point,
- ✓ utilisation du réseau déjà entraîné en fonctionnement selon les simulations et affichage des variables de sorties.

Les réseaux de neurones Adaline utilisés dans ce travail sont basés sur l'utilisation de deux entrées, la température et le flux de chaleur, et une sortie, la masse de dépôt ou le coefficient d'échange global.
Au début du processus, la surface de l'échangeur est propre et l'erreur est ainsi égale à 1. L'augmentation de la température du lait déclenche le processus de dégradation des protéines et de

sédimentation des sels, ce qui donne lieu au dépôt qui engendre l'encrassement et la chute du coefficient d'échange global. Dans cette situation, l'erreur tend vers zéro et le nettoyage des échangeurs du pasteurisateur devient nécessaire. L'encrassement dû au dépôt crée une nouvelle résistance thermique et cause ainsi la chute des échange thermiques entre les deux fluides de part et d'autre des plaques de l'échangeur (la conductivité thermique du dépôt est égale à 0.5 W/m°C [16]). Ceci donne lieu à une chute de température et donc une mauvaise pasteurisation.

La résistance thermique de cet encrassement est donnée par :

$$R = \frac{1}{U} - \frac{1}{U_0} \tag{9}$$

Le coefficient global de transfert après encrassement est donné par [1] :

$$U = \frac{U_0}{1 + Bi} \tag{10}$$

Le nombre de Biot est égal à zéro quand il n'y a pas d'encrassement et prend, en théorie, des valeurs infinies quand l'encrassement a lieu.

La masse, exprimée en Kg/m^2, de la crasse déposée sur une plaque de l'échangeur peut être calculée à l'aide de la relation [17] :

$$M(x, y) = \frac{\lambda_d Bi_p(x, y)\rho_d}{U_0} + t\, k_s \log\left[\frac{I}{L_L}\right] \tag{11}$$

La procédure suivante est utilisée pour estimer la masse du dépôt :

- ✓ Des résultats antérieurs de simulation (température et flux) sont utilisées comme données de base (référence) afin d'adapter le réseau neuronal Adaline choisi,
- ✓ les réseaux de neurones sont introduits pour mettre en évidence les conditions opératoires qui donnent une erreur minimale (équation 7),
- ✓ la masse de dépôt (équation 11) ou le coefficient d'échange global (équation 10) sont calculés.
- ✓ la valeur obtenue est comparée avec la valeur trouvée par simulation (prise comme référence).

4. Résultats et discussions

La première étape consiste à définir un critère pour choisir les paramètres de la structure en fonction des objectifs de modélisation. Ce critère est défini comme étant la différence entre la valeur issue de la simulation type génie chimique avec celle obtenue par la simulation à l'aide du réseau de neurones :

$$C = \frac{\sum_{i=1}^{n} \left(P_{ref}(t_i) - P(t_i)\right)^2}{\sum_{i=1}^{n} P_{ref}(t_i)^2} \tag{12}$$

Nous avons utilisé ce critère pour étudier l'influence de plusieurs paramètres. A titre d'exemple, nous présentons sur la figure 2, l'évolution du critère C pour le choix du nombre de neurones dans la couche cachée. On peut constater que le critère d'erreur se stabilise pour un nombre de neurones supérieur à 10. Au delà, les résultats obtenus se sont pas améliorés de façons significative et ce, en apprentissage et en validation.

Le but est de mettre en évidence l'encrassement dans l'échangeur à plaques section chauffage. Au départ, le système fonctionne dans les conditions normales de pasteurisation. Les résultats sont illustrés dans le tableau 1. Après un temps de fonctionnement de 220 minutes, le réseau de neurones détecte un encrassement et réclame l'arrêt du fonctionnement du système pour nettoyage, sans avoir la possibilité d'agir, car la surface d'échange minimale a été atteinte. Les valeurs de la masse de dépôt sont comparées à celle obtenue par les simulations antérieures, prises comme référence. Une bonne coordination est remarquée.

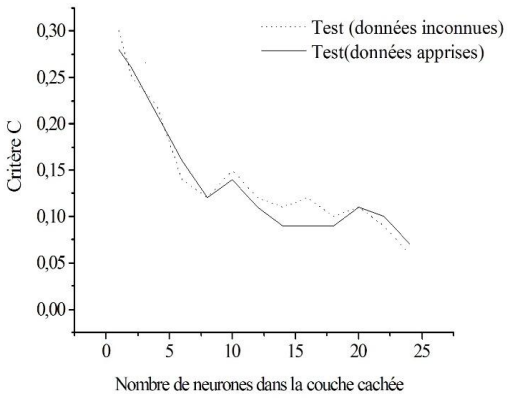


Figure 2. Evolution du critère de sélection de la structure en fonction du nombre de neurones dans la couche cachée.

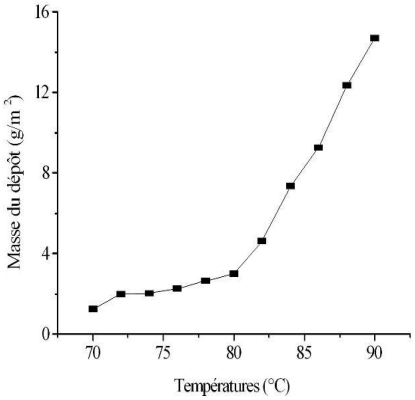


Figure 3. Evolution de la masse de dépôt en fonction de la température dans la section 2.

Temps (min)	$M_{simu}(g/m^2)$	$M(g/m^2)$
50	04.60	04.66
100	11.20	11.60
180	14.63	14.60
220	16.03	16.94

Table1. Résultats obtenus dans les conditions opératoires ordinaires (Température de pasteurisation égale à 90 °C)

La masse de dépôt dans la section de chauffage est représentée en fonction de la température sur la figure 3. Le résultat indique que l'augmentation de la température provoque l'augmentation de la quantité de dépôt. Ceci est dû d'une part à l'activation par la température du processus de dégradation de la protéine et par conséquent l'encrassement et d'autre part à la sursaturation des sels, mais sans atteindre le seuil de nettoyage préconisé par les réseaux de neurones.

L'analyse de ces résultats requiert la compréhension des causes de la diminution des performances du processus d'échange thermique. Pour ce faire, l'évolution du coefficient d'échange global est recherchée ; elle est présentée sur la figure 4. Au cours de l'évolution du système, l'augmentation de la température déclenche le processus de dégradation de la protéine et la sursaturation des sels, ce qui donne lieu à une couche non seulement encrassante mais également isolante, d'où la dégradation de la qualité des échanges. On peut remarquer qu'une diminution de l'ordre de 18% est observée au bout de 220 minutes de fonctionnement. Elle indique que la surface minimale a été atteinte ; les réseaux de neurones indiquent donc le nettoyage. Ceci confirme le résultat trouvé précédemment lors de l'estimation de la masse de dépôt (Tableau 1).

L'encrassement forme un dépôt qui entraîne un rétrécissement du diamètre accessible pour le passage du fluide, d'ou une perte de pression. Elle est représentée sur la figure 5. Sur la même figure est représentée pour comparaison l'évolution expérimentale de la perte de pression obtenue dans des conditions de fonctionnement similaires. Cette perte de pression est due au dépôt sur la surface de l'échangeur. L'observation des résultats expérimentaux montre une augmentation rapide de la perte de pression au début du processus, en raison d'un encrassement rapide qui se produit et qui est probablement dû à la mauvaise qualité de la poudre de lait. Pour ceux obtenus par les réseaux neuronaux, il apparaît que la perte de pression augmente plus discrètement puis reste sensiblement constante. Dans les deux situations, la perte de pression représente la formation d'un dépôt qui augmente pendant le fonctionnement du système par suite d'une accumulation du dépôt sur les surfaces du pasteurisateur. Cette crasse engendre des perturbations à l'écoulement et réduit ainsi la

section de passage du fluide. Les réseaux de neurones conseillent l'arrêt du fonctionnement du système au bout de 229 minutes. Il convient de noter que la perte de pression est calculée en utilisant l'équation de Darcy employée pour un système tubulaire circulaire et qui reste applicable pour un système à plaques. Cette approximation n'a pas beaucoup d'influencer sur le résultat [12].

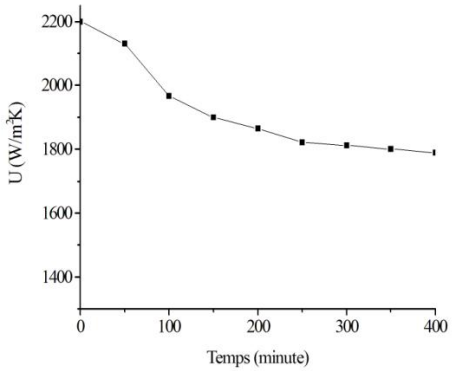


Figure 4. Evolution du coefficient global d'échange dans la section 2.

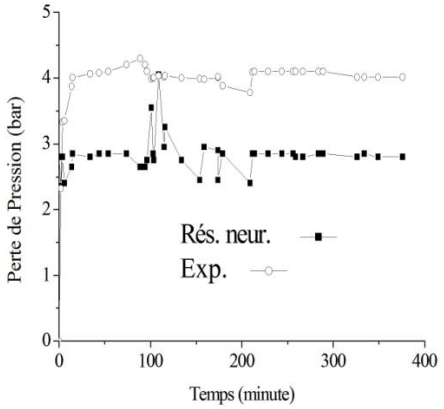


Figure 5. Evolution de la perte de pression dans La section 2.

La masse de dépôt accumulée sur la surface de l'échangeur le long de la section 2 du pasteurisateur en fonction du nombre de Reynolds est représentée sur la figure 6. Elle augmente le long de la section en raison de l'augmentation de la température qui fait suite à l'échange thermique accompli entre les deux fluides qui devient constante au bout du huitième canal. La masse de dépôt diminue avec l'augmentation du nombre de Reynolds. Ce résultat est en accord avec ceux de certains auteurs [16,19]. Ceci s'explique par le fait que la vitesse d'élimination du dépôt devient plus importante que la vitesse d'encrassement, avec l'augmentation du nombre de Reynolds. Cependant, ce dernier donne lieu à une perte de pression plus importante. Les réseaux de neurones recommandent l'arrêt du fonctionnement pour le nettoyage uniquement dans le cas ou le nombre de Reynolds est égal à 2000, car la surface minimale critique a été atteinte. Ceci suggère qu'un fonctionnement à des nombres de Reynolds supérieurs est souhaitable et se présente comme une alternative pour contrecarrer le problème de l'encrassement, mais entraînera probablement des pertes de pression supplémentaires.

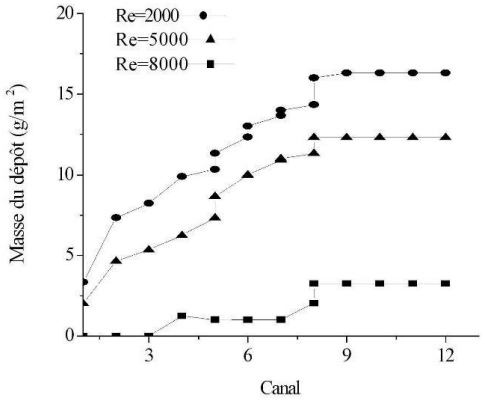


Figure 6. Evolution de la masse de dépôt en fonction du nombre de Reynolds dans la section 2.

5. Conclusion

Dans ce travail l'étude du phénomène de l'encrassement des échangeurs de chaleurs à plaques lors de la pasteurisation du lait à été considérée par la technique des réseaux de neurones en prenant en

compte les échanges thermiques entre les deux fluides, ainsi que les différentes interactions et réactions ayant lieu dans le lait. Des simulations de type génie chimique ont été considérées comme référence, pour évaluer ceux obtenues par les réseaux de neurones.

Les résultats montrent que l'encrassement augmente avec la température et provoque des perturbations hydrauliques et thermiques, suite au rétrécissement de la section de passage, qui se traduit par une diminution du coefficient d'échange globale et une perte de pression. Les résultats obtenus montrent également qu'une augmentation du nombre de Reynolds permet de réduire l'encrassement.

Il apparaît, à l'examen de ces résultats, qu'un compromis entre les différentes conditions opératoires est nécessaire afin de minimiser l'encrassement. L'approche par les réseaux de neurones permet, à chaque fois que c'est nécessaire, de mettre en évidence l'encrassement et de prévoir ainsi le nettoyage des échangeurs du pasteurisateur au moment opportun.

Nomenclature

a	Variable recherchée des réseaux neurones
b	Variable actuelle des réseaux neurones
Bi	Nombre de Biot
C	Critère de sélection de la structure des réseaux de neurones
E	Fonction erreur des réseaux neurones
f	fonction sigmoïde
I	Produit d'activité
i	Somme linéaire des entrées
K_s	Constante de dépôt du phosphate de calcium, kg/m^2s
L_L	Produit de solubilité
M	Masse du dépôt, g/m^2
m	Débit massique, Kg/s
Q	Flux thermique, w
P	Valeur du paramètre simulé
R	Résistance thermique, m^2 K/w
S	Surface de l'échange thermique, m^2
t	Temps, s ou min
T	Température, °C
ΔT	Différence de température, °C
ΔT_{log}	Différence de température logarithmique, °C
U	Coefficient global de transfert thermique, w/m^2K
X	Entrée réseaux neurones
Symboles grecs :	
Γ	Variable poids
$\Delta\Gamma$	Ajustement des variables poids
ξ	Constante de proportionnalité
ϑ	Seuil (-1,1)
ρ	Densité, kg/m^3
λ	Conductivité thermique, w/m°C
λ_j	Gain d'apprentissage

Indices / exposants :

c	Chaud
cr	Critique
d	Dépôt
e	Entrée fluide
f	Froid
j	Sortie neurone
i	Entrée des réseaux neurones
p	Plaque
ref	référence
s	Sortie fluide

simu	Simulation
0	Initial

Bibliographie

[1] Changani, S.D., Belmar-Beiny, M.T., & Fryer, P.Y. (1999). Engineering and chemical factors associated with fouling and cleaning in milk processing. *Experimental Thermal and Fluid Science*, (14) 392-406.

[2] Tissier, J.P., Lalande, M., & Corrieu, G. (1984). A study of milk deposit on a heat exchange surface during ultra-high-temperature treatment in engineering and food. *Engineering Sciences in the Food Industry*, Vol. 1, B. M. McKenna (Ed.), Applied Science Publishers.

[3] Lund, D.B., Sandu, C., Plett, C., & Jeurruink, T.J.M. (1995). Fouling and Cleaning in Food Processing. Fouling of Heat-Exchangers by Fresh and Reconstituted Milk and the Influence of Air Bubbles, Milchwissensch. *Milk ci. Int.*, (50) N 4 189-193.

[4] Lalande, M., & Corrieu, G. (1981). Fouling of a Plate Heat Exchanger by Milk In Fundamentals and Applications of Surface Phenomena associated with Fouling and Cleaning in Food Processing. B. Hallstrom, D. B. Lund, and C. Trigirdh, Eds., Univ. of Lund, Sweden 279-288.

[5] Dornier, M., Decloux, M., Trystam, M.G., & Lebert, A.(1995). Neural networks model cross flow micro filtration. *Membrane Technology*, (65) 8–9.

[6] Teodosiu, C., Pastravanu, O., & Macoveanu, M.(2000). Neural network models for ultra filtration and backwashing. *Water Research*, (34) 4371–4380.

[7] Shetty, G., & Chellam, S. (2003). Predicting membrane fouling during municipal drinking water Nan filtration using artificial neural networks. *Journal of Membrane Science*, (217) 69–86.

[8] Zhang, J., Morris, A., Martin, E., & Kiparissides, C. (1999). Estimation of impurity and fouling in batch polymerization reactors through the application of neural networks. *Computer and Chemical Engineering*, (23) 301–314.

[9] Hagglund, T. (1983). New estimation techniques for adaptative control. (1983). PhD Thesis, Department of Automatic Control, Lund University, Lund, Sweden.

[10] Lalande, M., Tissier, J.P., & Corrieu, G. (1984). Fouling of a Plate Exchanger Used in Ultra-High-Temperature Sterilisation of Milk. *Journal Dairy Res.*, (51) 557-568.

[11] Lalande, M., Tissier, J.P., & Corrieu, G. (1985). Fouling of Heat Transfer Surfaces Related to β-Lactoglobulin Denaturation during Heat processing of Milk. *Biotechnol. Prog*., (1) 131-139.

[12] Jeurrink, T.J.M. (1994). Fouling of Milk with Various Calcium Concentrations. Presented at Fouling and Cleaning in Food Processing, Jesus College, Univ. Cambridge, Cambridge, UK.

[13] Lalande, M., & Reno, F. (1984). Fouling by Milk and Dairy Product and Cleaning of Heat Exchange Surfaces In Fouling Science and Technology. L. F. Melo, T. R. Bott et C. A. Bernardo, Eds., NATO ASI Series E, Kluwer, Amsterdam, Netherlands, (145) 557-573.

[14] Vissier, J., & Jeurruink, T.J.M. (1997). Fouling of Heat Exchangers in the Dairy Industry. *Experimental Thermal Fluid Sciences*, (14) 407-424.

[15] Delaplace, F., Leulliet, J.C., & Tissier, J.P. (1994). Fouling Experiments of a Plate Heat Exchanger by Whey Protein Solutions Trans. I*ChemE*. 72(C) 163-169.

[16] De Jong, P. (1996). Modelling and optimisation of thermal treatments in the dairy industry. *NIZO Research* Report Ede, The NetherlandsV341 (p.165).

[17] Mahdi, Y., Mouheb, A. & Oufer, L. (2009). A dynamic model for milk fouling in a plate heat exchanger. *Applied Mathematical Modelling*, Elesevier Ed., (33) Issue 2 648-662.

[18] Nema, P.K., & Datta, A.K. (2005). Computer based solution to check the drop in milk outlet temperature due to fouling in a tubular heat exchanger. *Journal of Food Engineering*, (71) 141-156.

[19] Bott, T.R., & Melo, L.F., (1997). Fouling of heat exchangers. *Experimental Thermal and Fluid Science*, (14) 315-321.

Session 10 : Analyse de Risque III / *Risk Analysis III*

Élicitation pour l'évaluation des risques microbiologiques dans les aliments : vers une approche probabiliste de l'outil Risk Ranger

Elicitation for food microbial risk assessment: a probabilistic approach extending Risk Ranger proposal

Guillier Laurent[1], Kabunda Jean-Marc[2], Denis Jean-Baptiste[3] et Albert Isabelle[2]

[1] *Anses, Laboratoire de sécurité des aliments, 23 avenue du Général de Gaulle, 94700 Maisons-Alfort*
E-mail : *laurent.guillier@anses.fr*

[2] *INRA - Met@risk, 16 rue Claude Bernard, 75231 Paris cedex 05*
E-mail : *isabelle.albert@paris.inra.fr*

[3] *INRA - MIAJ, Domaine de Vilvert, 78352 Jouy-en-Josas cedex*
E-mail : *Jean-Baptiste.Denis@jouy.inra.fr*

Résumé
Ross et Sumner (2002) proposent un outil pratique sous la forme d'une feuille de calcul Excel, Risk Ranger, pour une approche rapide de l'évaluation des risques dans les aliments. L'outil est un moyen simple pour comparer et classer les risques liés à certains aliments et pour identifier les facteurs qui contribuent le plus à ces risques. La sortie de l'outil est un score calculé à partir des réponses renseignées sur 11 questions. L'objectif de ce travail est de faire évoluer l'outil Risk Ranger vers une version probabiliste. Pour cela, nous proposons ici une procédure d'élicitation de la variabilité à partir de deux quantiles de la distribution sous-jacente à la quantité d'intérêt dans la question. Les experts sont également interrogés sur leur degré de confiance pour chacun des quantiles donnés. Ce degré de confiance est utilisé pour intégrer un niveau d'incertitude à la quantité d'intérêt. Le nouvel outil permet à l'expert de modifier de vérifier graphiquement presque instantanément la conséquence de ses réponses sur l'incertitude et la variabilité de la quantité renseignée.
Mots-clés : élicitation, appréciation des risques, classification

Abstract
Ross and Sumner (2002) proposed a convenient tool, the Risk Ranger, for early-stage risk assessment of microbial hazards in food systems. The authors describe the tool as being a simple way of comparing and classifying food-related risks and highlighting main factors that contribute to food safety. The output of the tool is a risk score based on inputs obtained from answers to 11 questions. The objective of this work was to extend Risk Ranger in order to obtain a probabilistic version distinguishing uncertainty and variability. We propose an elicitation procedure where the expert is asked for two quantiles for assessing variability. Experts are also asked on their degree of confidence for the given quantiles. Degree of confidence were used to incorporate an uncertainty level. The new tool allows the expert to check graphically, almost instantly, the uncertainty and variability of the elicited variable and then to interactively modify it according his/her view.
Keywords : elicitation, risk assessment, ranking

1. Introduction

Quantitative microbial risk assessment (QMRA) aims to model the fate of pathogenic micro-organisms through the food production chain and to evaluate the health related risks. Moreover, it permits to estimate the impact of potential interventions measures on public health.

QMRA can be complex, time-consuming and expensive according to aims of risk managers. A QMRA can also in principle be simple especially when an order of magnitude estimate in expected. In such cases, point estimates and simplified model shall be used. In this context Ross and Sumner (2002) proposed a convenient tool, the Risk Ranger. The authors describe the tool as being a simple way of comparing food-related risks and classifying/ranking them and highlighting factors that contribute to food safety risks.

Risk ranger uses the principles of risk assessment, i.e. it incorporates the likelihood of exposure to a food-related risk, the prevalence of hazards in a food product when they exist, and the likelihood and severity of the consequences of a particular contamination level and frequency of exposure.

The tool requires that the user choose qualitative or quantitative statements concerning the factors that will affect the risk related to a specific food product and a specific hazard for a specific population, from production to consumption. An Excel worksheet converts the qualitative descriptions into numerical values and combines them with the quantitative statements in a series of mathematical and logical steps that use standard spreadsheet functions. Risk Ranger have been used for assessing risk for various pathogen and or food (e.g. Mataragas et al., 2008; Guillier et al., 2011, Sosa Mejia et al., 2011).

As Ross and Sumner (2002) pointed out, the tool can still be improved. They especially identified the possibility "*to enter a range of values, or distribution of values that would offer some of the benefits of stochastic modelling, but still in a relatively simple tool*". In this way, Davidson et al. (2006) proposed some modifications to create a fuzzy risk assessment tool.

The objective of our work was to generalize Risk Ranger (RR) in order to obtain a new probabilistic tool for early-stage risk assessment of microbial hazards in food systems taking into account the two major concept of uncertainty and variability.

2. The Risk Ranger Chain

RR can be interpreted as a Bayesian network with deterministic relationships (Figure 1), our attempt can be seen as introducing randomness in the ancestor (input) nodes.

2.1 Description of the input nodes

The calculation of the outputs of RR is based on inputs obtained from answers to 11 questions. The answer to the question number n, will be associated to the variable X_n.

The first question concerns the severity of the hazard considered. The hazard severity was in the first version of risk ranger arbitrarily weighted by factors of 10 for increasing levels of severity. In the current available version, it is assumed that a severe hazard (death for most victims) is 100 times more ''serious'' than a moderate case of illness.

Then, four questions concern the population exposed. Question 2 concerns the definition of the population of interest. The expert has to select it among four categories proposed, based on their susceptibility to illness. The weighting of relative susceptibility of these categories of consumers, with known predisposing conditions, was based on the relative risk of listeriosis. Question 3deals with frequency of consumption. The expert has five possible choices. He/she can choose within the following frequencies: daily, weekly, monthly, a few times per year or "other". For the four first choices, the number of days of consumption per year are 365, 52, 12 and 3 respectively. When "other" is selected, the expert has to indicate the number of days between two servings. The frequency of consumption per year is then calculated by dividing 365 by n3. Question 4deals with the proportion of population consuming the product. The expert has four possible followings choices : 100%, 75%, 25% and 5%. Question 5 concerns the size of the general population of interest. The expert can select a region ("Australia" or various Australian regions) or enter a specific population size.

The last six questions deal with the fate of the hazard in the food chain. Question 6 is about the proportion of raw product contaminated. The expert can either choose within five linguistic values: rare (1 product contaminated out of 1000), infrequent (1%), sometimes (10%), common (50%), all (100%) or give its own estimate. Question 7 deals with the effect of processing on the hazard. Here again the expert can either choose within seven linguistic values that represent the ability of processing step to reduce or to increase the level of the considered hazard. The following qualitative statements are proposed: reliability eliminates (100% reduction), usually eliminates (99% reduction), slightly reduces (50% reduction), has no effect (weight of 1), increases (multiplication by 10), greatly increases (multiplication by 1000). For question 8 the expert has to evaluate the frequency of recontamination after processing within the following statements: no (0%), minor (1%), major (50%) and other (% to assess). Question 9 considers the potential increase of the hazard during storage, distribution and retailing. Question 10 is about the ratio between the level of hazard in the product at consumption and the level thought to cause an illness in a consumer (with a susceptibility corresponding to the general population). Question 11: Effect of meal preparation. For this question, the expert can choose between statements that are almost the same as those proposed for question 7.

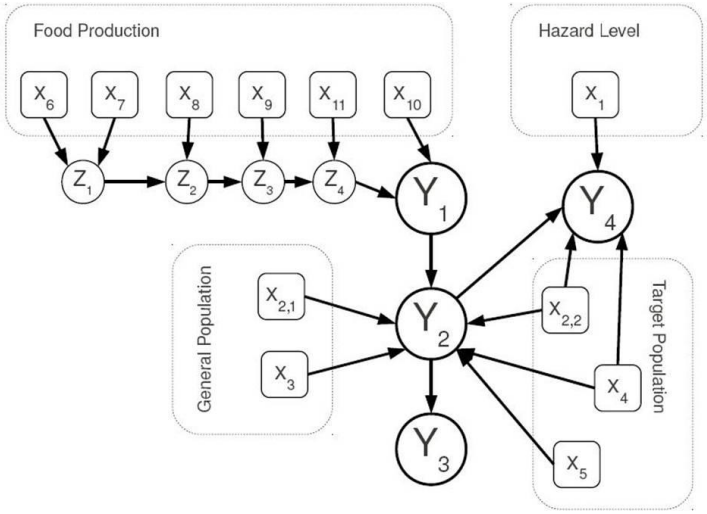


Figure 1: Structure of the Bayesian network associated to RR calculations

2.2 Description of the output nodes

Four outputs are produced by RR. Let us denote them by Y_1, Y_2, Y_3 and Y_4. Y_1 is the probability that a serving contains a dose of pathogen that would lead to illness it is defined as $min(1, \max(X_6X_7, X_8) \times X_9X_{10}X_{11})$.

Y_2 is defined as the ''*probability of illness per consumer per day*'' and is calculated as $min(1, Y_1X_{2,1}X_3)$.

Y_3 is the ''*total predicted illnesses/annum in population of interest*'' given by $365 \times Y_2X_{2,2}X_4X_5$

The last is the global risk score: $Y_4=100+\log_{10}(Y_2X_{2,2}X_4X_1)/17.56$. Y_4 is scaled between 0 and 100. A risk score of zero corresponds to a probability of illness equal to one case per 10 billion people per 100 years. For a risk score of 100, all the population is considered to eat daily a food that contains an illness-causing dose. A change of 6 in Y_4 corresponds to a tenfold difference in the absolute risk.

3. The Probabilistic Risk Ranger

3.1 From a unique value to a distribution of values for input nodes

In Risk Ranger, for a given question of the chain, the user is asked to propose a characteristic value, noted *x*. We propose now that the user gives four values to derive the probability distribution of the random variable *X*. These four quantities are q_l, q_u the assessed quantities for standard quantiles of *X* associated to known probability levels (α_l and α_u) ; and d_l, d_u the assessed associated degrees of confidence of the user in her/his assessment about the variability quantiles. We assume that it can vary from 1 (poor confidence) to 10 (perfect confidence).

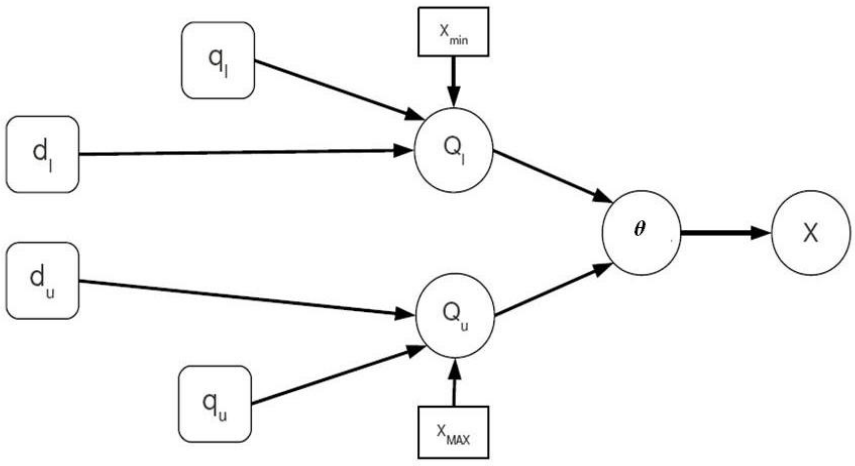


Figure 2: Graphical model illustrating the conditional dependencies between the four values given by the expert q_l, q_u, d_l, d_u and the random variable X. *d* stand for degree of confidence, *q* for assessed quantiles. X_{min}, X_{max} define the range of the random variable *X*.

Figure 2 displays the graph associated to the modeling for each question. For now, X is supposed to follow a Beta distribution defined on the support X_{min}, X_{max} and depending on two parameters ($\boldsymbol{\theta}$) which can be retrieved from two quantiles Q_l and Q_u (in practice, we propose to use $\alpha_l = 0.25$ and $\alpha_u = 0.75$ probabilities, but any couple of values can be used). The numerical procedure developed by van Dorp & Mazzuchi (2000) was used to obtain the two parameters of the Beta distribution given from two quantiles.

We also add uncertainty using the degrees of confidence given by the expert. The uncertainty on the two variability quantiles, Q_l and Q_u is modeled with a Uniform distribution, U(Q_{l1}, Q_{l2}) and U(Q_{u1}, Q_{u2}) respectively. The bounds for both Uniform distributions are calculated as follow:

$$Q_{l1} = q_l - (1-d_l/10) \cdot (q_l - X_{min}) \text{ and } Q_{l2} = q_l + (1-d_l/10) \cdot (q_m - q_l)$$
$$Q_{u1} = q_u - (1-d_u/10) \cdot (q_u - q_o) \text{ and } Q_{u2} = q_u + (1-d_u/10) \cdot (X_{max} - q_u)$$
$$\text{with } q_o = (q_u + q_l)/2$$

It must be underlined that when the expert chooses $q_l = q_u$ and $d_l=d_u=10$ then X is fixed and corresponds to x of the former version of Risk Ranger. When $q_l < q_u$ and $d_l=d_u=10$ then X represents only variability. When $q_l < q_u$ and d_l or $d_u < 10$ then X represents expert uncertainty and variability. An illustration of the expert uncertainty construction is shown on Figure 3.

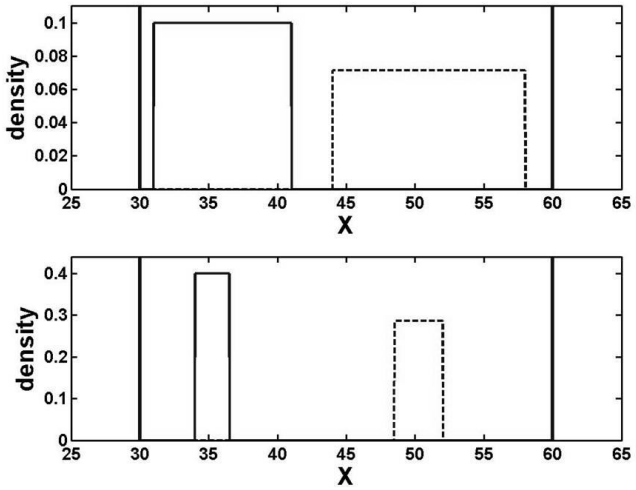


Figure 3: Example of uncertainty construction on Q_l (solid line) and Q_u (dashed line). Expert given quantiles q_l=35 q_u=50. X_{min}=30 and X_{max}=60. Upper graph $d_l=d_u$=2, bottom graph $d_l=d_u$=8.

3.2 The Excel worksheet of Probabilistic Risk Ranger: an interactive tool

The widespread use of Risk Ranger can be explained by its implementation in Excel. Indeed experts, who are generally food microbiologists are more likely to use a tool developed in a friendly environment. To keep this advantage we decided to develop the Probabilistic Risk Ranger in VBA Excel. The new tool allows the expert to check graphically, almost instantly, the uncertainty and variability of the elicited variable and then to interactively modify it according to the coherence between his/her opinion and that he/she sees on the graph. Figure 4 presents a screen shot of the new tool.

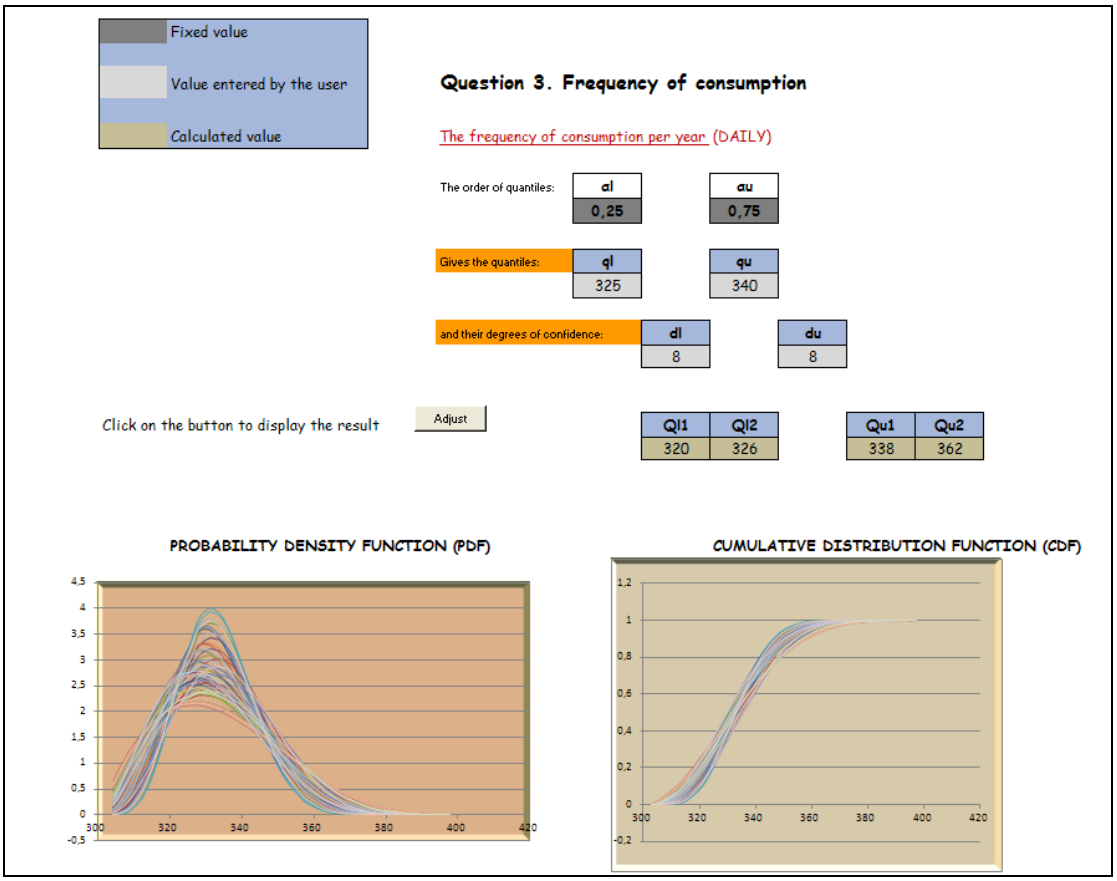


Figure 4: Screen shot of Question 3 worksheet of the Probabilistic Risk ranger. In the two diagrams, following the 2D simulation practice, the probability distributions are associated to the variability and their diversity to the uncertainty.

4. Conclusion

According to the level of the Risk score (Y_4), Sumner & Ross (2002) defined threshold values that help to define the importance of a food/hazard pair. For risk scores below 32, the food/hazard pair is considered as not significant. For scores above 48, the food/hazard pair is thought to present a major concern for public health.

In full deterministic version of Risk Ranger (unique values for inputs), the interpretation of the score is ambiguous as it is difficult to know is the expert has given a median estimate or an extreme one (worst-case scenario) for the different inputs.

With the Probabilistic Risk Ranger, we propose to use two dimensional (or second-order) Monte-Carlo simulations in which variability and uncertainty in the risk scores will appear. This will help the risk manager to take its decision.

Bibliography

Davidson, V. J., Ryks, J., & Fazil, A. (2006). Fuzzy risk assessment tool for microbial hazards in food systems. *Fuzzy Sets and Systems, 157*(9), 1201-1210.

Guillier, L., Thébault, A., Gauchard, F., Pommepuy, M., Guignard, A., & Malle, P. (2011). A risk-based sampling plan for monitoring of histamine in fish products. *Journal of Food Protection, 74*(2), 302-310.

Mataragas, M., Skandamis, P.N., & Drosinos, E.H. (2008). Risk profiles of pork and poultry meat and risk ratings of various pathogen/product combinations. *International Journal of Food Microbiology, 126*(1-2), 1-12.

Ross, T., and Sumner, J. (2002). A simple, spreadsheet-based, food safety risk assessment tool. *International Journal of Food Microbiology, 77*(1-2), 39-53.

Sosa Mejia, Z., Beumer, R. R., & Zwietering, M. H. (2011). Risk evaluation and management to reaching a suggested FSO in a steam meal. *Food Microbiology, 28*(4), 631-638.

Sumner, J., and Ross, T. (2002). A semi-quantitative seafood safety risk assessment. *International Journal of Food Microbiology, 77*(1-2), 55-59.

van Dorp, J. R., & Mazzuchi, T. A. (2000). Solving for the parameters of a beta distribution under two quantile constraints. *Journal of Statistical Computation and Simulation, 67*(2), 189-201.

Impact des incertitudes et des procédures d'estimation liées à la construction d'un modèle de microbiologie prévisionnelle sur la conformité avec un objectif de sécurité, dans une large gamme de conditions de conservation

Impact of uncertainties and estimation procedure inherent to predictive microbiology model construction on compliance with a food safety objective within a large range of preservative conditions

Laure Pujol[1,2], Sandrine Guillou[2,1] & Jeanne-Marie Membré[1,2]

[1] INRA, UMR1014 Secalim, Nantes, F-44307, France
[2] LUNAM Université, Oniris, Nantes, F-44307, France
E-mail : *laure.pujol@oniris-nantes.fr*

Résumé
La possibilité de réutiliser des données incomplètes est utile pour les acteurs de l'industrie agro-alimentaire évaluant l'effet combiné de conservateurs sur le développement de microorganismes pathogènes ou d'altération. A cet effet, l'impact des incertitudes et des procédures d'estimation sur la construction de modèles prédictifs avec de telles données a été évalué.
Des données de *Listeria monocytogenes* ont été simulées. Les incertitudes prises en compte ont été l'erreur de mesure et le défaut d'ajustement des modèles primaire et secondaire. L'estimation a été réalisée avec un traitement de données « simultané » ou « séquentiel », et selon trois procédures statistiques : inférence fréquentiste (critère des moindres carrés ou maximum de vraisemblance) et inférence Bayésienne.
Les résultats sont illustrés par le temps nécessaire à l'obtention de 100 ufc/g de *L. monocytogenes* dans un aliment. La méthode séquentielle associée à l'approche fréquentiste (maximum de vraisemblance, sous R) donne des résultats prometteurs.

Mots-clés : Incertitude, estimation de paramètres, microbiologie prévisionnelle, sécurité des aliments

Abstract
Re-using existing disparate data is particularly crucial for the food industry as it enables to save time when assessing the effect of preservative systems on pathogenic and/or spoilage microorganisms. In this context, the impact of uncertainties and estimation procedure inherent to predictive model construction with such data was assessed.
Various independent datasets of *Listeria monocytogenes* growth were generated by simulation. The uncertainties were the experimental error and the primary and secondary model (with or without interaction) lack-of-fitting. The estimation procedure issues were the data set management ("sequential" and "simultaneous"

estimation) and the statistical approaches (frequentist with least square or maximum likelihood criteria, Bayesian inference with MCMC).
The model output was expressed as the time to achieve 100 cfu/g of *L. monocytogenes* in food. An impact of secondary model uncertainty (under stressful conditions) was observed. The sequential dataset management associated with the maximum likelihood estimation procedure deployed in R seems valuable.

Keywords: Model uncertainty, estimation procedure, predictive microbiology, food safety

1. Introduction

The European Union regulation states that predictive mathematical modelling can be used by food business operators as one of the studies to investigate compliance with microbiological criteria throughout shelf-life. The modular models such as the gamma-type models developed by Zwietering et al. (Zwietering et al. 1992) allow the quantification of individual and combined preservative factors or hurdle effects on the bacterial growth rate. Subsequently, different combinations of hurdles (formulations) can be compared to each other to derive inhibitory effect equivalences, namely 'iso-hurdle rules'. A methodology to suggest and validate iso-hurdle rules has been recently developed (Pujol et al. 2012). It consisted in i) developing a predictive model based on existing but disparate datasets; ii) building an experimental design focused on the iso-hurdles using the model output; and iii) validating the model and the iso-hurdle rules with new data.

The first step, aiming to develop a predictive model re-using existing data is particularly crucial, as it enables to the food industry and other groups generating predictive models to save time and money when assessing the effect of preservative systems on pathogenic and/or spoilage microorganisms. Food companies generate and accumulate data on growth of microbiological contaminants associated with their food production, whether the purpose is for safety or quality checks. However, such data are often disparate in the sense that different data sets have been generated for different and specific purposes. For example one data set may have been generated to investigate the effects of temperature, pH and water activity (a_w) on a given microorganism. Another data set may have been generated to investigate the effects of temperature, pH and acetic acid on the same microorganism. How to build a model describing the effects of temperature, pH, a_w and acetic acid together on this microorganism?

In this context, the objective of our study was to determine the impact of uncertainties and estimation procedure inherent to predictive model construction with existing disparate data. Based on a full factorial design, various datasets were generated by simulation to reproduce *Listeria monocytogenes* growth on ambient stable product. This microorganism was chosen on the basis that it remains a serious threat for the safety of ready-to-eat foods. Assessing the accuracy of *L. monocytogenes* growth model is also important in a risk-based food safety management context, as an acceptable Food Safety Objective (FSO) of 100 cfu /g in ready-to-eat food products in Europe has been set. For potential use of results in an ambient stable dressing type product, the explanatory factors studied were temperature, pH, a_w, acetic, lactic and sorbic acids.

2. Materials and methods

2.1 Simulated data generation

First of all, various dataset were generated by simulation to reproduce *L. monocytogenes* growth rates. Five explanatory factors were incorporated in the experimental design: temperature, pH, a_w, acetic, lactic and sorbic acid. In the literature or in database, there is no study in which these factors

were studied all together; consequently, we have decided not to generate a full factorial design with five factors and repetitions but to create six dataset to explain each factor independently using a full factorial design for each. In the first one, the temperature (2, 5, 20, 35 and 40 °C) was study on its own with six repetitions (30 data), secondly the temperature (2, 5, 30 and 40°C) and pH (4.3, 5, 6.5, and 7.5) were considered with two repetitions (32 data). For the four other factors, the temperature and the pH were combined with a_w, sorbic, lactic and acetic acid as indicated in Table 1.

Factor studied and its levels	**Temperature levels associated to the factor studied (°C)**	**pH levels associated to the factor studied**
a_w : 0.94 ; 0.96 ; 0.99	1 ; 22 ; 30	5 ; 5.5 ; 6.5
Sorbic acid: 0.05 ; 0.1 ; 0.15 mM	8 ; 14 ; 24	5.2 ; 5.8 ; 6.4
Lactic acid: 0.5 ; 1 ; 3 mM	6 ; 12 ; 18	5 ; 5.5 ; 6
Acetic acid: 0.4 ; 0.8 ; 1.2 mM	12 ; 22 ; 32	5.5 ; 6 ; 6.5

Table 1: Temperature and pH levels for the studied factors and their levels

Once the experimental design built, growth rates were computed using the general gamma-type model (Zwietering et al. 1992) with an interaction term (Le Marc et al. 2002):

$$\sqrt{\mu_{\max}} = \sqrt{\mu_{opt} \times \gamma(T) \times \gamma(pH) \times \gamma(aw) \times \gamma(acids) \times \xi(T, pH, aw, acids)} \quad \text{(Eq1.)}$$

The cardinal model for temperature, pH and a_w, associated with the separate effect of the weak acid model (Coroller et al. 2012) were used to describe each γ term in Eq. 1.

To mimic the *L. monocytogenes* growth in ambient stable product, the cardinal values were set at realistic values, provided from an analysis of existing dataset. The values were set to -6.6, 37 and 45 °C for T_{min}, T_{opt} and T_{max} respectively, 3.9, 7, 9.4 for pH_{min}, pH_{opt} and pH_{max} and 0.90, 0.997 and 1 for a_{wmin}, a_{wopt} and a_{wmax} (Coroller et al. 2012).

The sorbic and lactic acid Minimum Inhibitory Concentrations (MIC) were also taken from Coroller et al. 2012. On the other hand, the acetic MIC was derived from Le Marc et al. 2002 considering a ratio of 2.5 between lactic and acetic MIC. The MIC values were 13mM, 8mM and 20mM for sorbic, lactic and acetic, respectively.

2.2 Uncertainty analysis

The uncertainties were the experimental error, the primary model and secondary model (with or without interaction) lack-of-fitting.

The experimental error was set to 0.5 log cfu/ml along the logcount curve, constant whatever the environmental condition. The logistic model with delay (Kono 1968) was set to derive the logcount value from the simulated growth rates (no lag, $\log(N_{max}) = 9$, $\log(N_0) = 3$, time scale from 0 to 500h, step of 2h), and to re-estimate the growth rates once the experimental error was applied to the logcount curves.

Since the data were generated with the gamma-model with interaction (Eq. 1), this model was used as benchmark versus two other secondary models. First, the gamma-model without interaction (Eq. 1 without the interaction term) was included in the uncertainty analysis; second, a simpler form of the cardinal model to be used in sub-optimal conditions was tested. The latter model named "Dalgaard", in our study, derives from Mejlholm and Dalgaard, 2007 (Mejlholm and Dalgaard 2007). It differs by the γ(T), γ(pH) and γ(a_w) terms. They were written as indicated below:

$$\gamma(T) = \begin{cases} \left[\dfrac{(T - T_{\min})}{(T_{opt} - T_{\min})}\right]^2 & \text{if } T > T_{\min} \\ 0 & \text{if T} \le T_{\min} \end{cases} \quad \text{(Eq2.)}$$

$$\gamma(pH) = \begin{cases} 1 - 10^{pH_{\min} - pH} & \text{if } pH > pH_{\min} \\ 0 & \text{if pH} \le pH_{\min} \end{cases} \quad \text{(Eq3.)}$$

$$\gamma(X) = \begin{cases} \dfrac{(a_w - a_{w\min})}{(a_{w\,opt} - a_{w\min})} & \text{if } a_w > a_{\min} \\ 0 & \text{if } a_w \le a_{w\min} \end{cases} \quad \text{(Eq4.)}$$

2.3 Estimation procedure analysis

Two different dataset management procedures, namely "sequential" and "simultaneous", were carried out. The "sequential" method consisted of estimating the effect of each gamma term successively, using its associated subset of data. The "simultaneous" method meant that all model parameters were estimated simultaneously using the whole datasets.

In term of statistical approach, a frequentist inference approach was run either with the least square criterion (deployed in Excel using the add-in Solver) or with the maximum-likelihood criterion (R software with nls function and nlstools package). These two methods were compared with a Bayesian inference approach, run through a Markov Chain Monte Carlo method, implemented in WinBugs. The prior distributions of the parameters were considered as relatively informative as food microbiologists have often a reasonable knowledge of the pathogen and/or spoilage microorganisms contaminating their products. Prior distributions are given in Table 2.

Parameter	**Distribution**	**Mean**	**Distribution Percentile**	
			2.5%	**97.5%**
$In\mu_{opt}$	dnorm(0.01, 0.01)	/	/	/
μ_{opt}	$\exp(In\mu_{opt})$	1	3.07	3.20×10^{8}
T_{min}	dnorm(-1.5, 1)	-1.5	-3.46	-0.46
T_{opt}	dnorm(37, 1)	37	35.04	38.96
pH_{min}	dnorm(4.2, 10)	4.2	2.24	6.16
pH_{opt}	dnorm(7, 10)	7	5.04	8.96
aw_{min}	dnorm(0.92, 10)	0.92	0.301	1.539
MIC_{sorbic}	dnorm(7, 1)	7	5.04	8.96
MIC_{lactic}	dnorm(7, 1)	7	5.04	8.96
MIC_{acetic}	dorm(22, 1)	22	20.04	23.96
σ	dunif(0, 10)	5	0.25	9.75

Table 2: Prior distributions of the parameters as used in the Bayesian inference

2.4 Model output

The uncertainty and estimation procedure analysis was performed by calculating the time to achieve 100 cfu/g (t_{100}), this latter value being set as an acceptable Food Safety Objective (FSO) for *L. monocytogenes* in ready-to-eat food products in Europe. t_{100} was derived from a loglinear model, i.e. without including any lag time and considering an initial bacterial load at 1 cfu/g.

The t_{100} values were calculated at various levels of inhibition ($\prod \gamma(.)$): 0.05, 0.1, 0.2, 0.3 and 0.5. The value of 0.05 corresponds to stressful inhibition conditions (growth rates estimated to be the 5th of its optimum) while the value of 0.5 corresponds to a moderate inhibition (half of the optimum growth rate). Each level of inhibition (each $\prod \gamma(.)$ level) was achieved by a combination of three preservative factors (among temperature, pH, a_w, acetic, lactic and sorbic acids), the remaining three being set to their optimum. The use of three factors in combination was chosen to mimic a food product stability due to a relatively complex hurdle preservative system. In total for each level of inhibition, twenty t_{100} values were generated (C_6^3), corresponding to twenty product formulations at identical level of inhibition, i.e. twenty iso-hurdles.

3. Results and Discussion

The impact of uncertainty and estimation procedure on the model parameter values is reported in Table 3. The growth rates generated by running the secondary model with interaction (Eq. 1) on the full factorial design datasets (Table 1) were used as response in a model, using the same secondary model (Eq. 1) but either the "simultaneous" or the "sequential" method. Results are presented in Table 3, row 1 and 2. Note that it was not possible to run the Bayesian inference method with the "simultaneous" method (systematic WinBugs error). The estimated parameter values are very close to target values, whatever the dataset management method and the statistical approach, showing that the estimation process error is negligible when substantial datasets are generated.

In addition, any significant difference on either parameter estimates values was noticed when a logcount error of 0.5 was applied to the data (Table 3, row 3), indicating that the consequence of an hypothetical primary model lack-of-fitting was negligible, with our datasets.

In terms of statistical approach and software used, reasonable results were obtained with Excel software (Solver add-in), which is an easy-to-use software, particularly for non-modellers. The software R, running a frequentist approach with the maximum-likelihood criterion (nls function and nlstools package) enabled to calculate the 95% confidence interval (and also the parameter correlation matrix, data not shown). That provides a more comprehensive piece of information than the only parameter estimates (Excel) when analyzing non-linear model outputs. This is definitively valuable when the predictive models are built on a limited and somewhat disparate dataset. In the Bayesian inference method, the impact of the prior distributions, although relatively informative, did not seem to be major. That is an important point to bear in mind when seeking food microbiologists, expertise on the pathogen and/or spoilage microorganism behaviour. The gamma model with interaction (Eq. 1) could not be run properly with the whole dataset ("simultaneous" method) with WinBugs ; that was likely to be due to the complexity of the mathematical expression of this interaction term. On the other hand, the estimation was perfectly correct with the "sequential" method.

The "sequential" method, assessed the effect of each preservative factor using only the subset of data in which this factor was studied; the "simultaneous" method, assessed the effect of the preservative factors altogether using the whole set of data. The first advantage of the "sequential" method lies in its simplicity in deployment: as the gamma model terms are estimated successively, running the fitting procedure is definitively straightforward. For instance, in our case-study, the Bayesian method could be deployed only with the "sequential" method. Also, generally in literature, the dataset associated with food-generic factors (e.g. temperature and pH) are more numerous than datasets dealing with food-specific factors such as organic acid or atmosphere modification. For these

two reasons, and as there was no difference between the model parameter values obtained from these two dataset management methods, one might recommend, when possible, to carry out a "sequential" method. This recommendation should come along a caveat on the data quality. Our case study is built on a combination of full factorial designs; however the "sequential" method might have some drawbacks when the dataset is of poor quality.

Model	Parameter	Target value	Frequentist				Bayesian inference		
			Least square criteria (Excel)	Maximum likelihood criteria (R)			MCMC (WinBugs)		
				Mean	Confidence interval		Mean	Credibility interval	
					2.50%	97.50%		2.5%	97.5%
« Simultaneous »	T_{opt}	37	37.0	37.0	37.0	37.0	/	/	/
	T_{min}	-6.6	-6.59	-6.60	-6.60	-6.60	/	/	/
	pH_{opt}	7	7.02	7.00	7.00	7.00	/	/	/
	pH_{min}	3.9	3.90	3.90	3.9	3.900	/	/	/
	aw_{min}	0.9	0.900	0.900	0.900	0.900	/	/	/
	CMI_{sorbic}	13	13.4	13.0	13.0	13.0	/	/	/
	CMI_{lactic}	8	8.21	8.00	8.00	8.00	/	/	/
	CMI_{acetic}	20	20.1	20.0	20.0	20.0	/	/	/
	μ_{opt}	1.5	1.50	1.500	1.50	1.50	/	/	/
	σ	/	0.001	0.000			/	/	/
« Sequential »	T_{opt}	37	37.0	37.0	37.0	37.0	37.0	37.0	37.0
	T_{min}	-6.6	-6.60	-6.60	-6.60	-6.60	-6.59	-6.60	-6.58
	pH_{opt}	7	7.00	7.00	7.00	7.00	7.0	7.0	7.0
	pH_{min}	3.9	3.90	3.90	3.90	3.90	3.90	3.90	3.90
	aw_{min}	0.9	0.900	0.900	0.900	0.900	0.900	0.900	0.900
	CMI_{sorbic}	13	13.0	13.0	13.0	13.0	13.0	13.0	13.0
	CMI_{lactic}	8	8.00	8.00	8.00	8.00	8.01	8.01	8.02
	CMI_{acetic}	20	20.0	20.0	20.0	20.0	20.0	20.0	20.0
	μ_{opt}	1.5	1.50	1.50	1.50	1.50	1.50	1.50	1.50
	σ	/	0.000	0.000			0.000	0.000	0.000
« Experimental error »	T_{opt}	37	37.0	37.0	37.0	37.1	/	/	/
	T_{min}	-6.6	-6.60	-6.60	-6.66	-6.54	/	/	/
	pH_{opt}	7	7.01	7.00	6.98	7.02	/	/	/
	pH_{min}	3.9	3.90	3.90	3.89	3.91	/	/	/
	aw_{min}	0.9	0.900	0.900	0.899	0.901	/	/	/
	CMI_{sorbic}	13	13.1	13.00	12.8	13.2	/	/	/
	CMI_{lactic}	8	8.00	7.99	7.90	8.09	/	/	/
	CMI_{acetic}	20	20.0	20.0	19.8	20.1	/	/	/
	μ_{opt}	1.5	1.50	1.50	1.50	1.50	/	/	/
	σ	/	0.004	0.004			/	/	/
« Dalgaard »	T_{opt}	37	53.395	/	/	/	/	/	/
	T_{min}	-6.6	-9.930	/	/	/	/	/	/
	pH_{min}	3.9	4.198	/	/	/	/	/	/
	aw_{min}	0.9	0.903	/	/	/	/	/	/
	CMI_{sorbic}	13	12.799	/	/	/	/	/	/
	CMI_{lactic}	8	6.861	/	/	/	/	/	/
	CMI_{acetic}	20	23.232	/	/	/	/	/	/
	μ_{opt}	1.5	2.649	/	/	/	/	/	/
	σ	/	0.055	/			/	/	/
« Without ξ »	T_{opt}	37	36.7	36.7	36.3	37.06	36.61	36.25	37
	T_{min}	-6.6	-5.59	-5.59	-6.02	-5.16	-5.46	-5.88	-5.05
	pH_{opt}	7	6.93	6.93	6.82	7.04	6.93	6.83	7.04
	pH_{min}	3.9	4.00	4.00	3.94	4.05	3.99	3.93	4.04
	aw_{min}	0.9	0.903	0.903	0.899	0.907	0.903	0.900	0.907
	CMI_{sorbic}	13	9.46	9.45	8.27	10.6	8.99	8.14	9.93
	CMI_{lactic}	8	7.43	7.42	6.55	8.29	7.16	6.65	7.45
	CMI_{acetic}	20	18.8	18.8	17.8	19.7	19.3	18.43	20.2
	μ_{opt}	1.5	1.52	1.52	1.49	1.55	1.52	1.50	1.55
	σ	/	0.031	0.031			0.031	0.029	0.034

Table 3: Parameter estimates obtained with the different model and statistical procedures.

Two secondary models were tested as alternative to the Eq.1. Both are simpler, either because they contain less parameters being applied only in sub-optimal conditions (model referenced as "Dalgaard") or because they do not include the interaction term (model referenced as "without ξ" in Table 3).

The "Dalgaard" model was not successfully deployed with our data set (row 4 in Table 3). Only Excel provided results, however even these results were not robust (estimation procedure issue: local minimum depending on initial parameter values). Parameter estimates might be considered as still relevant, for example T_{min} of -9.9 instead of -6.6 might be considered as acceptable. To pinpoint the consequence of this model uncertainty on potential application in food safety, the time to achieve 100 cfu/g (t_{100}) of *L. monocytogenes* was calculated for five iso-hurdles (each one based on 20 combinations of preservative factors). Results are depicted in Fig. 1. The time to achieve 100 cfu/g varied from 0 to 479 h in stressful conditions (target value 61h), showing a huge model output uncertainty due to secondary model lack-of-fitting.

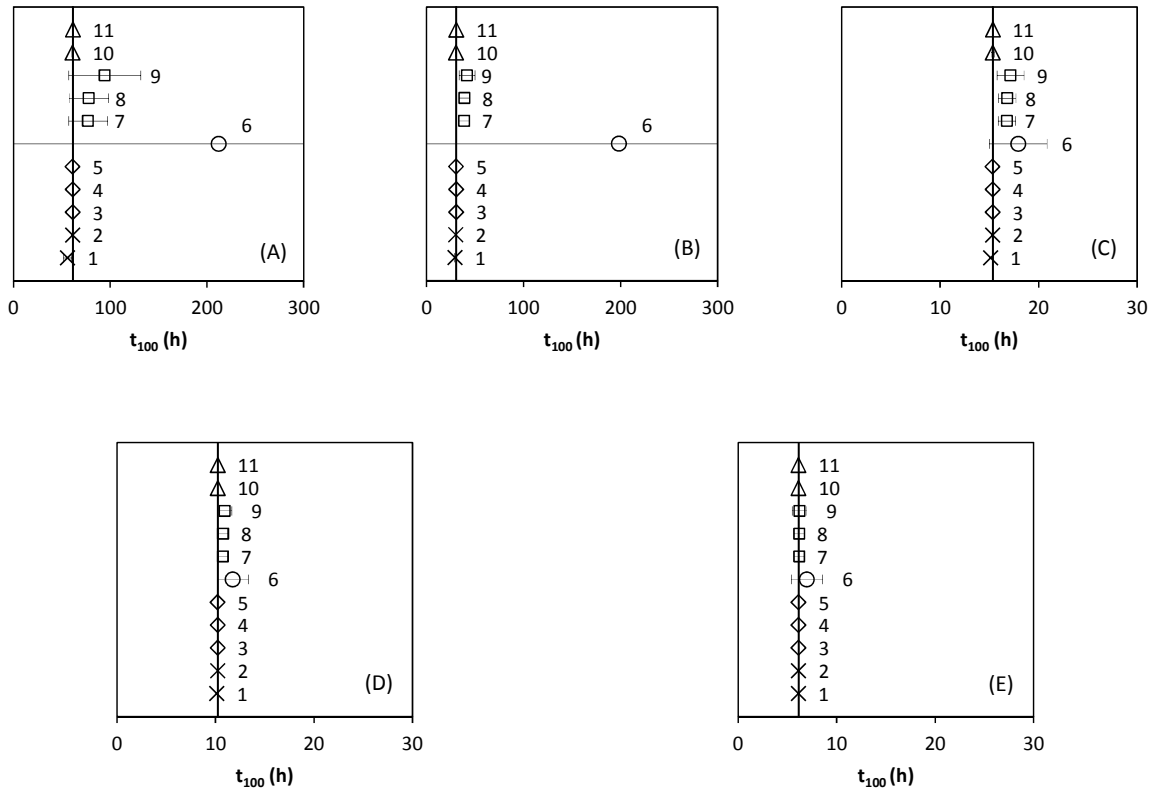


Figure 1 : Time to achieve 100cfu/g for each estimation procedure tested (uncertainty due to "simultaneous" method run with Excel (1) and R (2); uncertainty due to "sequential" method run with Excel (3), R (4) and WinBugs (5); uncertainty due to "Dalgaard" model run with Excel (6); uncertainty due to Eq.1 "Without ξ" run with excel (7), R (8) and WinBugs (9); uncertainty due to experimental error run with excel (10) and R(11)) for different inhibition tested (A: γ=0.05, B: γ=0.1, C: γ=0.2, D: γ=0.3, E: γ=0.5)

On the other hand, with the second simpler alternative ("without γ" model), the findings were worth being scrutinized in details. First of all, with the whole dataset used in one estimation step ("simultaneous" method), the three statistical approaches provided interpretable results, even the Bayesian inference run in WinBugs (whilst this latter one did not work properly in the case of Eq. 1 with interaction). The parameter estimates are different from the target, for instance the acetic CMI is estimated to 18.8, 18.8 and 19.3 mM with Excel, R and WinBugs respectively, while the target was 20 mM. That is due to removing the interaction term: the inhibitory effect has to be derived from each γ term (each factors has to be more inhibitor on its own), instead of the additional γ, leading to a stronger individual-factor inhibitory effect.

This change in parameter estimate has a direct consequence on food safety application. For stressful conditions, i.e. $\prod\gamma(.) \leq 0.2$, the impact of secondary model lack-of-fitting is significant. That is also illustrated in Fig. 2 where t_{100} is depicted as function of the iso-hurdles, for the Eq. 1 model with or without interaction.

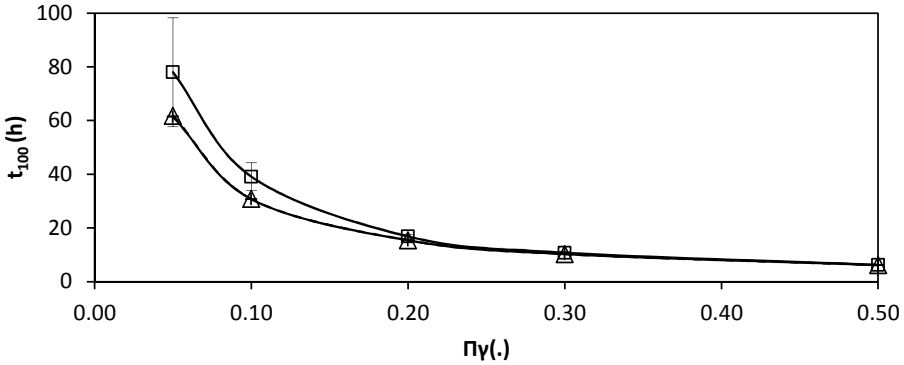


Figure 2: Time to achieve 100 cfu /ml for various iso-hurdle ($\prod\gamma(.)$=0.05 to 0.5) reported with mean and standard deviations (20 combinations of preservative factor levels for each iso-hurdle); + targeted time; Δ uncertainty due to experimental error run with R; □ uncertainty due to Eq1. "Without ξ" run with R

On the other hand, one might notice that the estimates provided by the Bayesian inference approach were slightly different from those obtained by the frequentist approach. These differences had no major consequence in term of food safety application as illustrated through the time to achieve 100 cfu/g (Fig. 1): for a given iso-hurdle ($\prod\gamma(.)$ constant), the results obtained with the three statistical approaches were similar.

The choice of the gamma-type secondary model for describing the effect of five or more environmental factors, especially the inclusion of the interaction term describing the interaction among environmental factors, is a topic of controversy. Recently, Biesta-Peters et al. (Biesta-Peters et al. 2010) studied the effect of pH and weak acid on *Bacillus cereus* growth and concluded that the addition of the interaction term ξ did not improve the prediction. On the other hand, when studying the effect of pH and water activity on *B. cereus* growth rates, the authors could not find a secondary model that gave fully satisfactory predictions, whether the interaction term was applied or not (Biesta-Peters et al. 2011). Nevertheless, the choice of the secondary model depends on the microbial stress(es) and the application. Indeed, if the model is planned to be used at conditions with low stress(es), e.g. in modelling spoilage organisms where some growth may be tolerated, a model without interaction might be sufficient. Conversely, if the model is planned to be used for the growth / no-growth area, a model with interaction seems to be much more appropriate (Le Marc et al. 2002, Mejlholm et al. 2010).

In conclusion, developing a predictive model re-using disparate dataset generate uncertainty and parameter estimation issues. The sequential dataset management procedure appeared to be relatively easy-to-be-implemented and appropriate to a food microbiology context where many data are generated for some generic factors (e.g. storage temperature) whilst only few for other ones. The maximum likelihood estimation run in R seems appropriate to perform secondary model comparison built on limited dataset (e.g. to scrutinize the model outputs with a comprehensive statistical toolbox). Overall, being able to generate robust predictive models on existing data will help the food industry and other groups involved in food safety management to save time and money when assessing the effect of a food preservative system.

Bibliography

Biesta-Peters, E. G., Reij, M. W., Gorris, L. G. M., & Zwietering, M. H. (2010). Comparing non-synergistic gamma models with interaction models to predict growth of emetic *Bacillus cereus* when using combinations of pH and individual undissociated acids as growth-limiting factors. *Applied and Environmental Microbiology,* 76**,** 5791-5801.

Biesta-Peters, E. G., Reij, M. W., Zwietering, M. H., & Gorris, L. G. M. (2011). Comparing non-synergy gamma models and interaction models to predict growth of emetic *Bacillus cereus* for combinations of pH and water activity values. *Applied and Environmental Microbiology,* 77**,** 5707-5715.

Coroller, L., Kan-King-Yu, D., Leguerinel, I., Mafart, P., & Membré, J.-M. (2012). Modelling of growth, growth/no-growth interface and nonthermal inactivation areas of *Listeria* in foods. *International Journal of Food Microbiology,* 152**,** 139-152.

Kono, T. (1968). Kinetics of microbial cell growth. *Biotechnology and Bioengineering,* 10**,** 105-131.

Le Marc, Y., Huchet, V., Bourgeois, C. M., Guyonnet, J. P., Mafart, P., & Thuault, D. (2002). Modelling the growth kinetics of *Listeria* as a function of temperature, pH and organic acid concentration. *International Journal of Food Microbiology,* 73**,** 219-237.

Mejlholm, O., & Dalgaard, P. (2007). Modeling and predicting the growth boundary of *Listeria monocytogenes* in lightly preserved seafood. *Journal of Food Protection,* 70**,** 70-84.

Mejlholm, O., Gunvig, A., Borggaard, C., Blom-Hanssen, J., Mellefont, L., Ross, T., Leroi, F., Else, T., Visser, D., & Dalgaard, P. (2010). Predicting growth rates and growth boundary of *Listeria monocytogenes* - An international validation study with focus on processed and ready-to-eat meat and seafood. *International Journal of Food Microbiology,* 141**,** 137-50.

Pujol, L., Kan-King-Yu, D., Le Marc, Y., Johnston, M. D., Rama-Heuzard, F., Guillou, S., McClure, P., & Membré, J.-M. (2012). Establishing equivalence for microbial-growth-inhibitory effects ("iso-hurdle rules") by analyzing disparate *Listeria monocytogenes* data with a gamma-type predictive model. *Applied and Environmental Microbiology**, In press*.

Zwietering, M. H., Wijtzes, T., de Wit, J. C., & van't Riet, K. (1992). A decision support system for prediction of the microbial spoilage in foods. *Journal of Food Protection,* 55**,** 973-979.

Les statistiques et la modélisation au service de la sécurité sanitaire des produits laitiers

Statistics and modeling for the microbiological safety of dairy products

Fanny Tenenhaus-Aziza[1], Valérie Michel[2], Hajer Souaifi[2], Frédérique Perrin[2,3], Moez Sanaa[3]

[1] *Centre National Interprofessionnel de l'Economie Laitière, 42 rue du Châteaudun, 75009, Paris*
E-mail : ftenenhaus@cniel.com
[2] *Actilait, Technopôle Alimentec, rue Henri de Boissieu, 01060, Bourg en Bresse cedex 9*
E-mail : v.michel@actilait.com; h.souaifi@actilait.com; f.perrin@actilait.com
[3] *ANSES, 27-31 av. du Général Leclerc, 94701, Maisons-Alfort cedex*
E-mail : moez.sanaa@anses.fr

Résumé
En cas de contamination, certains produits laitiers peuvent permettre le développement des bactéries pathogènes. La valorisation des données analytiques existantes, de la collecte du lait à la consommation, couplée aux connaissances sur les bactéries permet d'améliorer les mesures préventives et correctives vis-à-vis d'un problème sanitaire. La filière laitière a développé des outils permettant aux industriels d'optimiser la maîtrise de leur procédé de fabrication en termes de sécurité sanitaire : 1/ Outils d'analyses statistiques de données microbiologiques pour augmenter la qualité sanitaire du lait 2/ Outils de modélisation type Appréciation Quantitative des Risques, démarche reconnue et recommandée par les autorités sanitaires, pour estimer, à l'aide de la microbiologie prévisionnelle, les risques microbiologiques liés à un couple bactérie/matrice. L'utilisation de ces nouvelles approches par l'industrie agroalimentaire renforce l'importance des statistiques et de la modélisation comme outils d'aide à la décision.
Mots-clés : Appréciation quantitative des risques microbiologiques, analyse des données d'autocontrôles, analyse de tendance, produits laitiers

Abstract
In case of contamination, some dairy products can allow the growth of pathogenic micro-organisms. The statistical analysis of analytical data, from the milk collection to the product consumption, associated to the knowledge on bacteria's behavior can help optimizing preventive and corrective actions, toward a sanitary problem. The dairy field has developed some tools for stakeholders of the dairy sector in order to increase their process in terms of microbiological safety: 1/ Statistical tools for enhancing the quality of raw milk; 2/ Quantitative risk assessment modeling tools to assess the contamination level in the product all along the production chain. The quantitative risk assessment approach is recognized and recommended by safety authorities, at the national and international levels. The use of these methods by industrials reinforces the importance of statistics and modeling as decision-making tools.
Keywords: Microbiological quantitative risk assessment, statistical analysis of analytical data, tendency analysis, dairy products

1. Introduction

En cas de contamination, certains produits laitiers ont des caractéristiques qui peuvent permettre le développement des bactéries pathogènes. Le fabricant est garant de la salubrité de son produit envers le consommateur (articles 14 et 17 du règlement CE n°178/2002). Au quotidien, l'application des bonnes pratiques d'hygiène, le plan HACCP, le plan de maîtrise sanitaire et la traçabilité sont des outils indispensables pour prévenir les dangers et en limiter leur impact. A plus long terme, la valorisation des données analytiques existantes de la collecte du lait à la consommation, couplée aux connaissances sur les micro-organismes pathogènes permet d'optimiser les mesures préventives et correctives vis-à-vis d'un problème sanitaire. L'Appréciation Quantitative des Risques (AQR), démarche reconnue et recommandée par les autorités sanitaires, est l'outil répondant à cet objectif d'intégration des données et connaissances. Dans ce contexte, la filière laitière développe des outils permettant aux industriels d'optimiser la maîtrise de leur process en termes de sécurité sanitaire :
1/ Outils d'analyse de données pour optimiser la qualité sanitaire du lait : estimation de la prévalence de non-conformité microbiologique, au niveau « lait de producteurs » et au niveau « laits de mélange », analyse de tendances d'évolution des prévalences, simulation du tri du lait, mesure de l'impact des mesures de gestion (par exemple : saison, type d'alimentation).
2/ Outils de modélisation pour estimer le risque de contamination d'un produit fini, à l'aide de la microbiologie prévisionnelle : modèle de simulationspécifique à un couple bactérie pathogène/matrice fromagère permettant de tester des scénarios de contamination, des valeurs de process, d'estimer la prévalence de contamination en sortie usine et le risque consommateur, d'optimiser les plans de contrôle.

Le kit d'outils développés par le CNIEL, en collaboration avec Actilait, est transféré à l'équipe du Pôle Sanitaire d'Actilait, qui propose la réalisation de ce type d'étude aux professionnels du secteur laitier. Actilait accompagne plusieurs filières AOP dans cette démarche, ainsi que des industriels individuellement.

2. Analyse Quantitative des Données Laitières

L'outil d'Analyse Quantitative des Données Laitières (AQDL) permet de mettre en évidence des tendances dans l'évolution de la contamination microbiologique des laits, en tenant compte de l'échantillonnage et du nombre d'élevages présents dans une collecte. Il s'applique à tous les pathogènes ou indicateurs d'hygiène, sous réserve que des résultats d'analyses réguliers soient disponibles.

2.1 Matériel et méthodes

Les données nécessaires à l'analyse sont les résultats d'analyses microbiologiques par date, par élevage et par critère (Tableau 1), ainsi que le descriptif des mesures préventives ou correctives appliquées à l'échelle des élevages.

Tous les calculs sont implémentés sous SAS 9.1.3.
Pour tous les tests statistiques, le risque de première espèce est fixé à 5%.

Elevage	Date	Salmonelle	Staph300	Ecoli10
100005	14/06/2004	0	0	0
100005	18/06/2004	1	0	0
…				
100005	06/07/2004	0	0	1
100013	10/03/2005	0	0	0
100013	02/04/2005	0	1	1
100013	07/04/2005	1	0	0
100013	07/10/2005	0	1	1

Tableau 1. Exemple de jeu de données analysé avec l'outil AQDL. Critère Salmonelle : résultats de détection (Présence = 1 / Absence = 0) ; critères Staph300 et Ecoli10, respectivement : résultats de quantification pour *Staphylocoques aureus* et *Escherichia coli* respectivement : si le dénombrement est inférieur à 300 ufc/ml ou à 10 ufc/ml respectivement, alors Staph300 = 0 et Ecoli10 = 0 respectivement, 1 sinon

Le premier objectif est l'évaluation d'une tendance au cours du temps du pourcentage d'analyses non-conformes annuels (prévalence), tous producteurs confondus. On calcule la prévalence annuelle ainsi que son intervalle de confiance exact (utilisation de la loi binomiale). Des tests statistiques de comparaison de pourcentages annuelles sont effectués afin de :

- Mettre en évidence une éventuelle différence entre ces prévalences annuelles par un test exact de Fisher ou, si le nombre d'observation est trop élevé, un test du Khi-deux (test d'indépendance).
- Mettre en évidence une éventuelle évolution, à la hausse ou à la baisse, de la prévalence au cours du temps par un test de Cochran-Armitage (test de tendance).

Ensuite, un modèle de régression logistique de type GEE tenant compte des répétitions des observations par élevage sur l'année est appliqué aux données (Liang and Zeger, 1986). Il permet d'évaluer les risques relatifs entre les prévalences annuelles.
Le test du score généralisé (test de linéarité) permet de tester la régularité de l'évolution de la prévalence annuelle (Boos, 1992 ; Rotnitzky and Jewell, 1990). Si l'évolution est régulière, un risque relatif global est calculé sur l'ensemble des années, ainsi qu'une prédiction pour la période supplémentaire. Sinon, un risque relatif entre deux années consécutives est calculé.
Enfin, le test de Wald évalue si le(s) risque(s) relatif(s) calculé(s) est(sont) significativement différent(s) de 1 et donc, si la différence entre les prévalences de deux années consécutives est significative ou non. S'il n'y a pas de différence significative, une prévalence moyenne sur l'ensemble de la période est calculée.

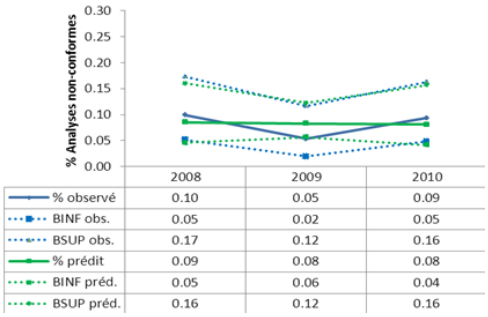
Le second objectif est de connaître la répartition des élevages suivant leur nombre de résultats d'analyses non-conformes annuel. On peut définir, à titre d'exemple, deux classes : classe n°1 si aucun résultat non-conforme dans l'année ; classe n°2 si au moins un résultat non-conforme dans l'année. Un test exact de Fisher ou de Khi-deux permet de comparer la répartition annuelle des élevages dans chaque classe et un test de Cochran-Armitage est effectué pour détecter une tendance.

2.2 Résultats

Dans cette section, nous présentons des exemples-type de résultats obtenus. Les analyses ont été réalisées sur des bases de données réelles, fournies par des entreprises de transformation laitière. Pour le critère *Listeria monocytogenes (L. monocytogenes)*, l'analyse est conforme s'il y a absence. Pour le critère Coliformes, l'analyse est conforme si le dénombrement est inférieur à 500 ufc/ml.

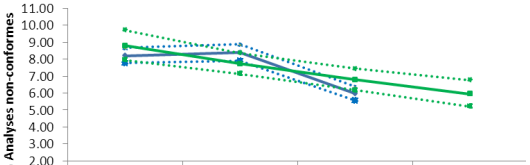
Pour *L. monocytogenes*, les tests d'indépendance et de tendance montrent qu'aucune prévalence annuelle ne diffère des deux autres et qu'il n'existe pas de tendance au cours du temps, lorsque tous les producteurs sont confondus. Le test de linéarité, tenant compte de la variabilité producteur, montre qu'il y a une évolution régulière dans le temps. Or, le risque relatif global calculé par le modèle n'est pas significativement différent de 1 d'après le test de Wald. Une prévalence globale a donc été calculée pour les trois années. La figure 3 montre l'évolution de la prévalence de *L. monocytogenes* et des Coliformes dans le lait entre 2008 et 2010.

Pour les Coliformes, les tests d'indépendance et de tendance montrent qu'il existe au moins une prévalence annuelle différente des deux autres et qu'il existe une tendance à la baisse au cours du temps. Le test de linéarité montre qu'il y a une évolution irrégulière dans le temps. Les risques relatifs ont donc été calculés pour chaque couple d'années. Pour le couple 2009/2008, il n'est pas significativement différent de 1. Une prévalence globale sur les deux années a donc été calculée. La figure 3 montre l'évolution de la prévalence de *L. monocytogenes* et des Coliformes dans le lait entre 2008 et 2010 et fournit une prédiction pour 2011. La figure 4 montre l'évolution de la prévalence pour le critère Coliformes dans le lait entre 2008 et 2010 et fournit une prédiction pour 2011.



	2008	2009	2010
% observé	0.10	0.05	0.09
BINF obs.	0.05	0.02	0.05
BSUP obs.	0.17	0.12	0.16
% prédit	0.09	0.08	0.08
BINF préd.	0.05	0.06	0.04
BSUP préd.	0.16	0.12	0.16

Figure 3. *L. monocytogenes* - Prévalences annuelles de non-conformité



	2008	2009	2010	2011
% observé	8.20	8.37	5.97	
BINF obs.	7.76	7.90	5.57	
BSUP obs.	8.66	8.87	6.39	
% prédit	8.79	7.72	6.78	5.94
BINF préd.	7.95	7.12	6.17	5.21
BSUP préd.	9.71	8.37	7.44	6.76

Figure 4. Coliformes - Prévalences annuelles de non-conformité et prévision pour l'année 2011

Années 2009/2008	Années 2010/2009
RR (prév. estimée%)	RR (prév. estimée%)
1/1.03 (0.08%) [1/1.73;1.63]	

Tableau 2. *L. monocytogenes* –

Risques relatifs et IC à 95%

Années 2009/2008	Années 2010/2009
RR (prév. estimée%)	RR (prév. estimée%)
1.03 (8.28%)	1/1.40

[1/1.08;1.14]	[1/1.59;1/1.23]

Risques relatifs et IC à 95%

Tableau 3. Coliformes –

Les figures 5 et 6 représentent la répartition annuelle des élevages suivant des classes de non-conformité. Pour le classement selon *L. monocytogenes* (0 analyse non-conforme par an ou au moins 1 analyse non-conforme par an), il n'y a pas d'évolution significative du pourcentage d'élevages dans la classe « absence de *L. monocytogenes* », qui reste stable autour de 99%. Pour le classement selon les Coliformes, le pourcentage d'élevages s'inscrivant dans la classe « 0 analyse non-conforme par an » augmente de manière significative au cours du temps.

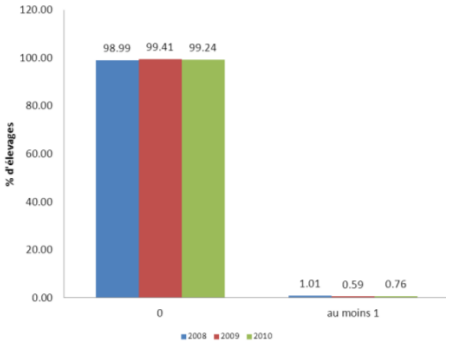


Figure 5. Répartition annuelle des élevages en fonction de classes de non-conformité en *L. monocytogenes*

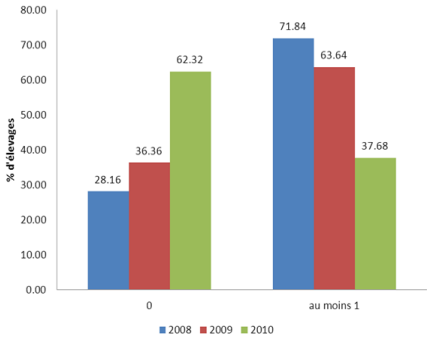


Figure 6. Répartition annuelle des élevages en fonction de classes de non-conformité en Coliformes

3. Modèle d'Appréciation Quantitative des Risques

Les modèles AQR permettent de simuler le comportement d'un germe indésirable tout au long du process, et ce jusqu'à la consommation, d'identifier les étapes à risque et d'optimiser les plans d'échantillonnage. Les résultats attendus sont : la probabilité d'effets néfastes, le pourcentage d'unités contaminées (fromage, portions de 25g), la distribution de la contamination dans un lot à chaque étape du process, le pourcentage d'unités respectant un objectif de sécurité sanitaire comme par exemple, 100 ufc/g pour *L. monocytogenes* au moment de la consommation, ou encore 10000 ufc/g au pic de concentration du process pour Staphylocoques à coagulase positive.

3.1 Matériel et méthodes

Pour construire ce type de modèle, les données nécessaires sont les paramètres technologiques du process, des résultats d'analyses microbiologiques (sur le lait cru, sur le produit en cours de process, sur le produit fini), et éventuellement des tests de croissance.

La microbiologie prévisionnelle est utilisée pour estimer les taux de croissance à partir de tests de croissance, et pour simuler l'évolution de la population microbienne (croissance, survie, latence) au cours du temps et en fonction des paramètres physico-chimiques. En particulier, les modèles primaires modélisent l'évolution de la population (ou de la concentration) bactérienne viable en fonction du temps. Le modèle primaire utilisé est le modèle logistique avec délai (Baranyi et al., 1993; Rosso, 1996), c'est-à-dire, avec un point de coupure entre la phase de latence et la phase exponentielle, sous sa forme différentielle (Equation 1).

$$\begin{cases} x(t) = x_0 & t = 0 \\ \dfrac{dx}{dt}(t) = 0 & \text{si } t \leq \lambda \\ \dfrac{dx}{dt}(t) = \mu.x(t).\left(1 - \dfrac{x(t)}{X_{max}}\right) & \text{si } t > \lambda \end{cases}$$

Equation 1. Modèle de croissance primaire logistique avec délai

Dans ce modèle, on considère un inoculum contenant x_0 micro-organismes, à l'instant $t = 0$. L'état physiologique des bactéries présentes dans l'inoculum détermine la durée de la phase de latence (λ) en heures.

Au cours de la phase exponentielle des divisions cellulaires se produisent, selon un taux ou une vitesse spécifique dit maximal de croissance (μ), en nombre de cellules sur une échelle logarithmique par unité de temps. Le paramètre μ dépend des conditions et de la matrice alimentaire contaminée. Les bactéries se multiplient jusqu'à atteindre une population ou une concentration maximale dans le milieu de culture (X_{max}), correspondant à la phase stationnaire du modèle. La valeur du paramètre X_{max} du modèle de croissance primaire varie en fonction du type de matrice alimentaire, solide ou liquide.

Le taux de croissance μ en fonction des paramètres physico-chimiques (notamment la température, l'activité de l'eau et le pH) est décrit par les modèles secondaires, en particulier les modèles cardinaux, présentant l'avantage d'être constitués de paramètres ayant un sens biologique (Le Marc, 2001). Le modèle utilisé tient compte des interactions entre les différents paramètres physico-chimiques (Augustin et al., 2005). Pour des valeurs de température, de pH et d'activité de l'eau données à l'instant t, respectivement T(t), pH(t) et aw(t), le modèle décrivant le taux de croissance à l'instant t, $\mu(t)$, est donné par l'équation 2.

$$\mu(t) = \mu_{opt} \times CM_2(T(t)) \times CM_1(pH(t)) \times SR_1(aw(t)) \times \xi(T(t), pH(t), aw(t)),$$

avec pour tout X et Y $\in \{T(t), pH(t), aw(t)\}$

$$\xi = \begin{cases} 1 & \text{si } \psi \leq 0{,}5 \\ 2 \times (1 - \psi) & \text{si } 0{,}5 < \psi < 1 \\ 0 & \text{si } \psi \geq 1 \end{cases}, \quad \psi = \sum_X \frac{\phi(X)}{2 \times \prod_{X \neq Y} (1 - \phi(Y))}, \quad \phi(X) = \left(\frac{X_{opt} - X}{X_{opt} - X_{min}}\right)^3$$

$$SR_n(X) = \begin{cases} 0 & \text{si } X < X_{min} \\ \left(\dfrac{X - X_{min}}{X_{opt} - X_{min}}\right)^n & \text{si } X_{min} < X < X_{max} \end{cases},$$

$$CM_n(X) = \begin{cases} 0 & \text{si } X \le X_{min} \\ \dfrac{(X - X_{max})(X - X_{min})^n}{(X_{opt} - X_{min})^{n-1}[(X_{opt} - X_{min})(X - X_{opt}) - (X_{opt} - X_{max})((n-1)X_{opt} + X_{min} - nX)]} & \text{si } X_{min} < X < X_{max} \end{cases}$$

Equation 2. Modèle de croissance secondaire proposé dans (Augustin et al., 2005)

Les paramètres X_{min}, X_{max} et X_{opt} sont les valeurs cardinales de croissance pour $X \in \{\text{Température}, X_{min}, X_{max}\}$: X_{min} et X_{max} sont les valeurs minimale et maximale de croissance, respectivement, et X_{opt} est la valeur du paramètre X à laquelle le taux de croissance μ est optimal. Les fonctions modulaires CM_1, CM_2, SR_1 représentent les effets relatifs des différents paramètres physico-chimiques sur le taux de croissance. L'effet combiné de ces facteurs est obtenu par le produit des effets séparés et de la fonction ζ modélisant l'interaction entre les paramètres (Le Marc, 2001).
Les valeurs des paramètres de microbiologie prévisionnelle peuvent provenir de la littérature scientifique ou être calculés à partir des bases de données en ligne (Sym'Previus, ComBase).

D'autres modèles sous forme de système dynamique peuvent être intégrés pour simuler les contaminations croisées en cours de process ou pendant la préparation chez le consommateur (Aziza et al., 2006).

Les simulations de Monte Carlo permettent de tenir compte de la variabilité et l'incertitude sur les différents paramètres (Vose, 2000). La probabilité d'effets néfastes est calculée en utilisant les modèles doses-réponses disponibles dans la littérature (Buchanan et al., 2000), couplée aux données de consommation sur le produit alimentaire d'intérêt.

Dans les résultats, nous présenterons un exemple d'évolution du niveau de contamination d'un produit, pour un process donnée depuis le lait cru mis en œuvre jusqu'au fromage à la sortie de l'usine.

3.3 Résultats

Nous examinons les résultats d'un modèle simulant la mise en œuvre d'un lait servant à fabriquer un fromage de type pâte pressée non cuite, contaminé à 10^{-5} ufc/ml. 100 itérations sont réalisées tenant compte de la variabilité de ces paramètres en cours de process. On obtient ainsi, pour un lot de fromages fabriqués, la distribution empirique de la concentration dans les portions de 25 grammes, ainsi que le pourcentage de portions contaminées.

La figure 5 montre l'évolution de la concentration au cours du process pour trois itérations, à titre d'exemple. La croissance se produit principalement pendant le saumurage et l'affinage. La concentration dans le produit sortie usine peut attendre 10^5 ufc/g. En moyenne, cette concentration est de 1 ufc/g.

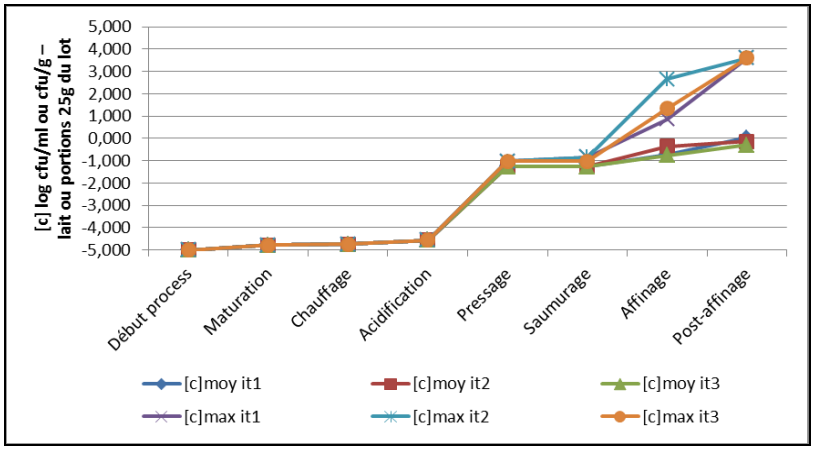


Figure 7. Evolution de la concentration au cours du process dans le lait et les portions de fromages de 25 grammes contaminées

Cette distribution concerne seulement 0,5% des portions de 25 grammes. De plus, entre 62 et 69% des fromages sont contaminés. Enfin 72 et 78% des portions de 25 grammes contaminées ont une concentration inférieure à 100 ufc/g.

3. Discussion

Les outils présentés sont évolutifs et répondent à une demande terrain. L'outil AQDL permet à ce jour aux industriels laitiers d'identifier des leviers de maitrise sur la base de résultats scientifiquement validés. En AQR, les modèles permettent d'identifier les étapes à risque dans un process mais peuvent encore être améliorés, notamment en ce qui concerne la décroissance athermique. La facilité d'accès aux simulations est une issue importante pour les industriels. La filière laitière a donc développé une interface de simulation des modèles (www.aqr.maisondulait.fr). Enfin, les résultats de simulations nécessitent aussi des traitements statistiques spécifiques, telle que l'analyse de sensibilité. Ces perspectives sont aujourd'hui en cours de développement.

Bibliographie

Augustin, J. C., Zuliani, V., Cornu, M. and Guillier, L. (2005). Growth rate and growth probability of *Listeria monocytogenes* in dairy, meat and seafood products in suboptimal conditions. *Journal of Applied Microbiology*, 99, 1019-1042.

Aziza, F., Mettler, E., Daudin, J.-J. and Sanaa, M. (2006). Stochastic, compartmental, and dynamic modeling of cross-contamination during mechanical smearing of cheeses. *Risk Analysis*, 26, 731-745.

Baranyi, J., Roberts, T. A. and McClure, P. (1993). A non-autonomous differential equation to model bacterial growth. *Food Microbiology*, 10, 43-59.

Boos, D. (1992), On Generalized Score Tests. *The American Statistician*, 46, 327–333.

Buchanan, R. L., Smith, J. L. and Longa, W. (2000). Microbial risk assessment: dose-response relations and risk characterization. *International Journal of Food Microbiology,* 58, 159-172.

Le Marc, Y. (2001). Développement d'un modèle modulaire décrivant l'effet des interactions entre les facteurs environnementaux sur les aptitudes de croissance de *Listeria*, Thèse de doctorat. Université de Bretagne Occidentale.

Liang, K. Y. and Zeger, S. L. (1986), Longitudinal Data Analysis Using Generalized Linear Models *Biometrika*, 73, 13–22.

Rosso, L. (1996). Indices for performance evaluation of predictive models in food microbiology. *Journal of Applied Microbiology,* 81, 501-508.

Rotnitzky, A. and Jewell, N. P. (1990), Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data. *Biometrika*, 77, 485–497.

Vose, D. J. (2000). Risk analysis: a quantitative guide. John Wiley & Sons.

Session 11 : Sensométrie III / *Sensometrics III*

Plans expérimentaux pour les essais sensoriels: des règles abstraites et des exigences pratiques

Experimental Designs for Sensory Trials: abstract rules and practical requirements

Joachim Kunert[1],

[1] *Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany*
E-mail : *joachim.kunert@udo.edu*

Abstract
The paper wishes to discuss some problems with experimental designs used for sensory trials. It seems that it has become the fashion among sensometricians to rely too much on some abstract mathematical concepts, instead of thinking about the practical requirements it might take to make the experiment work. Following abstract mathematical concepts can certainly be helpful for finding an appropriate design that fits to the purpose of the experiment. However, the concepts have to be followed with thought. The ideas behind this statement will be exemplified with two examples. One challenge for the experiment is carryover effects. I was rather surprised when I read a draft for an international standard on consumer trials, where I found a terribly lengthy elaboration of neighbour balanced designs, including a discussion of mutually orthogonal Latin squares - but no mention of washout periods. Another challenge is order effects. Yes, balancing the order of products over the assessors generally is a good idea. However, there are circumstances when it is not. If you want to compare the assessors to each other, not the products, then it is indeed better to treat all assessors in the same way. And that may include presenting the products to each assessor in the same order.

Keywords: Randomized Design; Systematic Design; Carryover Effects; Order Effects; Experiments with Consumers; Unbiased Estimates

1 Introduction

Maybe the most important principle that an experimenter should follow when planning a sensory experiment, is to first ask three questions. The first question should be: "What do I want to find out with the help of my experiment?"; the second question then is to ask "Will my design provide an answer to the first question?" and the third question should be: "What circumstances might corrupt my results and how can I avoid that?"

A good experimental design is very important for the success of a sensory trial. If the experiment is poorly planned, this may introduce additional variance or even bias. There are a number of possible problems of a sensory trial which any good experimenter would try to avoid. One of these problems to be avoided might be a bias due to packages. Therefore, experimenters will try to have a blinded trial, using identical packages for all the products. Another one may

be mutual dependence between the answers from different assessors. It seems clear that the single assessments are no longer independent if one assessor knows the assessments of others. In most cases, therefore, experimenters may try to avoid letting assessors discuss their results with each other. A third problem may be increased variability. It seems clear that any surroundings that may distract the assessors or reduce their concentration will increase the variability of the assessments. There are many more methods that good experimenters use to improve the quality of the responses.

However, experimenters know very well that there may be situations where one or the other of these methods is not sensible. For instance, if experimenters want to find out what influence the packages have on the sensory perception of the products in the study, they will have to use different packages.

It seems to me, however, that this sound attitude is not used with respect to the statistical experimental design. Instead, the idea seems to be popular among sensometricians that it is sensible to use systematic designs which are balanced in as many ways as possible, without thinking much about their background. It seems that experimenters hope that a highly balanced design will always be good and protect against all kinds of problems.

Unfortunately, this is an illusion.

2 Randomized versus systematic designs

Consider an experiment with consumers to compare v products. Most likely, the experiment is done to compare the liking of the products. It then will make sense to get data from each assessor about each product. There clearly are differences between the assessors, for instance in their use of the scale. Some will generally give better marks than others. To avoid the risk that some products may have an unfair advantage by being tasted by assessors who generally give better marks, the experiment is normally done as a crossover experiment. That is, each assessor tastes each product, one after the other. There are, however, some problems with crossover designs. For instance, there may be order effects. It is well-known that assessors tend to give the first few products better marks on liking than later products. Furthermore, there is a problem of carryover effects. It is possible, that some problems may influence the assessments of the products which are tasted afterwards. An often quoted example is the case of a very bitter product, which may influence the perception of the bitterness of the next sample. A third problem may be sensory fatigue. This may increase the random component in later observations: the variance of the observations will increase over time.

A simplified model for a single assessment in a sensory experiment could be

$$y = \tau + \phi + e$$

where τ is the theoretical (ideal) observation, ϕ is the systematic bias introduced by external factors and e is the random error.

To explain what is meant with these terms, return to the crossover experiment and consider the i-th observation in the study. Assume that this observation is done with product 1, say, which in the population of interest will be rather popular and would get an average of, say, 8 if an infinite number of assessors could be asked. Then for this observation we would have $\tau = \tau_1 = 8$. Now assume that the observation is done by an assessor who normally gives rather poor marks, further assume that it is done at a later period, i.e. the assessor has already given several assessments before. Both these circumstances may mean that the observed response is

smaller than it might be. Therefore, the ϕ for this observation will be some negative number. Finally, there are all sorts of random influences acting on the assessment. Since we consider a situation where the assessor has tasted several other products before, we may assume that the assessor may be tired and therefore will pay less attention. This means that the random error may be relatively large in absolute size.

There are several methods discussed in the literature to deal with these problems. Some of these methods are statistical methods. For instance, it is a good idea to have each assessor taste each of the products, such that the difference between the products can be estimated free of the differences between the assessors. It is, however, also very important to use technical means to reduce the size of the systematic influences. For instance, the systematic differences between the assessors will certainly be larger in a study where the experimenter fails to introduce the assessors properly to their task. Disturbances like carryover effects can be reduced by so-called washout, i.e. by giving the assessors some food with a neutral taste between the products, aiming to re-normalize their sensory perception. The size of the error can be reduced, for instance by providing a proper surroundings with less noise. As said in the introduction, good experimenters have a large number of effective measures which will definitely help in reducing the size of the bias and of the error.

In addition to these technical solutions there are a number of statistical methods to increase the precision of the estimates. One of them is blocking which I have already mentioned: if an assessor tastes both products 1 and 2, then the difference between these assessments cancels the effect of the assessor.

Some other statistical methods are not so obvious. There is a never ending discussion whether to use a systematic or a randomized design. The idea of randomization is to shift a part of the systematic bias ϕ into the error. This will increase the variance of the error - but it will help to avoid systematic bias. We have tried to explain how this works in several earlier contributions, (e.g. Kunert, Meyners & Erdbrügge, 2002 or Kunert, 2007).

As opposed to that, a systematic design will try to model the systematic component ϕ. This leaves the variance of the error untouched. It may, however, lead to bias if the assumptions about ϕ in the model are wrong. As an (overly simplified) example, we consider an experiment with 4 products and 4 assessors, where each assessor evaluates each product once. A possible systematic design would be the Latin Square

$$d_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

where the rows indicate the assessors and the columns indicate the periods. We use simulations for this small example to show some ideas behind the discussion.

Assume in a first step that the $\phi_{i,j}, i, j = 1, ..., 4$ follow exactly a row-column model, i.e. we have $\phi_{i,j} = \alpha_i + \beta_j$, where α_i is the effect of the i-th assessor and β_j is the effect of the j-th period. For our simulations, we arbitrarily chose $\alpha_i = i$ and $\beta_j = j$ such that the matrix of all $\phi_{i,j}$ equals

$$\phi = \begin{pmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{pmatrix}.$$

To each of these numbers, we add a $\mathbb{N}(0,1)$-distributed random number and assume that the result is the response from an experiment run under the design d. Note that this simulates data under the null-hypothesis that there are no differences between the products. To simulate the effect of the products, we would add τ_r to each observation (i,j) receiving product r according to the plan d. This then gives the simulated responses $y_{i,j}$. For the present paper, however, we have simulated the case were all products are equal.

We then can analyse the simulated responses with the row-column model $y_{i,j} = \tau_{d(i,j)} + \alpha_i + \beta_j + e_{i,j}$. Assume we want to estimate, say, the difference between products 1 and 3. We then calculate

$$\widehat{\tau_1 - \tau_3} = \frac{1}{4}(y_{1,1} + y_{2,3} + y_{3,2} + y_{4,4} - (y_{1,3} + y_{2,4} + y_{3,1} + y_{4,2})),$$

i.e. the mean of the observations under product 1 minus the mean of the observations under product 3. Note that this completely cancels out the $\phi_{i,j}$. Each of the α_i and of the β_j gets added and subtracted exactly once.

We have repeated this procedure 1000 times, producing 1000 data-sets, all assuming that we have used design d_1. Table (1) gives the mean and the standard deviation of the 1000 realizations of the estimate $\widehat{\tau_1 - \tau_3}$ over all these experiments. This mean is near enough to 0 to indicate that the estimate is indeed unbiased.

However, another experimenter may assume that the period effect is negligible. Then he might decide to use a simple block design. One possible design would then be

$$d_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

If there are period effects, then a systematic design like d_2 clearly is a bad idea. To see this, look at observations simulated in the same manner as before, i.e. with additive assessor and period effects. Using d_2 and assuming a model without period effects, we would estimate the difference between products 1 and 3 by calculating the difference between the corresponding means, i.e. by the estimate

$$\widehat{\tau_1 - \tau_3} = \frac{1}{4}(y_{1,1} + y_{2,1} + y_{3,1} + y_{4,1} - (y_{1,3} + y_{2,3} + y_{3,3} + y_{4,3})).$$

Then the period does not cancel out. The estimate is biased by the period effects. Even if there are no true differences between treatments 1 and 3, when using design d_2 the estimate $\widehat{\tau_1 - \tau_3}$ will in the average equal -2. If we use a systematic design and there is something in ϕ that we have overlooked, then this will lead to biassed estimates.

However, instead of d_1 or d_2 we might use a randomized design. A randomized complete block design would present the products to each assessor in his/her own random order, randomized independently for each assessor. The randomization process should be such that each possible order gets chosen with the same probability. An example of a design derived with this randomization procedure is

$$d_3 = \begin{pmatrix} 2 & 4 & 1 & 3 \\ 4 & 1 & 2 & 3 \\ 3 & 1 & 2 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

design	mean	standard deviation
systematic Latin square d_1	0.009	0.711
randomized complete block	-0.015	1.16
randomized Latin square	0.005	0.718

Table 1: Mean and standard deviation of estimates under an ideal row column model

Note that using this design, the estimate

$$\widehat{\tau_1 - \tau_3} = \frac{1}{4}(y_{1,3} + y_{2,2} + y_{3,2} + y_{4,1} - (y_{1,4} + y_{2,4} + y_{3,1} + y_{4,3})),$$

is also not free from the period effect: for instance, the effect of the second period effect appears in two of the positive terms and in none of the negative. Randomization does not guarantee that an individual design produces estimates which are untouched by the systematic disturbances ϕ. Instead, randomization theory considers a single experiment as a member of a series of hypothetical experiments. It only guarantees that the average of the estimates over this series of hypothetical experiments is unbiased.

Note that this is the usual approach of statistical methods: the observed value of an estimate is never equal to the true parameter. If the estimate is unbiased, then the average over all hypothetical observations of the estimate equals the true value. A randomized experimental design therefore does not guarantee that a given experiment gives an estimate that is "near" the true product difference. It only guarantees that there is no systematic advantage of one product over the other.

Again, we have simulated 1000 data sets, each of them according to the row-column model. For each of them we have also simulated a new set of random orders for the four assessors. The results for the estimate $\widehat{\tau_1 - \tau_3}$ for this so-called randomized complete block design are in the second row of Table 1.

A comparison of the first and second row of Table 1 shows that the randomized design also produces an unbiased estimate, even though it neglects the period effect in the data. This is the important advantage of a randomized experiment: even if there is a structure in ϕ that the experimenter wishes to neglect (or is not aware of), this structure does not introduce systematic bias. A comparison of the standard deviations shows where the neglected period effect has gone to: if a randomized design neglects a structure present in the ϕ of the experiment, then this inflates the variances of the estimates.

If the experimenter wishes to use a randomized design and take account of the period effect, then a randomized Latin square could be used. This design would randomly permute the order of the rows and the columns of a Latin square. Again simulating 1000 data-sets according to the row-column model and 1000 randomized Latin squares, we get the entries in the third row of Table 1. Our simulations show that if the data are constructed under the ideal row-column model, then the randomized Latin square and the systematic Latin square both perform equally well.

However, the data of a true experiment will not have an ideal row-column structure. There will be a general tendency of assessors to give better marks in early periods but this will most likely not be a simple additive structure. We therefore have repeated our simulations under a

design	mean	standard deviation
systematic Latin square d_1	1.756	0.706
randomized complete block	0.006	2.58
randomized Latin square	0.011	1.84

Table 2: Mean and standard deviation of estimates under a nonlinear structure

second matrix ϕ, namely

$$\phi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{pmatrix}.$$

Again, we have simulated 1000 data sets for the systematic design d_1 and for the randomized complete block design and the randomized Latin square. The results are reported in Table 2.

We can see from Table 2 that the systematic Latin square has a problem if the model for the ϕ is not exactly correct. There is a clear bias. The two randomized designs, however, remain unbiased. Note that the randomized Latin square has a considerably smaller variance than the randomized complete block design: there is a clear period effect in the data.

A very important aspect of an experiment is the estimation of the variance. Use of a randomized design allows to get an unbiased estimate of the variances of treatment estimates. We have discussed this in more detail in earlier papers, e.g. Kunert, Meyners & Erdbrügge (2002) or Kunert (1998).

3 Carryover effects

If the taste of a product is liable to influence the assessments of products to be tasted later, this may invalidate all comparisons in the experiment. Therefore, some experimenters try to reduce the size of these so-called carryover effects e.g. by presenting neutral substances between the samples to wash out the carryover effect.

Since the appearance of the paper by MacFie et al (1987), use of neighbour balanced designs has become very popular. As expressed in Kunert (1998), I think that uncritical use of neighbor balanced designs has several disadvantages. Firstly, some experimenters may think that use of neighbour balanced designs completely resolves the problem of carryover. At least, I have heard a lot more discussions about the construction of neighbour balance than about ways to avoid carryover in the first place. Secondly, if they use a neighbour balanced design, experimenters tend to use it as a systematic design and therefore do not benefit from the advantages of randomization. And thirdly, as pointed out in Kunert (1998), there are a number of problems, even if we assume that a simple model holds where the carryover effects are additive. For instance, even for neighbour-balanced designs, the direct and carryover effects are not orthogonal to each other. This implies that the uncorrected estimates of treatment differences are biased by carryover effects. If, however, we use the corrected estimate, then the usual estimate of its variance is biased (see Kunert and Utzig, 1993). Furthermore, it is not clear at all that the model with additive carryover effects is a good approximation to the data. For instance, Senn (2004, p. 298 ff) gives "five reasons for believing that the simple carry-over model is not useful".

If there are carryover effects, then there even is another problem. As pointed out by e.g. Bailey and Druilhet (2004) the experimenter has to decide whether the direct effects of the treatments are really of interest. If there are carryover effects and a product is assessed twice in a row by the same assessor, then the second assessment might be quite different. This is because at the second assessment, the assessor's response contains the so-called permanent effect, which is the sum of the direct effect of the product plus the carryover effect of this product. If in commercial use the consumers are expected to eat this product twice or more often in a row, then maybe these permanent effects are what we are really interested in. However, if we are interested in permanent effects, then neighbour balanced Latin squares or similar designs are not useful. Instead, we should then use a design that allows to estimate the permanent effects properly.

For instance, it was shown by Bailey and Druilhet (2004) that the design

$$d_4 = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 4 & 4 \\ 2 & 2 & 3 & 3 \\ 2 & 2 & 4 & 4 \\ 3 & 3 & 4 & 4 \\ 4 & 4 & 1 & 1 \\ 3 & 3 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 4 & 4 & 2 & 2 \\ 3 & 3 & 2 & 2 \\ 4 & 4 & 3 & 3 \end{pmatrix}.$$

is optimal for the estimation of total effects if there are four periods and four products. Again, rows indicate assessors. Note that this design makes sense from a practical viewpoint: If there are carryover effects and the experimenter wants to find out what is the permanent effect of a certain product, then the assessors must taste each product more than once.

In all, neighbour balanced designs should be used with care. If there is proper washout they may be unnecessary. Depending on the purpose of the experiment they may be inefficient, even if there are carryover effects following an ideal model.

4 Period effects

As pointed out above, balance for period effects will often increase the efficiency of a design. However, there are situations when using the same order for each assessor may be more efficient.

The aim of a study reported in Alexy et al (2011) was to analyze the association between body weight status on the one hand and sensory preferences and discrimination abilities on the other. This was done in a large cross-sectional study with children and adolescents, using 9 experimental tests, covering 4 different taste categories. Before the work on this project had started, the authors thought it was most likely that children who have a preference for sugar or fat are more liable to become overweight than others. However, we thought it might also be possible that the main causes for overweight are something else, like, for instance, too little physical exercise. Or, maybe the sensory preferences of the parents are more important than those of the children themselves. We therefore thought that, at least with younger persons,

maybe there is no difference between the weight groups, neither in sensory perception nor in preference.

The authors therefore decided to design the experiment in such a way that there are two possible outcomes: a) maybe we can prove a difference between the weight groups, b) maybe we cannot prove a difference. An important concern of the experiment was that, if we should have outcome a) this should be due to true differences, not due to some bias induced by the experiment. To allow for the case that outcome b) should happen, the authors thought it might be important to design the experiment in such a way that failure to prove a difference is some evidence that any true differences between the weight groups must be "small".

We tried to design the experiment in such a way that both, sensory discrimination ability and sensory preference, should be covered. Furthermore, the basic tastes (sweet, salty, sour and bitter) as well as the perception of fat should be considered. On the other hand, the design had to avoid making the task for the children too complicated. Therefore, we decided to ask each participant to do a series of 9 paired comparisons. The series of comparisons consisted of 5 paired comparisons in the form of preference tests and 4 paired comparisons in the form of sensitivity tests. In the preference tests, we used food which would really be consumed in everyday use. For the tastes sweet, salty and sour, two different samples could be produced by adding different amounts of sugar, salt and citric acid, respectively. Finding a good product to compare two concentrations of fat turned out to be more difficult. Since we could not add fat to an existing product, we had to use two samples that should only differ in the amount of fat. We found a gouda-cheese and a salami-sausage, both of which existed in a low-fat and a high-fat version of the same brand. We were aware that both products (in particular the sausage which was based on pork) might be refused by parts of the participants and therefore decided to use both. It later turned out that most participants were willing to taste the cheese, but a large number refused the sausage. In the sensitivity tests, we used water solutions with different concentrations of sugar, salt or citric acid. For the sensitivity test for fat, however, we used milk with different fat contents. An originally planned test for the taste "bitter" was prohibited by the ethical committee, because it might be unpleasant for the participants.

Each assessor performed all tests in a single session, and the order of the tests was the same for each participant. This confounds the test with a possible order effect. Note, however, that we do not wish to compare the different tests to each other. While our design makes the comparison of one test to the other impossible, confounding the tests with order effects reduces the variance of the comparison between the weight groups: the order effects subtracts out in this comparison. Hence, running the tests in the same order for each participant increases the power of the comparison between the weight groups.

The main purpose of the study was to compare the weight groups to each other: Is there a significant difference in the proportion of decisions for the more intense sample between the weight groups in any of the nine tests? Since there is a clear ordering of the three weight groups, we did each of these comparisons with an extension of Fisher's exact test for two-by-three tables, see Agresti (2002, p. 97 - 98).

The test session started with the preference tests. Different products were used for the preference tests and for the sensitivity tests, to avoid a possible bias. We were concerned that the answers might be biased in the preference tests, if the participants knew which attributes varied between the two samples in a given test (see e.g. Earthy et al, 1997). In particular, we were concerned that this bias might be different for the different weight groups - for instance overweight children might hear more often that they should not eat sweeter foods. For the sensitivity tests, the differences in the concentration of the stimulus were planned in such a

way that about 50% of the participants should be able to sense the difference. This would result in about 75% of correct answers. With this proportion of "identifiers", we expected the highest chance of finding possible differences between the groups. We used a larger difference in concentrations of the stimulus within the pairs used for the preference tests: preferably all assessors in all groups should be able to experience a difference. However, we had to limit the concentrations in such a way that both samples would still be acceptable: the chance to find differences between the groups is largest if about 50% of all participants prefer sample A and 50% prefer sample B.

Before the start of the proper experiment, the experimental design was pre-tested by our study group asking 7th graders from another school to do these tests (with 22 participants). Except for some minor changes, the methods turned out to be suitable for testing the age group of interest outside of special laboratories. The most noteworthy change after the pre-test was a change in the phrasing of one question: we were surprised to find during the pre-test that there were less than 50% of correct identifications of the high fat milk as fattier (German: "fettiger"). It seems that many participants could not relate the pleasant taste of high fat milk with the negative term "fettiger". We therefore rephrased the question to "Which of the two samples is creamier?" (German: "sahniger"). This increased the number of correct answers largely.

Subjects with a BMI larger than the 90% quantile of the German reference (Kromeyer-Hauschild et al, 2001) were defined as overweight, those with a BMI of more than the 97% -quantile were defined as obese. To achieve a sufficient power for the comparison of the three weight groups, we had planned to do the study with a minimum of 500 participants. We assumed that the group of obese children then should be approximately 50, expecting that far more than 3% of our target population should be obese. A total of 574 fifth to ninth graders participated in the study, about half of them (284) were boys. However, some of the participants omitted some of the tests. Therefore, for each single test the number of participants was less than 574. In particular, only 410 assessors took part in the test with salami. However, for the eight other tests, the sample size was above the 500 that we had planned. In the sample, 426 subjects were classified as normal weight (74.2%), 94 as overweight (16.4%) and 54 as obese (9.4%). This agreed well with the expectation that the proportion of overweight and obese children should be higher than in the reference population.

Detailed results of the study are reported in Alexy et al (2011). There was no indication of a difference between the weight groups. We think that the design of the study with the same order of the tests for each assessor improved the precision of the comparison between the weight groups and therefore was adequate.

5 Conclusions

The paper tried to make clear that the design has to fit to the research question. A design which would be rather poor to answer one kind of questions might be quite good for another kind. In any case, the experimenter should take some time to think about the aims of the experiment before choosing a design.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis.* Hoboken, N.J.: John Wiley & Sons, 2nd edition.

Alexy, U., Schaefer, A., Sailer, O., Busch-Stockfisch, M., Huthmacher, S., Kunert, J. & Kersting, M. (2011). Sensory Preferences and discrimination ability of children in relation to their body weight status. *Journal of Sensory Studies* 26, 409-412.

Bailey, R.A. & Druilhet, P. (2004). Optimality of neighbour balanced designs for total effects. *Annals of Statistics* 32, 1650-1661.

Earthy, P.J., MacFie, H.J.H. & Hedderley, D. (1997). Effect of question order on sensory perception and preference in central location trials. *Journal of Sensory Studies* 12, 215 - 237.

Kromeyer-Hauschild, K., Wabitsch, M., Kunze, D., Geller, F., Geiß, H.C., Hesse, V., von Hippel, A., Jaeger, U., Johnsen, D., Korte, W. et al. (2001). Perzentile für den Body-mass-index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben. *Monatsschrift Kinderheilkunde* 149, 807 - 818.

Kunert, J. & Utzig, P. (1993). Estimation of variance in cross-over designs. *Journal of the Royal Statistical Society* 55, 919-927.

Kunert, J. (1998). Sensory Experiments as Crossover Studies. *Food Quality and Preference* 9, 243-253.

Kunert, J., Meyners, M., & Erdbrügge, M. (2002). On the applicability of ANOVA for the analysis of sensory data. *7e Journées européennes agro-industrie et méthodes statistiques*, 129-134.

Kunert, J. (2007). Randomization in Experimental Designs. In: *Encyclopedia of Statistics in Quality and Reliability*, Ruggieri, F., Kenett, R. & Faltin, F.W. (Eds.) John Wiley and Sons Ltd., Chichester, UK, 1559-1563.

MacFie, H.J.H., Bratchell, N., Greenhoff, K. & Vallis, L.V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in Hall tests. *Journal of Sensory Studies* 4, 129-148.

Senn, S. (2002). *Crossover Trials in Clinical Research.* Chichester, UK: John Wiley & Sons, 2nd edition.

CRAGGING: une nouvelle approche pour évaluer la qualité des vins italiens

CRAGGING: a novel approach for inspecting Italian wine quality

Eugenio Brentari[1], Maurizio Carpita[1] & Marika Vezzoli[1]

[1] *Department of Quantitative Methods, University of Brescia*
C.da S. Chiara, 50 - 25122 Brescia, Italy
E-mail: {*brentari, carpita, vezzoli*}*@eco.unibs.it*

Abstract

Assessing the wine quality is a challenging task due to the multifaceted nature of such a concept. Indeed, subjective evaluations and objective features are mixed together in order to get fair judgements and effective ranking of wines. To asses wine quality, chemical and sensory tests are commonly used. These tests usually collect a relatively high number of variables among which some certainly play the role of key factors underlying the essence of the wine quality. It is then extremely important to identify these fundamental attributes, since they can lead to significant improvements in the understanding process of the wine quality.

Using a data mining approach, in this paper we inspect the quality of Italian red and white wines trying to identify which of the sensorial and chemical-type variables have a major impact on it. In detail, we analyze the dataset used by Altroconsumo, an Italian independent consumer's association, for the *Guida Vini 2011*, containing 231 wines grouped with respect to the type of grapes used by producers. Since the dataset has a hierarchical structure, we use a new algorithm, called CRoss-validation AGGregatING (CRAGGING), well suited for this type of data. In particular, we extract a synthetic model able to identify the predictors and correspondent thresholds useful for showing the "true path" towards the quality.

Résumé

Evaluer la qualité des vins est une tâche difficile en raison de la nature multiforme d'un tel sujet. En effet, les évaluations subjectives et les caractéristiques objectives sont mélangés afin d'obtenir des jugements justes et un classement effectif des vins. Pour évaluer la qualité du vin, des tests chimiques et sensoriels sont couramment utilisés. Ces tests recueillent habituellement un nombre relativement élevé de variables dont certains jouent certainement le rôle des facteurs clés qui soulignent l'essence de la qualité du vin. Identifier ces attributs fondamentaux devient alors extrêmement important, car ils peuvent conduire à des améliorations significatives dans le processus de compréhension de la qualité du vin.

En utilisant un approche data mining, dans cet article nous examinons la qualité de vins rouges et blanc d'Italie et nous essayons d'identifier quelles variables sensorielles et chimiques en produisent le plus fort impact. Nous analysons en détail les données utilisées par Altroconsumo,

une association indépendante italienne de consommateurs, pour la *Guida Vini 2011*, qui contient 231 vins groupés par type de raisins utilisés par les producteurs.

Comme le dataset suit une structure hiérarchique, nous utilisons un nouvel algorithme, appelé CRoss-validation AGGregatING (CRAGGING), bien adapté pour ce type de données. On a extrait un modèle synthétique qui nous permet d'identifier les variables et leur relatives seuils de référence pour améliorer la qualité du vin.

1 Introduction

Once viewed as a luxury good, today wine is enjoyed by an increasing number of consumers. To support its growth, the industry is investing in new technologies to improve making and selling processes both implemented having the objective to increase the quality of the wine. While it is clear that wine quality appears as a key element for producers and sellers, its concept is in some sense elusive and experts rely on chemical and sensory tests for its measurement. Chemical tests include determination of density, alcohol or pH values, while sensory tests mainly rely on human experts. Since taste is the least understood of the human senses (Smith and Margolskee, 2006), the evaluation of the wine quality is a difficult task. Moreover, such tests collect a huge number of variables among which it is essential to select the most important ones in order to improve the wine quality.

Using a data mining approach the objective of this paper is to provide a variable selection among potential drivers of the wine quality.

The issue is interesting both from a methodological and economic perspectives. Indeed, only recently some data mining approaches have been applied to inspect the wine industry. In Brentari and Zuccolotto (2010), for *e.g.*, Random Forests and their variable importance measures (Breiman, 2001) have been applied to understand the major determinants of the low-priced Italian red wines.

Based on the same philosophical approach, while using a different methodology, the quality of Italian red and white wines is explored with the end to identify which of the sensorial and chemical-type variables have a major impact on it. The dataset used in our study derives from Altroconsumo, an Italian independent consumer's association, for the *Guida Vini 2011*[1]. The quality of wine was used as response variable measured with a composite score from 0 (lowest quality) to 100 (highest quality). In total we analyzed 231 wines grouped in 49 clusters with respect to the type of grapes used by producers, together with 7 chemical and 26 sensorial variables. The algorithm used in the analysis is the CRoss-validation AGGregatING (CRAGGING henceforth) (Vezzoli and Stone, 2007; Vezzoli, 2011). Such algorithm was introduced to handle with hierarchical data structures maintaining the "inner" homogeneity of observations belonging to the same group. As the Random Forests, the CRAGGING is an ensemble of trees (Breiman *et al.*, 1984) and its main concern is the lack of interpretability. For this reason, Vezzoli and Stone (2007) suggested to extract a single tree from the ensemble, called Final Model, in order to obtain a synthetic predictor.

In our results, the Final Model highlights that 5 sensory variables (aromatic richness, harmony, attraency, structure and frankness) and a chemical-type feature (total sulphur anhydrides) seem to be the key factors explaining the wine quality.

The paper is organized as follows. Section 2 provides the description of the CRAGGING

[1]The authors thank Luigi Odello, chairman of Centro Studi Assaggiatori of Brescia, who supplied data, and Altroconsumo for the allowance to use them.

algorithm. In Section 3 we describe the dataset and we discuss the results obtained in the empirical analysis while Section 4 concludes.

2 The CRAGGING algorithm

Tree-based methods are suited for the analysis of complex data which require flexible and robust analytical methods to deal with nonlinear relationships, high-order interactions, and missing values. Trees explain the variation of a response variable y by repeatedly splitting the data into more homogeneous groups, using combinations of explanatory variables that may be categorical and/or numeric. The splitting criterion allows to select at each tree node the best covariate and the cut-off point along it (Breiman *et. al*, 1984). Trees are a popular tool in many scientific areas since their representation is easy to understand and unambiguous.

One of their major concern is the instability. In other terms, trees produce models that can change dramatically with small changes in the data undermining the ability of the predictor to produce knowledge (Dietterich, 1996). An approach that mitigates this problem and increases the accuracy of the predictors consists of computing a high number of simple models (*e.g.* trees), called base or weak learners, within the perturbed training set and combining them to obtain a univariate and stable predictor. These multiple models, called ensemble learning, include Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1996) and Random Forest (Breiman, 2001). Since these algorithms do not take into account the inner structure of the data, when dealing with hierarchical datasets (in which the dependencies between the observations could play a key role) they remove fundamental features to be used to better describe complex relationships. The CRAGGING algorithm was introduced to handle this concern and indeed is conceived in such an extent to maintain the inner structure of the data.

We briefly describe the CRAGGING algorithm using the same notation as in Vezzoli and Zuccolotto (2011). Let first denote by $(y, \mathbf{x})$ a hierarchical dataset with J groups and N observations, where y is the dependent variable and $\mathbf{x}$ is the matrix of the R predictors. Each group is composed by n_j observations and $N = \sum_{j=1}^{J} n_j$. Let us denote with $\mathcal{L} = \{1, 2, \cdots, J\}$ the set of groups and with $x_{ji} = (x_{1ji}, x_{2ji}, \cdots, x_{rji}, \cdots, x_{Rji})$ the vector of predictors for i-th subject of group j where $j \in \mathcal{L}$ and $i = 1, 2, \cdots, n_j$. The set $\mathcal{L}$ is randomly partitioned in V subsets denoted by $\mathcal{L}_v$ with $v = 1, \cdots, V$, each one containing J_v groups. For each v, let $\mathcal{L}_v^c$ be the complementary set of $\mathcal{L}_v$, containing J_v^c groups. Moreover, let $\mathcal{L}_{v\setminus\ell}^c$ be the set obtained by removing the ℓ-th group from $\mathcal{L}_v^c$. For each $\mathcal{L}_v$ and for each $\ell \in \mathcal{L}_v^c$, let $\hat{f}_{\alpha,\mathcal{L}_{v\setminus\ell}^c}(\cdot)$ be the prediction function of a single tree (base learner) trained on data $\{y_{ji}, x_{ji}\}_{j\in\mathcal{L}_{v\setminus\ell}^c,\ i=1,\cdots,n_j}$ and pruned with cost-complexity parameter α. The corresponding prediction for the observation not used to grow the tree is given by

$$\hat{y}_{ji,\alpha\ell} = \hat{f}_{\alpha,\mathcal{L}_{v\setminus\ell}^c}(x_{ji}), \quad \text{with} \quad j \in \mathcal{L}_v, \quad \text{and} \quad i = 1, 2, \cdots, n_j. \tag{1}$$

An aggregated prediction over the groups contained within the test set $\{y_{ji}; \mathbf{x}_{ji}\}_{j\in\mathcal{L}_v,\ i=1,\cdots,n_j}$ is obtained by the average of functions (1):

$$\hat{y}_{ji,\alpha} = \frac{1}{J_v^c} \sum_{\ell\in\mathcal{L}_v^c} \hat{f}_{\alpha,\mathcal{L}_{v\setminus\ell}^c}(x_{ji}) \quad \text{with} \quad j \in \mathcal{L}_v \quad \text{and} \quad i = 1, 2, \cdots, n_j. \tag{2}$$

The procedure is repeated for different values of α and finally the algorithm chooses the optimal tuning parameter α^* which corresponds to that value for which the out of sample error

estimation over all $\mathcal{L}_v$ is minimized:

$$\alpha^* = \arg\min_{\alpha} L(y_{ji}, \hat{y}_{ji,\alpha}) \quad \text{with} \quad j \in \mathcal{L}, \quad i = 1, 2, \cdots, n_j \tag{3}$$

where $L(\cdot)$ is a generic loss function (usually MSE for regression and misclassification rate for classification). The CRAGGING predictions are given by

$$\tilde{y}_{ji}^{\text{crag}} = \hat{y}_{ji,\alpha^*} \quad \text{with} \quad j \in \mathcal{L}, \quad i = 1, 2, \cdots, n_j.$$

The generalization error of the algorithm is measured by the loss function $L_{\alpha^*} = L(y_{ji}, \hat{y}_{ji,\alpha^*})$ that is based on the predictions of a generic subject i computed using trees grown with training sets not containing that subject.

As suggested by its name, the CRAGGING increases the diversity between the trees using a double cross-validation. In fact, the algorithm uses the *leave-one-unit-out* cross-validation when removes the ℓ-th group from the training set and the v-fold cross-validation when repeats the procedure for each v-set.

As mentioned before, ensemble learning algorithms are excellent predictors but their main concern is the lack of interpretability. On this issue, some authors studied methods for extracting a simpler and more comprehensible model from an overly complex one (Evans and Fisher, 1994; Fayyad *et al.*, 1996; Breiman and Shang, 1997). Vezzoli and Stone (2007) proposed to combine the results of CRAGGING with a single tree. In detail, they replace the dependent variable y with the predictions $\tilde{y}_{ji}^{\text{crag}}$, and they grow a single regression tree with cost complexity parameter α^* on $(\tilde{y}_{ji}^{\text{crag}}, \mathbf{x})$. As a result they obtain a simple model, called Final Model, namely a predictor with a good interpretability and whose accuracy is better than the accuracy of a single tree.

Following the same philosophy of the CRAGGING, Vezzoli and Zuccolotto (2011) modified the Mean Decrease in Accurary (MDA henceforth) measure of variable importance, proposed by Breiman (2001), in order to perturb the data without destroying their inner structure. In detail, at each tree of CRAGGING, in correspondence of the optimal tuning parameter (α^*), the values of the r-th variable are randomly permuted and new predictions are obtained with this new data set $(y, \mathbf{x})_r$. Hence, a new loss function $L_{\alpha^*,r}$ is computed and it is compared with L_{α^*}.

In particular, the authors randomized the values of r-th variable conditionally to the J groups of the data set. In other words, for each variable X_r a permutation $p = \{p_1, p_2, \cdots, p_J\}$ of the sequence $\{1, 2, \cdots, J\}$ is randomly selected. The values of X_r are randomized in the data set according to the following rule:

$$\{x_{rji}\}_{j \in \mathcal{L},\ i=1,2,\cdots,n_j} = s(x_{rp_j}), \tag{4}$$

where $s(\cdot)$ denotes a sampling with replacement from a set of values and $x_{rp_j} = \{x_{rp_j i}\}_{i=1,\cdots,n_{p_j}}$. This randomization should be particulary useful if $f(X_r|j_1) \neq f(X_r|j_2)$ for all $j_1 \neq j_2$, as frequently happens in the application domain of the CRAGGING. The procedure of sampling is repeated k times and the MDA measure for the r-th variable is given by the following average (denoted as av) on k:

$$MDA_r = av_k(L_{\alpha^*,r} - L_{\alpha^*}).$$

3 Empirical Analysis

The analysis was carried on the dataset that Altroconsumo, an Italian Independent Consumers' Association, uses for its guide (*Guida Vini 2011*). Each year, about 300 wines (red and white)

are bought and their chemical and sensorial characteristics are evaluated. Wines are chosen in order to represent the variety of Italian wines as regards vineyards, producers and region of origin, and the most part of them cost less than 15 euro (Brentari and Levaggi, 2010; Brentari and Zuccolotto, 2010; Brentari *et. al*, 2011).

3.1 Dataset

In our study, we analyzed 231 red and white wines[2] grouped in 49 clusters defined by the type of grapes used by producers. Although the dataset of Altroconsumo collects several types of information, we focused only on chemical and sensory characteristics, since we expect these features could be the major drivers of the wine quality.

Chemical variables measure objective characteristics of the wine. They include the verified alcoholic strength (***Alcohol***), the residual sugar[3] (***Residual Sugar***), the reducer sugar (***Reducer Sugar***), the total and the volatile acidity[4] (***Acidity tot*** and ***Acidity vol***, respectively), the ratio between free and total sulphur anhydrides[5] ($\boldsymbol{SO_2}$) and the total sulphur anhydrides[6] ($\boldsymbol{SO_2}$ ***tot***).

Sensory variables were collected with the help of Brescia's Centro Studi Assaggiatori which assesses the sensory characteristics of the wine selected by Altroconsumo. Experienced judges, divided in panels balanced for age and sex, express their vote about the following sensory variables divided in four groups: (*i*) visual characteristics (***Color saturation***, ***Green reflection***, ***Gold reflection***, ***Violet reflection***, ***Garnet reflection***, ***Visual sparklingness***, ***Attraency***[7]), (*ii*) olfactory characteristics (***Floral***, ***Fruit***, ***Vegetal***, ***Spicy***, ***Olfactory intensity***, ***Olfactory quality***, ***Olfactory frankness***, ***Perception***, ***Harmony***), (*iii*) gustatory characteristics (***Structure***, ***Harmony***, ***Acidity***, ***Bitterness***, ***Sweet***, ***Astringency***, ***Aromatic Richness***) and (*iv*) Intense Aromatic Persistence (***Persistence***, ***Frankness***, ***Quality***). The perception of each descriptor has been registered using a 0-9 scale where 0 denotes the lowest and 9 the highest score.

Altroconsumo provide also a global evaluation of the wine quality attributing a composite score from 0 (lowest quality) to 100 (highest quality). It is just this score that we used as response variable in our analysis.

3.2 Results

In order to inspect the wine quality assessing the contribution played by the chemical and sensory drivers, and given the structure of the dataset, which is clearly hierarchical, we run the CRAGGING algorithm using the indicator of the wine quality as dependent variable and the 7 chemical together with the 26 sensory variables as predictors.

To understand how the characteristics impact on the wine quality we run the Final Model, then giving a possible mapping of the quality based on a series of rule of thumbs. In other terms, such a Final Model could be used as a tool to drive the quality towards high scores, also

[2]For obtaining a balanced dataset, we excluded from our analysis the wines belonging to the 2011 dossier. In addition, we discarded also the wines whose classification was difficult and ambiguous.

[3]It determines the organoleptic characteristics of the wine.

[4]The total acidity influences the flavour of the wine while the volatile acidity signals how well the wine is preserved and how it fermented.

[5]It allows to evaluate the quality of the technology used for wine making.

[6]It helps in the wine making process.

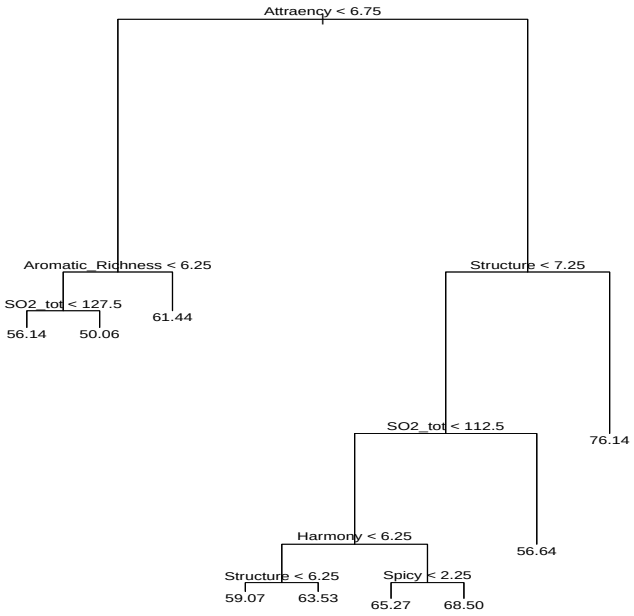
[7]It measures how pleasant is the aspect of the wine.

avoiding scarce products, based on the monitoring of the variables selected by the procedure (regression tree) with the corresponding thresholds.

It is interesting to note that in the Final Model reported in Figure 1 the wines with highest score (76.14) have a pleasant aspect (***Attraency*** > 6.75) and a good flavour (***Structure*** > 7.25). In order to have high scores in Altroconsumo's *Guida Vini* it is therefore necessary to have a pleasant-looking wine and a rich profile, both on the tactile and the taste side.

On the contrary, the wines with lowest score (50.6) do not have an attractive aspect (***Attraency*** < 6.75), do not have a pleasant flavour (***Aromatic richness*** < 6.25) and the level of the total sulphur anhydrides is very high (***SO_2 tot*** > 127.5). In particular, the total sulphur anhydrides plays an important role in the Altroconsumo ranking. According to the EU law, usually the maximum level of ***SO_2 tot*** is 160 mg/l for red wines and 210 mg/l for white and rosé wines. The higher the level of the total sulphur anhydrides, then, the more the wine is penalized.

Figure 1: Final Model



4 Concluding remarks

In this study we handle the problem of the wine quality evaluation, which is a complex task due to the fact that extrinsic and intrinsic characteristics interact with each other in a non-simple way. To make clear how these features impact on the wine quality, we implemented a data mining technique pertaining to the ensemble learning. Such an approach appeared to be well

suited to describe the relationship between the wine quality and its most important drivers also showing which is the "true path" towards better products in terms of high levels of quality.

In particular, winemakers have to pay attention to the flavour and the pleasant aspect of the wine. Moreover, other four out of twenty-six sensory variables show substantial importance, *i.e.*, two gustatory characteristics (***Aromatic richness*** and ***Structure***), an olfactory characteristics (***Harmony***) and the intense aromatic persistence (***Frankness***).

In addition, the synthetic model extracted from the ensemble shows a negative relationship between the quality and $\boldsymbol{SO_2}$ ***tot***. Hence, in order to improve the wine quality, winemakers should monitor such a chemical additive maintaining the corresponding values within specific ranges endogenously identified by the synthetic model.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26, pp. 123-140.

Breiman, L. (2001). Random Forest. *Machine learning*, 45(1), pp. 5-32.

Breiman, L., Shang, N. (1997). Born again trees. *Technical report*, Statistics Department, University of California Berkeley, Berkeley, CA.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J (1984). *Classification and regression trees*, Monterey, California, USA, Wadsworth.

Brentari, E., Levaggi, R. (2010). Hedonic price for Italian Red Wine: a panel analysis. *Proceedings of the 11th European Symposium on Statistical Methods for the Food Industry*, Academy School, Afragola (Napoli), pp. 249-258.

Brentari, E., Levaggi, R., Zuccolotto, P. (2011). Pricing strategies for Italian red wine. *Food Quality and Preference*, 22, pp. 725-732.

Brentari, E., Zuccolotto, P. (2010). The implicit value of chemical and sensorial quality in the hedonic analysis of low-priced Italian red wines. *Proceedings of 11th European Symposium on Statistical Methods for the Food Industry*, Academy School, Afragola (Napoli), pp. 269-276.

Dietterich, T.G. (1996). Editorial. *Machine Learning*, 24, pp. 91-93.

Evans, B., Fisher, D. (1994). Overcoming process delays with decision tree induction. *IEEE Expert*, 9, pp. 60-66.

Fayyad, U.M., Djorgovski, S.G., Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, CA: AAAI Press, pp. 471-493.

Freund, Y., Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, San Francisco: Morgan Kaufman, pp. 148-156.

Guida Vini (2011), Altroconsumo Edizioni, Milan, Italy.

Smith, D., Margolskee, R. (2006). Making sense of taste. *Scientific American, Special issue*, 16(3), pp. 84-92.

Vezzoli, M., Stone, C. J. (2007). CRAGGING. In *Book of Short Papers CLADAG (Classification and Data Analysis Group) 2007*, pp. 363-366, EUM.

Vezzoli, M. (2011). Exploring the facets of overall job satisfaction through a novel ensemble learning. *Electronic Journal of Applied Statistical Analysis*, 4(1), pp. 23-38.

Vezzoli, M., Zuccolotto, P. (2011). CRAGGING measures of variable importance for data with hierarchical structure. In *New Perspectives in Statistical Modeling and Data Analysis*, pp. 393-400, S. Ingrassia, R. Rocci, M. Vichi (Eds.), Springer.

Prédiction de la qualité sensorielle de la viande bovine basée sur une approche d'analyse d'images multi-échelle

Prediction of the sensory quality of bovine meat based on mutiscale image analysis approach

Mohammed EL Jabri[1,2], Saïd Abouelkaram[3] & Daniel Roux[2]

[1] *ADRIA Développement, Créac'h Gwen, F-29196 Quimper, France*
[2] *Laboratoire de Mathématiques, Université Blaise Pascal. 63177 Aubière France*
E-mail : *mohammed.eljabri@adria.tm.fr*

[3] *INRA, UR370 Qualité des Produits Animaux, F-63122 Saint Genès Champanelle, France*

Résumé
Une approche par analyse d'images mutli-échelle est proposée en vue de prédire la qualité de la viande bovine en particulier la tendreté. Les images analysées ont été acquises sous un éclairage en lumière blanche polarisée. Elles ont été analysées en s'appuyant sur un Modèle de Vision Multi-échelle (MVM) basée sur la transformée en ondelettes discrète, notamment l'algorithme "à trous". Ce modèle permet d'isoler les structures significatives du Tissu Conjonctif (TC) du muscle. L'information, retenue des images analysées, est la distribution des surfaces d'objets correspondant aux éléments du TC. La méthode de Régression sur les Composantes Principales (RCP) a été appliquée aux données issues des étapes de traitement des images. Cinq variables d'images ont été sélectionnées avec cette méthode parmi treize. Leur combinaison linéaire a permis une bonne prédiction de la tendreté de la viande (R^2=0.92).
Mots-clés : Tissu Conjonctif (TC), Tendreté, Modèle de Vision Multi-échelle (MVM), Régression sur les Composantes Principales (RCP).

Abstract
A multi-scale based image analysis approach was used to predict beef quality particularly tenderness. The images were acquired under polarized visible lighting. They were analyzed leaning on a Multi-scale Vision Model (MVM) based discrete wavelet transform, especially "à trous" algorithm. This model allowed to extract significant structures of muscle considered for Connective Tissue (CT). The information, retained from analyzed images, is the distribution of considered object surfaces corresponding to CT elements. Principal Component Regression (PCR) was applied to data issued from image processing steps. Five variables were selected with this method from thirteen. Their linear combination has a good prediction of meat tenderness (R^2=0.92).
Keywords : Keywords : Connective Tissue (CT), Tenderness, Multi-scale Vision Model (MVM), Principal Component Regression (PCR).

1 Introduction

La qualité de la viande bovine englobe une multitude de propriétés d'origines diverses plus ou moins bien maîtrisées par le producteur, le transformateur et même le consommateur lors de la préparation de la viande. Pour la majorité des consommateurs, le principal critère de qualité est la tendreté de viande de boeuf fraîche concernant surtout les morceaux à griller. La dureté de la viande résulte de la résistance mécanique des deux principales composantes structurales du muscle : les fibres contractiles musculaires et le Tissu Conjonctif (TC). En maîtrisant le conditionnement *post mortem*, une grande partie de la dureté liée aux fibres musculaires peut être contrôlée (Maunier-Sifre, 2005). En revanche, pour le TC sa dureté reste constante au cours du temps; de plus il présente d'importantes variations dans sa constitution et dans son organisation spatiale. Le TC se compose principalement de collagène dont le contenu et la distribution contribuent significativement à la dureté intrinsèque de la viande. La teneur en collagène endomysial varie peu entre différents types de muscle, et ne semble pas impliquée dans la variabilité de la texture de la viande (Nakamura et al., 2003). Comme l'épimysium qui enveloppe le muscle est paré par le boucher lors de la découpe des morceaux de viande, celui-ci n'est pas consommé (Maunier-Sifre, 2005). En revanche, le périmysium constitue une part majoritaire, 90% du tissu conjonctif intramusculaire (McCormick et al., 1994) ; il joue un rôle prépondérant dans le déterminisme de la texture de la viande. C'est d'ailleurs sur cet aspect que se sont penchées la plupart des études portant sur les relations entre les caractéristiques du tissu conjonctif et la texture de la viande. Plusieurs auteurs ont mené des études sur des outils permettant la prévision de la qualité de la viande par l'analyse d'image. Une étude par analyse d'image du tissu conjonctif a permis d'estimer le contenu en collagène total (Abouelkaram et al., 2003). La couleur et persillé ont été utilisés pour prédire la tendreté de la viande (Lu et al., 1998). Cette technique a été améliorée en y incorporant les paramètres de texture, permettant ainsi d'obtenir des valeurs de R^2 atteignant 0.70 (Li et al., 1999). Le but de la présente étude est de développer des outils de prévision de la tendreté de la viande en utilisant une technique d'analyse d'image basée sur une approche multi-échelle. On démontre ici que la tendreté est sensiblement liée aux caractéristiques du tissu conjonctif. Pour cela, à partir des images de muscle de bovin, nous avons extrait les paramètres les plus significatifs liés au tissu conjonctif qui ont été ensuite introduits dans le modèle de prévision de la tendreté.

2 Approche multi-échelle

2.1 Algorithme "à trous"

L'algorithme à trous a été développé par (Holdschneider et al., 1989). Nous avons exposé de façon détaillé les fondements de cet algoritme dans (El Jabri, 2008). Dans le cadre de cet article nous allons rappeler les éléments essentiels. On considère que les données discrètes $c_0(k)$ (plan lissé à la résolution 0) sont définies comme un produit scalaire à la position k entre le signal $f(t)$ et la fonction d'échelle $\phi(t)$:

$$c_0(k) =< f(t), \phi(t-k) > \tag{1}$$

La fonction d'échelle doit satisfaire l'équation de dilatation :
$\frac{1}{2}\phi(\frac{t}{2}) = \sum_{l\in\mathbb{Z}} h_l \phi(t-l)$, où h est un filtre passe bas discret. Le plan lissé $c_j(k)$ à la résolution j et la position k est donné par :

$$c_j(k) =< f(t), 2^{-j}\phi(\frac{t-k}{2^j}) > \tag{2}$$

L'équation de dilatation permet d'obtenir la récurrence suivante :

$$c_{j+1}(k) = \sum_{l\in\mathbb{Z}} h_l c_j(k + l2^j) \tag{3}$$

L'approximation c_{j+1} est obtenue à partir du plan lissé c_j par convolution avec le filtre h et un pas de 2^j entre coefficients créant ainsi des trous dans le filtre d'où la dénomination de l'algorithme.
En pratique, $2^j - 1$ zéros sont insérés dans le filtre h à chaque résolution j, ce qui permet d'obtenir, par convolution avec le signal de départ, le plan lissé c_{j+1}.

Nous considérons ici une fonction ondelette ψ (filtre passe haut) telle que : $\frac{1}{2}\psi(\frac{t}{2}) = \sum_{l\in\mathbb{Z}} g_l \phi(t-l)$, nous avons :

$$w_j(k) =< f(t), \frac{1}{2^j}\psi(\frac{t-k}{2^j}) > \tag{4}$$

Nous obtenons une récurrence similaire à celle de l'équation (3) :

$$w_{j+1}(k) = \sum_{l\in\mathbb{Z}} g_l w_j(k + l2^j) \tag{5}$$

Le plus simple, pour le choix du filtre g, consiste à effectuer la différence entre approximations successives :

$$\hat{g}(\xi) = 1 - \hat{h}(\xi) \tag{6}$$

L'algorithme permet d'obtenir une pyramide de résolution contenant des approximations successives. La différence entre une approximation à l'échelle $j-1$ et celle d'une échelle immédiatement supérieure donne ce qu'on appelle le plan d'ondelette à l'échelle j :

$$w_j(k) = c_{j-1}(k) - c_j(k) = \frac{1}{2^{j-1}} < f(t), \phi(\frac{t-k}{2^{j-1}}) - \phi(\frac{t-k}{2^j}) > \tag{7}$$

Nous pouvons reconstruire le signal d'une façon simplifiée en additionnant les plans d'ondelettes avec la dernière approximation :

$$c_0 = c_p + \sum_{j=1}^{p} w_j \tag{8}$$

L'algorithme à trous est facilement étendu dans le cas bidimensionnel. On considère la fonction $\phi(t,u)$ telle que :

$$\frac{1}{4}\phi(\frac{t}{2},\frac{u}{2}) = \sum_{l,m\in\mathbb{Z}} h_{l,m}\ \phi(t-l, u-m) \tag{9}$$

Les données discrètes $c_0(l,m)$ sont définies comme le produit scalaire à la position l et m entre la fonction image $f(t,u)$:

$$c_0(l,m) =< f(t,u), \phi(t-l, u-m) > \tag{10}$$

Comme dans le cas monodimensionnel, la suite d'approximations successives de l'image est calculée par la récurrence :

$$c_{j+1}(k,k') = \sum_{l,m\in\mathbb{Z}} h_{l,m}\ c_j(k+l2^j, k'+m2^j) \tag{11}$$

On choisit une fonction ondelette ψ engendrée par la fonction d'échelle ϕ, soit :

$$\frac{1}{4}\psi(\frac{t}{2},\frac{u}{2}) = \sum_{l,m\in\mathbb{Z}} g_{l,m}\ \phi(t-l,u-m) \tag{12}$$

Les plans d'ondelettes sont obtenus par la récurrence suivante :

$$w_{j+1}(k,k') = \sum_{l,m\in\mathbb{Z}} g_{l,m}\ c_j(k+l2^j, k'+m2^j) \tag{13}$$

De nombreuses fonctions échelles répondent aux critères décrits ci-dessus. Pour utiliser facilement la transformation en ondelettes en 2D et en 3D, on choisit une fonction échelle à variables séparées : $\phi(t,u) = \phi(t)\phi(u)$ (ou $\phi(t,u,v) = \phi(t)\phi(u)\phi(v)$).

2.2 Modèle de Vision Multi-échelle (MVM) appliqué à l'analyse de la structure musculaire

Le concept de vision multiéchelle basé sur les ondelettes a été ennoncé par Meyer en 1992 (Meyer, 1992). Pour un signal donné quelconque, sa transformée en ondelettes mesure l'importance locale de ses différentes fréquences à travers des convolutions avec une série de filtres se déduisant les uns des autres par dilatation. Pour analyser les images de viande, nous avons choisi de mettre en oeuvre l'algorithme *à trous*. La fonction échelle employée est une fonction B-spline cubique (Unser & Aldroubi, 1992). Le filtre h associé à cette fonction correspond à un filtre binomiale d'ordre 4 :

$$h(0) = \frac{3}{8},\quad h(\pm1) = \frac{1}{4},\quad h(\pm2) = \frac{1}{16},\quad h(n) = 0 \quad si \quad |n| > 2$$

Dans le cas de nos images, il serait intéressant de pouvoir réaliser une analyse quasi isotrope. La seule fonction isotrope et séparable est la gaussienne, pourtant cette fonction ne satisfait pas l'équation de dilatation. Notons qu'il n'existe pas de fonction compacte isotrope satisfaisant l'équation de dilatation à deux dimensions. Nous avons choisi la fonction B-spline cubique qui est proche de la gaussienne, son utilisation en pratique conduit à une analyse quasi-isotrope. Les images que nous étions amenées à analyser proviennent d'un éclairage en lumière dans le visible. La figure 1 donne un exemple de la transformée en ondelettes d'une image de viande issu de ce type d'éclairage.

2.3 Détection des structures significatives

La détection des sructures significatives dans nos images de viande, consiste à seuiller les coefficients d'ondelettes directement à un seuil S convenablement déterminé en fonction du modèle de bruit. Ces sructures pour chacun des plans d'ondelettes w_j ont été obtenues en ne conservant que les valeurs $w_j(x,y) > S$ et en remplaçant les autres par 0. Nous avons choisi comme valeur pour S, le seuil universel donné par Donoho (Donoho et al., 1995), $S = \sigma\sqrt{2log(n)}$, où n désigne

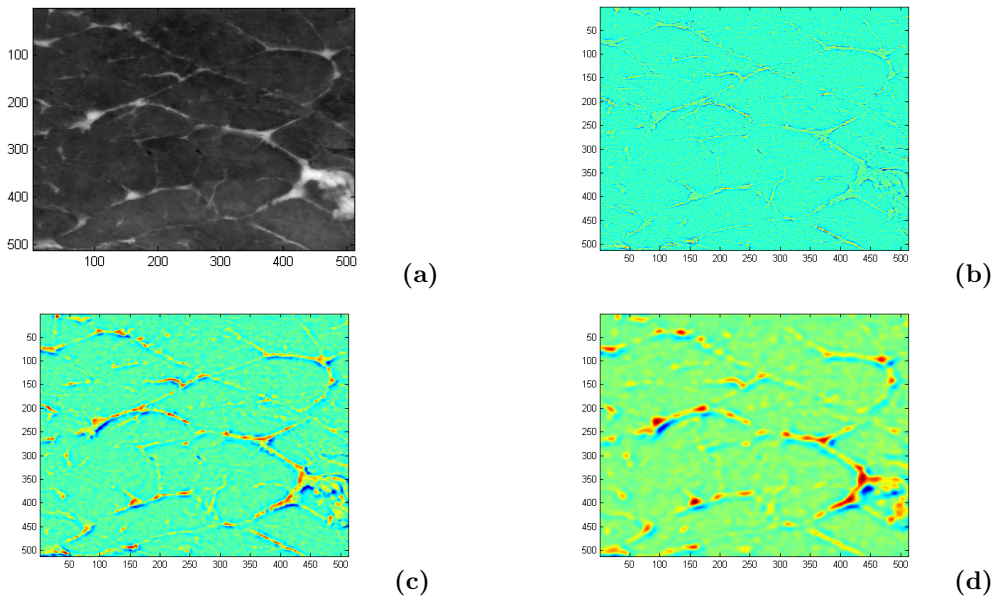


Figure 1: Résultat de la tansformée en ondelettes. Les 4 sous-figures représentent, dans le sens des aiguilles d'une montre, l'image brute (a) et ses plans d'ondellettes obtenues par l'algorithme "à trous" avec comme fonction d'échelle la B-spline cubique. Seules sont visualisées les trois premières échelles.

le nombre de points du signal.
On sait qu'un bruit gaussien ε centré de variance σ^2 n'est pas borné, néanmoins :

$$\lim_{n\to\infty} P\Big\{ \max_{1\leq i\leq n} |\varepsilon_i| > \sigma\sqrt{2log(n)} \Big\} = 0$$

C'est à dire qu'on peut considérer essentiellement que : $|\varepsilon_i| < \sigma\sqrt{2log(n)}$, ainsi on pourrait mettre à zéro tous les coefficients (w_j) qui pourraient être attribuable au bruit (Misiti et al., 2003). Notons que la décomposition en ondelettes de l'image à analyser est très creuse et ne subsistent que quelques coefficients de détail de niveau 1 (noté w_1) qui sont attribuables à ε. Une approche de premier ordre consiste à estimer le bruit à cette échelle (la plus fine, *i.e.* haute fréquence) dont le bruit domine le signal. Dans le cas d'un bruit gaussien, l'estimateur de l'écart-type est donné par la formule suivante :

$$\hat{\sigma} = MED(|w_1|)/0.6745$$

MED est la fonction médiane .

2.3.1 Algorithme de segmentation

Variables

- $h = [1, 4, 6, 4, 1]/16$ correspond au noyau de lissage associé à la B-spline cubique. Pour une fonction d'échelle bidimensionnelle $h_{2D} = T_h \times h$,

- j échelle d'ondelette,
- $c(j)$ approximation de l'image à l'échelle j,
- $w(j)$ plan d'ondelette à l'échelle j.

Algorithme

1. $c(0) = I$, I correspond à l'image de départ. Ainsi on considère que l'image à analyser est suffisamment régulière,
2. $j = 1$,
3. $c(j)$: convolution de $c(j-1)$ avec h,
4. $w(j) = c(j-1) - c(j)$,
5. Insérer $2^j - 1$ zeros entre les coefficients du filtre h *ie:* le filtre se dilate par un facteur de 2^{j-1} à l'échelle j créant ainsi des *trous*,
6. $j = j + 1$,
7. Aller à l'étape 3 jusqu'à obtenir le nombre d'échelles désiré J,
8. Aplliquer un seuil à chacun des plans d'ondelettes pour la détection des structures significatives. Cette sélection repose sur un test binaire (*hard thresholding*) vis-à-vis d'une valeur critique [1].

2.4 Résultats de segmentation

Nous illustrons les résultats de segmentation d'images de viande, issues de l'éclairage en lumière dans le visible, par l'exemple donné dans la figure 2. L'image originale est affiché à gauche. Le résultat de segmentation est affiché à droite. Les structures significatives ont été assez bien détectées avec cette méthode, donnant une image avec beaucoup d'information sur le réseau conjonctif du muscle (Périmysium).

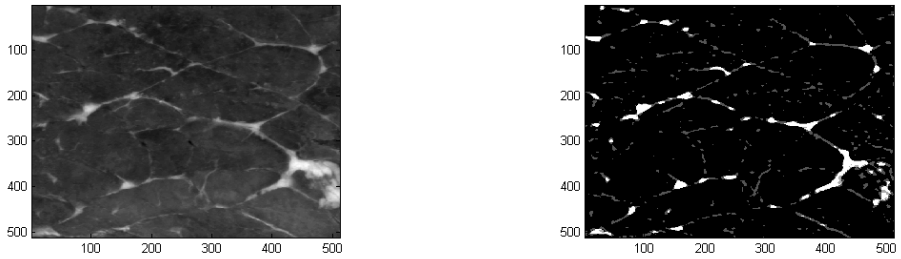


Figure 2: Résultat de segmentation

[1] La valeur critique que nous avons appliqué pour la segmentation des images de viande, issues de l'éclairage en lumière visible, se base sur le seuil universel de (Donoho et al, 1995).

3 Préparation des données d'analyse d'images

3.1 Extraction de données

Les images segmentées ont ensuite été binarisées. L'ensemble de données retenues de ce ces images représente les surfaces d'objets en pixels. Elles ont été calculées à l'aide d'un algorithme de quantification basée sur la notion de voisinage pixel. En 2D, on utilise très souvent la relation des 4 plus proches voisins, caractérisée par deux directions horizontale et verticale, ou celle des 8 plus proches voisins, caractérisée par trois directions horizontale, verticale et diagonale. La figure suivante décrit cette notion :

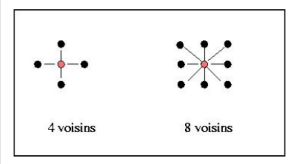


Figure 3: Notion de connexité : 4 et 8 plus proches voisins.

3.2 Transformation des données

C'est une étape de prétraitement de données qui consite en une action importante, prise avant le démarrage du processus d'analyse des données réelles (Famili at al., 1997). Il s'agit d'une transformation T qui transforme les vecteurs des données brutes des images en un jeu de données $Y = T(X_{in_i})$ avec :

- X_{in_i} représente la surface du n_i ème objet dans l'image i,
- $i = 1..n$, avec n : nombre d'observations, qui correspond aussi au nombre d'images traitées,
- $\sum_i n_i = N$ le nombre totale d'objets en tenant compte de toutes les images.

Cette transformation préalable vise à homogénéiser les données, dans le but de concentrer l'information. Il existe des logiciels qui déterminent automatiquement la transformation la plus adaptée utilisant l'algorithme de Box-Cox ou la loi de Taylor (Tufféry, 2007). Pour homogénéiser les données issues de l'analyse de nos images, nous avons opté pour une transformation *Log*.

3.3 Discrétisation des données

Afin de réaliser une discrétisation des données, il faut choisir le nombre de classes n_c et les bornes de classes. Le nombre de classes n_c a été choisi en se référant à la formule de Sturgers :

$$n_c = 1 + Log_2(N) \tag{14}$$

Les méthodes de bornes de classes supposent que le nombre de classes est fixé *a priori*. La technique de bornes de classes choisies consiste en une augmentation de l'amplitude des classes

selon une progression arithmétique de raison r. r étant la dynamique de la série divisée par l'addition des classes. Par exemple, pour 5 classes :

$$r = \frac{Max - Min}{1 + 2 + 3 + 4 + 5} \tag{15}$$

Ainsi, les bornes de classes seront : Min, Min+r pour la première classe, Min+r, Min+r+2r pour la deuxième classe et ainsi de suite. Cette méthode est connue sous le nom de la méthode par progression arithmétique.

3.4 Mise en tableau de données

Une fois que le nombre de classes et le mode de discrétisation sont définis, l'étape suivante consiste à identifier, pour chaque image de mucle, la distribution de ses données d'analyse d'images suivant discrétisation choisie. Celle-ci est basée sur l'ensemble de données (tailles d'objets), provenant de la même modalité d'acquisition (éclairage en lumière visible). Les données en question caractérisent les images de coupes de muscles. Elles sont présentées sous forme de tableau (tableau 1) :

Variables / Individus	$C_1 = [a_0, a_1]$	...	$C_j =]a_{j-1}, a_j]$	...	$C_{n_c} =]a_{n_c-1}, a_{n_c}]$
Muscle 1					
Muscle 2					
...					
Muscle i			$P(T(X_{ik}) \in C_j)$		
...					
Muscle n					

Table 1: Tableau de données

Chaque muscle est caractérisé par sa distribution de tailles d'objets dans l'image. Les données de ce tableau vont servir à une analyse statistique de données dans un objectif prédictif. Ci-après la définition de chaque terme générale de ce tableau :
X_{ik}: la surface d'objet en *pixel* de l'image du muscle i, retenue après traitement d'image. k représente le nombre d'objets dans l'image.
T: la transformation effectuée sur les données.
C_j: classe j d'objets définie après étape de discrétisation.
P: représente la proportion d'objets appartenant à C_j.

4 Application à la prédiction de la tendreté de la viande bovine

4.1 Matériel animal et imagerie

Les analyses ont été effectuées sur des échantillons de muscles de bovin de deux races : Holstein et Salers. Le type de muscle étudié est le *Semimembranosus* (SM). L'étude était basée sur 20 animaux (11 Holstein) et (9 Salers). Nous avons travaillé sur des tranches de muscle, afin de mettre en évidence les composants du muscle, particulièrement le tissu conjonctif. Pour chaque

muscle d'animal, une image en noire et blan a été acquise (Abouelkaram et al. 2003). La tendreté de la viande a été mesurée par un panel de dégustateurs entraînés selon la méthode préconisée dans (Dransfield et al., 2003).

4.2 Analyse des données

Le modèle de prévision de la tendreté, établis en utilisant les paramètres d'images, a été déterminé par la méthode de Régression sur les Composantes Principales (RCP). Les composantes principales retenues pour le modèle finale ont été sélectionnées en se basant sur le critère de minimisation du CP de Mallows. La figure suivante illustre la variation de ce critère en fonction du nombre de composantes principales introduites dans le modèle.

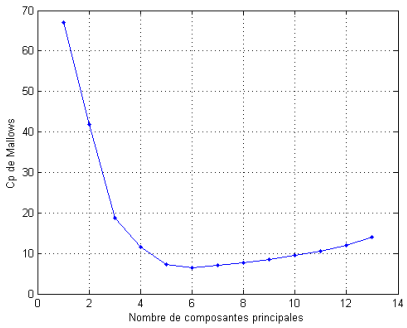


Figure 4: Cp de Mallows en fonction du nombre de composantes principales.

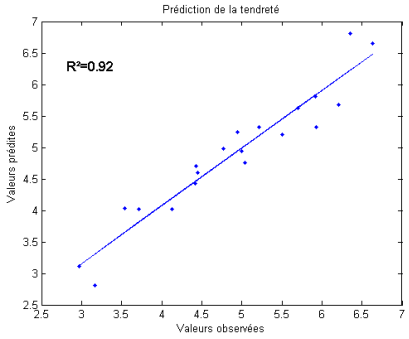


Figure 5: Prévision de la tendreté de la viande avec 5 paramètres d'images.

Cinq composantes principales parmi treize ont été choisies pour le modèle final de prévision. Le coefficient de détermination trouvé pour ce modèle est aux alentours de 0.92, caractérisant une bonne qualité d'ajustement. La figure 5 montre le nuage de points des valeurs prédites en fonction des valeurs observées. Ce modèle possède une bonne qualité prédictive.

5 Conclusion

Nous venons de montrer les résultats que nous avons trouvés avec notre approche basée sur une analyse d'images multiéchelle ainsi que les résultats d'analyse des données qui en découlent. La technique de segmentation développée permet d'extraire une information pertinente sur le réseau conjonctif du tissu musculaire. L'information retenue, à partir des images segmentées, est la distribution des tailles d'objets formant le périmysium. Cette analyse a été réalisée en s'appuyant sur la technique de nombre de classes de *Sturgers* et la technique de *progression arithmétique* comme méthode de bornes de classes. Cinq composantes principales ont été selectionnées pour le modèle finale permettant de prédire 92% de la variabilité de la tendreté de la viande.
Le modèle développé dans la présente étude permet une bonne prédiction de la tendreté de la viande. Il dépend en grande partie de la segmentation d'images ainsi que de la classification des tailles d'objets. Il serait aussi intéressant de passer en routine de grands échantillons de données incluant les facteurs biologiques tels que la race, le sexe ou encore l'âge de l'animal.

Bibliographie

Abouelkaram S., Berge P., Hocquette J. F., Culioli J. & Listrat A., 2003. Image study analysis of the relationship between collagen content and distribution of perimysial connective network in a bovine muscle. *Sciences des Aliments,* **231***, 166-170.*

Dononho D.L., Johnstone, I.M, Kerkyacharian, G. & Picard, D. (1995). *Wavelet Shrinkage: Asymptotia.*

Dransfield, E., Martin, J, F., Bauchard, D., Abouelkaram S., Lepetit, J., Culioli, J., Jurie, C. & Picard, B. (2003). Meat Quality and composition of three muscles from French cull cows and young bulls. *Animal Science,* **76***, 387-399*

El Jabri, M. (2008). Etude de l'organisation spatial du tissue conjonctif par analyse d'image basée sur une approche multiéchelle. Application à la prédiction de la tendreté de la viande bovine. *Thèse de doctorat, Université Blaise Pascal, N° d'ordre : D.U. 1831.*

Famili, A., Shen, W.-M., Weber, R. & Simoudis, E. (1997). Data processing and intelligent data analysis, *International journal on intelligent data* **1(1)***, 1-28.*

Holdschneider, M., Kronland-Martinet, R., Morlet, J. & Tchamitchian, Ph., 1989. A real time algorithm for the Signal Analysis with the help of the wavelet transform in Wavelets. pp.286-297 ed. J.M. Combes et al. Springer-Verlag Berlin.

Li, J., Tan, J., Martz, F., & Heymann, H. (1999). Images texture features as indicators of beef tenderness. *Meat science,* **53***, 17-22.*

Lu, J., Tan, J., Gao, X. & Gerrard, G.E. (1998). ASAE Mid-Central Conference (paper no. MC98131), St. Joseph, MI: ASAE.

Maunier-Sifre L. (2005). Organisation spatiale du tissu conjonctif intramusculaire : relation avec la texture de la viande bovine. *Thèse de doctorat, Université Blaise Pascal, N° d'ordre : 422.*

McCormick, R., J. (1994). The flexibility of the collagen compartment of muscle. *Meat Science,* **36***, 79-91.*

Meyer, Y. (1992). Les Ondelettes, Algorithmes et Applications., *Paris, Armand Colin.*

Misitti, M., Misitti, Y., Oppenheim G. & Poggi J-M. (2003). Les ondelettes et leurs applications, *Paris, Lavoisier.*

Nakamura, Y. N., Iwamoto, H., Ono, Y., Shiba, N., Nishimura, S. & Tabata, S. (2003). Relationship among collagen amount, distribution and architecture in the M. Longissimus thoracis and M. Pectoralis profundus from pigs. *Meat Science,* **64***, 43-50.*

Shensa, M.J. (1999). The discrete wavelet transform: Wedding the à trous and Mallat algorithms. *IEEE Transctions on Signal Processing.* **40***, 2464-2482.*

Tufféry, S. (2007). Datamining et statistique décisionnelle. *Technip.*

Unser, M. & Aldroubi, A. (1992). Polynomial splines and wavelets - A signal processing Perspective. *Wavelets : A Tutorial in Theory and Applications. ed. C.K. Chui Academic Press New York, 91-122.*

Session 12 : Posters / *Posters*

A comparison between a non parametric approach and a multivariate technique for evaluating the production of agribusiness products

Alibrandi Angela[1], Giacalone Massimiliano [2]

[1] *Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences (S.E.FI.S.A.S.T.), University of Messina, Via dei Verdi 75, 98122 Messina,*

E-mail : aalibrandi@unime.it

[2] *Department of Public Organization Law, Economy and Society (D.O.P.E.S), University of Catanzaro "Magna Graecia", Campus of Germaneto, 88100 Catanzaro,*

E-mail : maxgiacit@yahoo.it (*Corresponding author*) – Tel. +39 347/7734769

Home Address: Via Marchese Ugo, 74, 90141 Palermo (ITALY)

Abstract

The present study regards the analysis of the visits (and the relative disputes) for checking the preparation of agribusiness products, noticed by the ICQ on the whole national territory, occurred in the period 1999-2008.

The research has been performed on the available data for year and for sector. Particularly the examined sectors has been: Milk-Cheese, Feed and Integrators, Eggs, Honey and Meats. First of all, we have realized the "dispute rates", dividing the number of dispute to the total number of inspection visits for each sector and year.

In this way, we focused our attention both on the temporal variations of the dispute rates in the years and on the comparison among the different sectors of production.

Since we haven't guarantee about asymptotically valid results we used a nonparametric approach and a multivariate one,: in particular we applied the Cox and Stuart test for trend, the NPC test, the NPC Ranking, and the Correspondance Analysis-

Keywords : production sectors, inspection visits, Cox and Stuart test, NPC test, NPC Ranking, Correspondance Analysis.

Régression Bêta PLS

PLS Beta Regression

Frédéric Bertrand[1], Myriam Maumy-Bertrand[2], Nicolas Meyer[3] & Michèle Beau-Faller[4]

[1] *Université de Strasbourg et CNRS UMR7501*
E-mail : *frederic.bertrand@math.unistra.fr*

[2] *Université de Strasbourg et CNRS UMR7501*
E-mail : *myriam.maumy@math.unistra.fr*

[3] *Hôpitaux Universitaires de Strasbourg et Faculté de Médecine*
E-mail : *nmeyer@unistra.fr*

[4] *Hôpitaux Universitaires de Strasbourg et INSERM U682*
E-mail : *Michele.Faller@chru-strasbourg.fr*

Résumé
De nombreuses variables d'intérêt, comme par exemple des résultats expérimentaux ou des indicateurs économiques, s'expriment naturellement sous la forme de taux, de proportions ou d'indices dont les valeurs sont nécessairement comprises entre zéro et un. La régression Bêta permet de modéliser ces données avec beaucoup de souplesse puisque les densités des lois Bêta peuvent prendre des formes très variées. Toutefois, comme tous les modèles de régression usuels, elle ne peut s'appliquer directement lorsque les prédicteurs présentent des problèmes de multicolinéarité ou pire lorsqu'ils sont plus nombreux que les observations. Ces situations se rencontrent fréquemment de la chimie à la médecine en passant par l'économie ou le marketing. Pour circonvenir cette difficulté, nous formulons une extension de la régression PLS pour les modèles de régression Bêta. Celle-ci, ainsi que plusieurs outils comme la validation croisée et des techniques bootstrap, est disponible pour le langage R dans la bibliothèque `plsRbeta`.
Mots-clés : Régression Bêta. Régression PLS. Régression Bêta PLS. Validation croisée. Techniques bootstrap. Langage R.

Abstract
Many responses, for instance experimental results or economic indices, can be naturally expressed as rates or proportions whose values must lie between zero and one. The Beta regression often allows to model these data accurately since the shapes of the densities of Beta laws are very versatile. Yet, as any of the usual regression model, it cannot be applied safely in case of multicollinearity and not at all when the model matrix is rectangular. These situations are frequently found from chemistry to medicine through economics or marketing. To circumvent this difficulty, we derived an extension of PLS regression to Beta regression models. It, as well as several other tools, such as cross validation or bootstrap techniques, is available for the R language in the `plsRbeta` package.
Keywords : Beta Regression. PLS Regression. PLS Beta Regression. Cross validation. Bootstrap techniques. R language.

1 Introduction

La régression PLS a déjà été étendue avec succès aux modèles linéaires généralisés par Bastien *et al.* (2005) et aux modèles de Cox par Bastien (2008).

Nous proposons ici une extension de la régression PLS aux modèles de régression Bêta. En effet, l'intérêt pratique de la loi Bêta a été plusieurs fois affirmé par exemple par Johnson *et al.* (1995, p. 235) : "Beta distributions are very versatile and a variety of uncertanties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications".

Plusieurs articles récents se sont intéressés à l'étude de la régression Bêta et de ses propriétés. Mentionnons en particulier, l'article de Ferrari et Cribari-Neto (2004) pour une introduction à ces modèles et ceux de Kosmidis et Firth (2010), Simas *et al.* (2010) et Grün *et al.* (2011) pour des extension ou l'amélioration des techniques d'estimation de ces modèles.

Nous supposerons dans la suite que la réponse étudiée est à valeurs dans l'intervalle $[0;1]$. Le modèle que nous proposons peut bien sûr s'utiliser dès que la réponse Y est à valeurs dans un intervalle borné $[a;b]$, avec $a<b$ fixes et connus, en étudiant $(Y-a)/(b-a)$ à la place de Y.

2 Régression Bêta PLS

2.1 La régression Bêta

Lorsqu'elle est non nulle, la densité de la loi Beta(p,q) est donnée par :

$$\pi(y;p;q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1},\quad 0<y<1 \tag{1}$$

avec $p>0$, $q>0$ et $\Gamma(\cdot)$ la fonction gamma d'Euler. Si Y suit une loi Beta(p,q), son espérance et sa variance sont égaux à :

$$\mathbb{E}(Y)=\frac{p}{p+q}\quad\text{et}\quad \mathrm{Var}(Y)=\frac{pq}{(p+q)^2(p+q+1)}\cdot \tag{2}$$

Afin de pouvoir appliquer des techniques similaires à celles utilisées pour les modèles linéaires généralisés par McCullagh et Nelder (1989), Ferrari et Cribari-Neto (2004) proposent de reparamétriser la loi Bêta de la manière suivante. En posant $\mu=p/(p+q)$ et $\phi=p+q$, c'est-à-dire

$p = \mu\phi$ et $q = (1-\mu)\phi$, l'Équation (2) devient :

$$\mathbb{E}(Y) = \mu \quad \text{et} \quad \text{Var}(Y) = \frac{V(\mu)}{1+\phi}$$

où $V(\mu) = \mu(1-\mu)$. Ainsi μ est la valeur moyenne de la réponse et ϕ peut être interprété comme un paramètre de précision puisque, pour un μ fixé, plus la valeur de ϕ est élevée, plus la variance de la réponse est petite. Avec ces nouveaux paramètres, la densité donnée à l'Équation (1) est égale à :

$$\pi(y;\mu;\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1. \tag{3}$$

Soit $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et distribuées suivant la densité donnée à l'Équation (3) de moyenne μ_t et de précision inconnue ϕ.

Nous obtenons le modèle de régression Bêta en supposant que la moyenne de Y_t, $1 \leqslant t \leqslant n$, peut s'écrire :

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_k \tag{4}$$

où $\beta = (\beta_1, \ldots, \beta_k)'$ est un vecteur de paramètres de régression inconnus ($\beta \in \mathbb{R}^k$) et $x_{t1,}, \ldots, x_{tk}$ sont les observations des k prédicteurs avec $k < n$ qui sont supposées connues et fixes. Enfin, $g(\cdot)$ est une fonction de lien strictement monotone, deux fois dérivable, surjective et définie sur l'intervalle $]0;1[$ et à valeurs dans $\mathbb{R}$. La variance de Y_t est une fonction de μ_t et de ce fait dépend de la valeur des covariables. Par conséquent, le modèle prend automatiquement en compte les éventuels défauts d'homoscédasticité.

Il existe plusieurs choix usuels pour la fonction de lien $g(\cdot)$. Par exemple, le lien logit $g(\mu) = \log\mu(1-\mu)$, le lien probit $g(\mu) = \Phi^{-1}(\mu)$, avec Φ la fonction de répartition de la loi normale centrée et réduite, le lien complémentaire log-log $g(\mu) = \log(-\log(1-\mu))$, le lien log-log $g(\mu) = -\log(-\log(\mu))$. Une étude détaillée de ces liens a été réalisée par McCullagh et Nelder (1989) et Atkinson (1985) en a proposé d'autres encore. Ici aussi l'utilisation du lien logit permet d'interpréter l'exponentielle des coefficients des covariables en termes d'odds ratio.

2.2 La régression PLS

Considérons les variables centrées y, x_1, …, x_j, …, x_p. Soit X la matrice des prédicteurs x_1, …, x_j, …, x_p. La régression PLS est bien connue et décrite de manière exhaustive notamment par Höskuldsson (1988) et Wold *et al.* (2001). La présentation classique de la régression PLS est sous forme algorithmique. Nous n'en rappellerons que les éléments utiles pour la suite. La régression PLS est un modèle non-linéaire qui permet de construire des composantes orthogonales t_h obtenues en maximisant les quantités $cov(y, t_h)$. Soit T la matrice formée de ces composantes, nous avons :

$$y = T^t c + \epsilon, \tag{5}$$

où ϵ est le vecteur des résidus et ${}^t c$ le vecteur des coefficients des composantes, ^{t} désignant la transposée.

En posant $T = XW^*$, où W^* est la matrice des coefficients des variables x_j dans chaque composante t_h, nous avons l'expression directe de la réponse y à l'aide des prédicteurs x_j :

$$y = XW^{*t} c + \epsilon. \tag{6}$$

En développant le membre de droite de l'Équation (6), nous obtenons pour chaque composante y_i de y :

$$y_i = \sum_{h=1}^{H} \left(c_h w_{1h}^* x_{i1} + \cdots + c_h w_{ph}^* x_{ip} \right) + \epsilon_i, \tag{7}$$

H étant le nombre de composantes retenues dans le modèle final avec $H \leqslant \mathrm{rg}(X)$, H étant en général très inférieur au rang de X et p étant égal au nombre de variables contenues dans la matrice X. Les coefficients $c_h w_{jh}^*$, où $1 \leqslant j \leqslant p$, suivant la notation avec $*$ de Wold *et al.* (2001), traduisent la relation entre le vecteur y et les variables x_j à travers les composantes t_h.

2.3 La régression Bêta PLS

La régression Bêta PLS de la réponse y sur les variables $x_1, \ldots, x_j, \ldots, x_p$ avec H composantes $t_h = w_{1h}^* x_{i1} + \cdots + w_{ph}^* x_{ip}$ s'écrit :

$$g(\mu) = \sum_{h=1}^{H} \left(c_h \sum_{j=1}^{p} w_{jh}^* x_{ij} \right), \tag{8}$$

où μ est l'espérance de y. Le lien $g(\cdot)$ est à choisir parmi les liens logit, probit, complémentaire log-log, log-log, cauchit et log en fonction du type de données et de la qualité de l'ajustement du modèle au données. Les composantes PLS t_h sont orthogonales. L'algorithme permettant de déterminer les composantes PLS t_h d'un modèle de régression Bêta PLS est le suivant :

- Calcul de la première composante PLS t_1 :
  1. Calculer le coefficient a_{1j} de x_j dans la régression Bêta de y sur x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
  2. Normer le vecteur colonne a_1 : $w_1 = a_1 / \|a_1\|$.
  3. Calculer la composante $t_1 = 1/({}^t w_1 w_1) X w_1$.
- Calcul de la deuxième composante PLS t_2 :
  1. Calculer le coefficient a_{2j} de x_j dans la régression Bêta de y sur t_1 et x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
  2. Normer le vecteur colonne a_2 : $w_2 = a_2 / \|a_2\|$.
  3. Calculer la matrice résiduelle X_1 de la régression linéaire de X sur t_1.
  4. Calculer la composante $t_2 = 1/({}^t w_2 w_2) X_1 w_2$.
  5. Exprimer la composante t_2 en termes de prédicteurs X : $t_2 = X w_2^*$.
- Nous supposons construites les $h-1$ composantes $t_1, \ldots, t_{h-1}$.
  Calcul de la $h-$ème composante PLS t_h :
  1. Calculer le coefficient a_{hj} de x_j dans la régression Bêta de y sur $t_1, t_2, \ldots, t_{h-1}$ et x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
  2. Normer le vecteur colonne a_h : $w_h = a_h / \|a_h\|$.
  3. Calculer la matrice résiduelle X_{h-1} de la régression linéaire de X sur $t_1, t_2, \ldots, t_{h-1}$.

4. Calculer la composante $t_h = 1/({}^t w_h w_h) X_{h-1} w_h$.
5. Exprimer la composante t_h en termes de prédicteurs X : $t_h = X w_h^*$.

Il est facilement possible de modifier l'algorithme précédent pour pouvoir traiter les jeux de données incomplets.

3 Bootstrap, validations croisées et implémentation logicielle

3.1 Bootstrap

Nous supposons avoir retenu le nombre m adéquat de composantes d'un modèle de régression Bêta PLS de y sur $x_1, \ldots, x_j, \ldots, x_p$. Nous proposons l'algorithme suivant pour construire des intervalles de confiance et des tests de significativité pour les prédicteurs x_j, $1 \leqslant j \leqslant p$, à l'aide de techniques de bootstrap.
Soit $\widehat{F}_{(T|y)}$ la fonction de répartition empirique étant données la matrice T formées des m composantes PLS et la réponse y.

Étape 1. Tirer B échantillons de $\widehat{F}_{(T|y)}$.

Étape 2. Pour tout $b = 1, \ldots, B$, calculer :

$$c^{(b)} = ({}^t T^{(b)} T^{(b)})^{-1 t} T^{(b)} y^{(b)} \quad \text{et} \quad b^{(b)} = W^* c^{(b)},$$

où $[T^{(b)}, y^{(b)}]$ est le $b-$ème échantillon bootstrap, $c^{(b)}$ est le vecteur des coefficients des composantes et $b^{(b)}$ est le vecteur des coefficients des p prédicteurs d'origine pour cet échantillon et enfin W^* est la matrice fixe des poids des prédicteurs dans le modèle d'origine comportant m composantes.

Étape 3. Pour chaque j, notons Φ_{b_j} l'approximation de Monte-Carlo de la fonction de répartition de la statistique bootstrap de b_j.
Pour chaque b_j, des boîtes à moustaches et des intervalles de confiance peuvent être construits à l'aide des percentiles de Φ_{b_j}. Un intervalle de confiance peut être défini par $I_j(\alpha) = \Phi_{b_j}^{-1}(\alpha), \Phi_{b_j}^{-1}(1-\alpha)$ où $\Phi_{b_j}^{-1}(\alpha)$ et $\Phi_{b_j}^{-1}(1-\alpha)$ sont les valeurs obtenues à partir de la fonction de répartition de la statistique bootstrap de telle sorte qu'un niveau nominal de confiance de niveau $100(1-2\alpha)\%$ soit atteint. Afin d'améliorer la qualité de l'intervalle de confiance en termes de taux de couverture, c'est-à-dire la capacité de $I_j(\alpha)$ à fournir les taux de couverture attendus, il est possible d'utiliser plusieurs techniques de construction : normale, percentile ou BC_a (Efron et Tibshirani 1993 ou Davison et Hinkley 1997). Les intervalles ainsi obtenus ne sont pas conçus pour servir à réaliser des comparaisons multiples ou deux à deux et doivent être interprétés séparément.

3.2 Points forts de l'implémentation logicielle

La bibliothèque de fonctions `plsRbeta` pour le langage R implémente les modèles de régression Bêta PLS. Elle utilise la régression Bêta implémentée dans la bibliothèque `betareg` pour réaliser l'étape **1.**.

- Modèles de régression Bêta PLS avec des données complètes ou incomplètes.

- Choix du nombre de composantes grâce à différents critères AIC, BIC, R^2 modifié, arrêt de significativité de la composante t_{m+1} lorsqu'aucun des coefficients a_{m+1} n'est plus significatif dans le modèle ou en utilisant un critère Q^2 estimé par validation croisée.
- Validation croisée « repeated k-fold cross-validation » avec des données complètes ou incomplètes.
- Bootstrap des coefficients des prédicteurs pour des modèles de régression Bêta PLS avec des données complètes ou incomplètes. Différentes constructions d'intervalles, détaillées dans Efron et Tibshirani (1993) ou Davison et Hinkley (1997), sont disponibles et reposent sur la bibliothèque de fonction `boot` (Canty et Ripley 2009).

4 Exemple d'application en médecine

Les tumeurs cancéreuses représentent l'une des trois principales causes de mortalités dans le monde occidental. La compréhension des mécanismes des pathologies cancéreuses reposent actuellement sur l'étude des relations entre elles des anomalies génétiques acquises, apparaissant dans les tissus au cours du processus de la cancérisation. Ces anomalies sont fréquemment analysées par allélotypages, permettant de déterminer pour un nombre plus ou moins important de sites géniques, la présence ou non d'une modification du nombre de copie de chaque gène. La description multivariée de ces anomalies est informative sur le processus de cancérogénèse. Par ailleurs, l'ensemble de ces sites géniques porteur ou non d'anomalie peut être utilisé pour tenter de prédire certaines caractéristiques cliniques ou biologiques de la tumeur telles que le taux de cellules tumorales sur la biopsie d'une lésion. La modélisation dans un modèle statistique de taux, variable dont l'espace de variation est contenu dans l'intervalle fermée $[0; 1]$ comme variable prédite suggère l'utilisation d'une régression Bêta. Par ailleurs, les données d'allélotypage sont caractérisées par une fréquente colinéarité et par une proportion importante de données manquantes. De plus la matrice des données a souvent des dimensions $(i; j)$ telles que $j > i$, ce qui rend la matrice non-inversible, posant des difficultés dans l'ajustement d'un modèle de régression. La régression Bêta de type PLS que nous avons développée est donc particulièrement adaptée pour traiter les données d'allélotypage dans le contexte particulier de la prédiction d'une variable de type taux.

L'exemple proposé et détaillé sera celui de données d'allélotypage obtenues sur une série de 93 patients atteints de différents types de cancer du poumon. La variable prédite est le taux de cellularité tumorale du prélèvement peropératoire de la tumeur. En fonctions des critères, 6 ou 8 composantes sont à retenir aussi bien pour un lien logit que pour un lien log-log.

5 Conclusion et perspectives

Notre objectif a été de proposer une extension de la régression PLS aux modèles de régression Bêta, puis de la mettre à la disposition des utilisateurs du logiciel libre R.

Nous offrons ainsi la possibilité de travailler, pour modéliser des taux ou des proportions, avec des prédicteurs colinéaires, difficulté inévitable dans le cas de la modélisation des mélanges ou lors de l'analyse de spectres, de l'étude de données génétiques, protéomiques ou métabonomiques.

De plus, la régression Bêta PLS peut être aussi appliquée à des jeux de données incomplets. Il est également possible dans ce cas, comme dans celui des données complètes, de sélectionner le nombre de composantes par validation croisée « repeated k-fold cross-validation ».

Enfin, nous proposons des techniques bootstrap afin de, par exemple, tester la significativité de chacun des prédicteurs présents dans le jeu de données et ainsi valider les modèles construits.

Bibliographie

Bastien, Ph., Esposito Vinzi, V. & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, 48(1), 17-46.

Bastien, Ph. (2008). Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 91(1), 78-86.

Canty, A., & Ripley, B. (2009). `boot`*: Bootstrap R (S-Plus) Functions.* R package version 1.2-37.

Cribari-Neto, F. & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2), 1-24.

Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Ferrari, S.L.P. & and Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, 31(7), 799-815.

Grün, B., Kosmidis, I. & Zeileis, A. (2011). Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. Working Paper 2011-22. *Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics*, Universität Innsbruck.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2, 211-228.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed. New York, Wiley.

Kosmidis, I. & Firth, D. (2010). A Generic Algorithm for Reducing Bias in Parametric Estimation. *Electronic Journal of Statistics*, 4, 1097-1112.

McCullagh, P. & Nelder J.A. (1989). *Generalized Linear Models.* 2nd ed. Boca Raton: Chapman & Hall/CRC.

Simas, A.B., Barreto-Souza, W. & Rocha, A.V. (2010). Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, 54(2), 348-366.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.

Migration of engineered silver nanoparticles from PVC nanocomposite

Maeve Cushen*, Enda Cummins
Biosystems Engineering, School of Agriculture, Food Science and Veterinary Medicine, Agriculture and Food Science Centre, University College Dublin, Belfield, Dublin 4, Ireland.
*Corresponding author (Tel.: 00353 1 716 7458; e-mail: maeve.cushen@ucd.ie)

Nanotechnology is the manipulation of matter at the nanoscale, generally between 1 – 100 nm. Discoveries of unique nanomaterial properties have lead to novel applications in the food industry, one of which is antimicrobial food packaging materials. This type of value added packaging is likely to be most suited to high value perishable foods such as meat and poultry. These nanocomposites have antimicrobial nanoparticles either anchored to the packaging surface or incorporated within the polymer matrix. The objective of this study is to evaluate the potential for nanoparticle migration from this novel food packaging to food. Raw chicken breasts were packaged with plasticised polyvinyl chloride (PVC) silver nanoparticle nanocomposites. The chicken was subsequently analysed for potential migration of nanoparticles using inductively coupled plasma mass spectrometry (ICPMS). ICPMS analysis is a quantitative technique which combines a high temperature ICP source with a mass spectrometer. It is commonly used to detect *ultra trace* elements in complex matrices such as foods. Matrix interferences are minimized due to the high temperature of the ICP source. The limit of detection for silver using ICPMS analysis is < 5 μg/kg. The effects of nanoparticle size (diameter), nanoparticle concentration (w/w) in the nanocomposite (fill), time and temperature were also investigated in this experiment. The experiment was set up as a multi-factorial design. This maximised statistical power and resources. Treatments were the various combinations of the levels of different factors: diameter: 10 nm and 50 nm; fill: 0.5 % and 5%; time: 1.1, 2, 3.1 and 4 days; Temperature: 6.5, 7.2, 19.9 and 24.1. Results were statistically analysed using SAS. Particles were found to migrate ($p < 0.001$) at very low levels. All factors, except diameter, were found to significantly impact the resulting silver concentration of the chicken breasts in contact with the nanocomposites. Migration is influenced by fill ($p < 0.01$); time ($p < 0.01$) and temperature ($p < 0.05$). This study highlights the use of statistical methods to analyse experimental results which indicate possible migration of silver nanoparticles from a PVC nanocomposite to a food matrix. This study highlights the need to investigate possible human exposure resulting from such migration and whether such migration constitutes a human health risk.

Title:

Remote Difference Tests through Internet around the world: Sensodist in France, Italy, Madagascar & Vietnam

Authors & affiliations:

C. Dacremont[1], F. Sorrentino*[2], A. Pecourt[1], E. Monteleone[3], V. Ramaroson[4], D. Hoang Nguyen[5], D. Valentin[1]

[1]CSGA – AgroSup Dijon, France, [2]Biosystèmes, France, [3]University of Florence, Italy, [4]LAS/FOFIFA, Madagascar, [5]HoChiMinh city, University of Technology, Vietnam,

Abstract:

The objective is to explore the experimental conditions in which distant sensory tests may be conducted through internet. The experiment was conducted in four countries: France, Italy, Madagascar, and Vietnam with various internet access. We studied the impact of communication language: Domestic vs. English; length of the instructions: triangle test vs. A – not A test, and complexity of sample preparation: mineral water (no preparation) vs. instant coffee (to be prepared with hot water).

Six groups of about 60 participants were recruited in each country. They performed one triangle test and one A – not A test (two samples of each product) on either mineral water or coffee. Products were domestic mineral water with 0 vs. 11 mM NaCl and Nescafé instant coffee 1.6g plus 1.2 vs. 2.0g of sugar to be diluted with 70g of 70°C water. Participants in the two experimental groups collected the samples from a central location and performed the tests from home through internet with Fizz Web. On-screen instructions were in local language for one group and in English for the other group. The third group was a control. Participants performed the tests in sensory booths with Fizz Network/paper in domestic language.

Results for triangle test show no decrease in performance for tests performed by internet compared to control. Performance is similar whatever the language for instructions. An anecdotal difference was observed between countries for mineral water: the discrepancy being explained by water mineral content. Results for A - not A tests lead to similar conclusions.
Overall this study demonstrates the potential for remote tests via internet. Performance at difference tests are not impaired even when sample required careful preparation.

Impact de l'offre et de la spéculation sur la volatilité des marchés agricoles: Application au marché du cacao

Impact of supply and speculation upon the volatility of agricultural markets: Application to the cacao market

Dr. Marius-Cristian Frunza[1]*a*, Dr. Marius-Cristian Frunza[1] & Lionel Lecesne[2]

[1] *AgroSigma, 34 quai de Dion Bouton, 92800 La Défense Puteaux, France;* [a] *CNAM, 13 rue des Jeuneurs, 75002, Paris, France;*

E-mail : marius-cristian.frunza@polytechnique.org

[2] Dauphine University
E-mail : *lecesne.lionel@hotmail.fr*

Abstract
The aim of this paper is to assess the relationship between the volatility structure of the agricultural markets and the speculation effects juxtaposed over supply shortage. It is well accepted that the pure financial players (hedge funds) on the agricultural markets bring both liquidity and risk appetite, thereby implying an increase of the volatility. Thus we show via econometric studies that the agricultural markets have a non-Gaussian behavior exhibiting fat tails, and volatility clustering. Further we develop a volatility model depending on supply level, market liquidity and open interest. We apply this theoretic framework fro the Ivory Coast cacao crisis from the beginning of 2011. We show that the presence of the speculation during a supply shortage amplified the bullish trend of prices and the volatility clustering effect.
Keywords : Volatility, Structural breaks, Cacao crisis, convenience yield, changing regimes

1 Introduction

It is well accepted that the pure financial players (hedge funds) on the agricultural markets bring both liquidity and risk appetite, thereby implying an increase of the volatility. These effects become overwhelming in shortage situation like we witnessed on the Ivory Coast cacao crisis. Thus we try to show how the pure speculative trading could amplify both trend and volatility. On December 2nd 2011, the Electoral Commission declared that Ouattara had won the election. In response, the Gbagbo-aligned Constitutional Council rejected the declaration, and the government announced that country's borders had been sealed. Supplies from Ivory Coast have been disrupted since Ouattara imposed an export ban on January 24th 2011 in a bid to cut off funds for incumbent President Laurent Gbagbo.

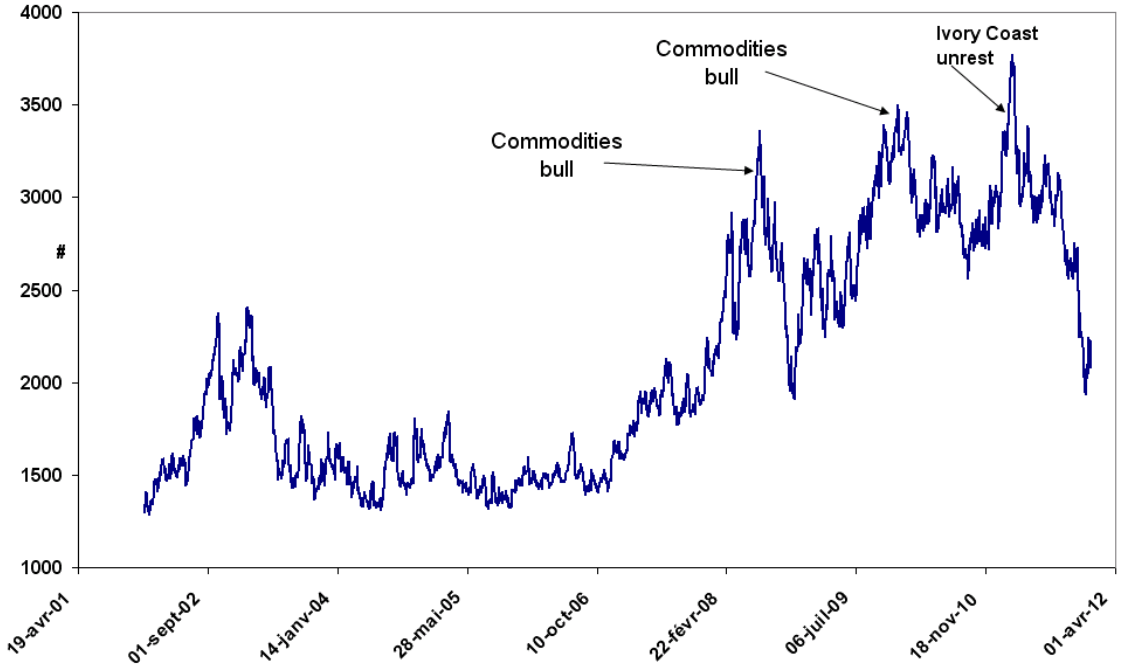


Figure 1: The Evolution of Cocoa Prices: 2001-2011

The main purpose of this paper is to evaluate which is the impact of the crisis on the cocoa market ? We analyze the futures and options prices since 01 January 2007, quoted on the International Continental Exchange. We compare two periods: January 2007 - April 2011 versus November 2010 - April 2011. The work is organized ias following. After an introduction we discuss the strucural breaks in both returns and volatility of the cocoa futures. In Section 3 we develop the topic of the convenience yield. Section 4 concludes.

2 Structural Breaks

Structural change is a statement about parameters, which only have meaning in the context of a model. To focus our analysis, we will discuss structural change for the case of linear model:

$$y_t = \alpha + \bar{x}_t + \epsilon \tag{1}$$

$$E(\epsilon^2) = \sigma^2 \tag{2}$$

The classical test for structural change is typically attributed to Chow (1960). His famous testing procedure splits the sample into two subperiods, estimates the parameters for each subperiod, and then tests the equality of the two sets of parameters using a classic F statistic. This test was popular for many years and was extended to cover most econometric models of interest. However, an important limitation of the Chow test is that the breakdate must be known a priori. Gregory and Hansen (1996) residual-based tests for structural breaks centers on deriving an alternative hypothesis. According to Gregory and Hansen (1996), the power of the Engle-Granger (1987) test of the null of no cointegration is substantially reduced in the presence of a break in the cointegrating relationship. To overcome this problem, Gregory and Hansen (1996) extended the Engle-Granger test to allow for breaks in either the intercept or the intercept and trend of the cointegrating relationship at an unknown time.

We build a model where we study the dependency of cocoa prices and a commodities index Goldmans Sachs Commodity index (GSCI). We study the relationship of this two times series during the Ivory Coast unrest in the beginning of 2011.

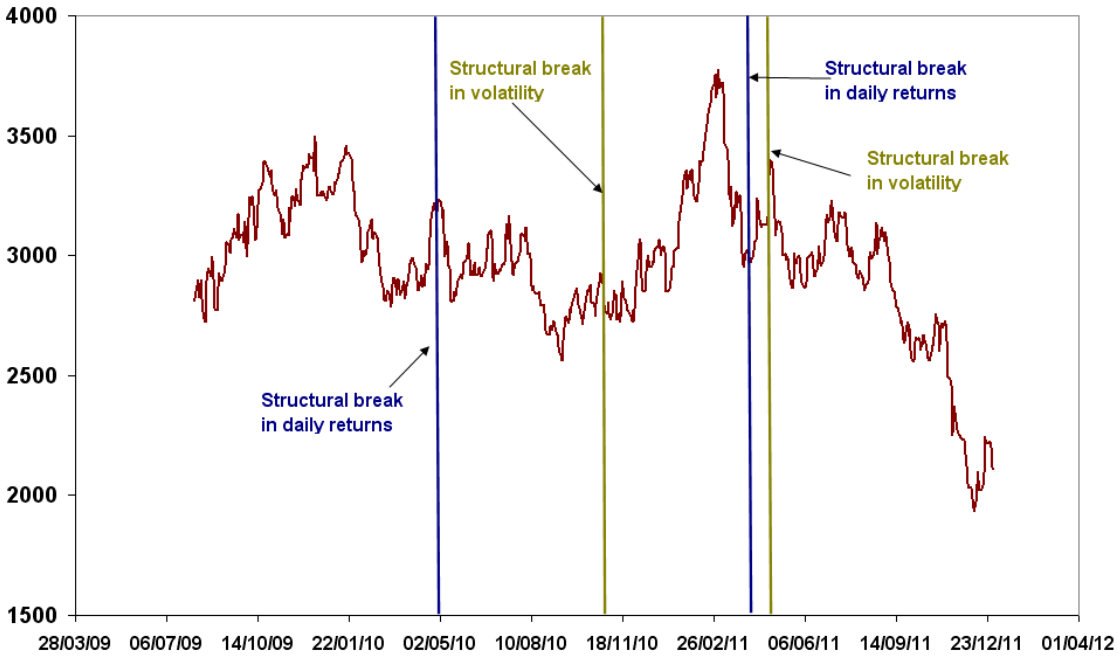


Figure 2: The Evolution of Cocoa Prices during the Ivory Coast crisis: 2010-2011

3 Convenience Yield

A second focus of this is around the convenience yield. If we apprehend the cocoa as a classic commodity like oil, gas or gold we should find similarities in economic interpretation. On the one hand, the agent has the option of flexibility with regards to consumption (no risk of commodity shortage). On the other hand, the decision to postpone consumption implies storage expenses. The net cost of these services per unit of time is termed the convenience yield δ. Intuitively, the convenience yield corresponds to dividend yield for stocks, thereby the price of a forward contract is given by:

$$F_{t,T} = S_t \cdot \exp((r_{t,T} - \delta_{t,T}) \cdot (T - t)) \tag{3}$$

where $F_{t,T}$ is the value at the moment t of the future contract for the maturity T, S_t is the spot value at time t, $r_{t,T}$ and $\delta_{t,T}$ are respectively the values of the rate and convenience yield for the maturity T. Here, the physical cocoa holders will not sell their stock to realize an arbitrage opportunity (by selling the quota and buying futures contracts). Consequently they "value" their owner-right and the convenience yield is a major element while modeling cocoa prices. Using the futures prices we construct a spot and a convenience yield timeseries using the following relationship. Thus we are able to find the evolution of the convenience yield during the crisis.

$$F(t, T_1) = S(t) \exp[r(t, T_1) - c(t))(T_1 - t)] \text{ and}$$
$$F(t, T_2) = S(t) \exp[r(t, T_2) - c(t))(T_2 - t)]$$

Our results shows that the markets passed from contango to backwardation in the early 2011 and come back to contango in march 2011, after the end of the Ivory Coast unrest. Thus the convenience yield become positive underlining the preference of markets for the spot or for short term delivery futures. The forward structures anticipated the end of the unrest as the market returned to contango in March 2011.

4 Results

Our econometric model shows that a structural break appeared on the cocoa market during the Ivory Coast political crisis on the cocoa. The direct impact was that the prices have raised significantly and retreated after the end of the crisis. But paradoxically, the observed physical volatilities did not exhibit a significant variation nor a switch in the volatility regime. On the other hand the implied volatility of the option market was higher after the beginning of the crisis. One possible explanation is that the physical market was not impact significantly market as operators have hedged their positions previously. Nevertheless the structural changes were probably generated by the impact of speculative trading that pushed up the drift on a constant volatility. Further they most-likely used options (for hedging or speculation), thereby increasing the neutral measure volatility.

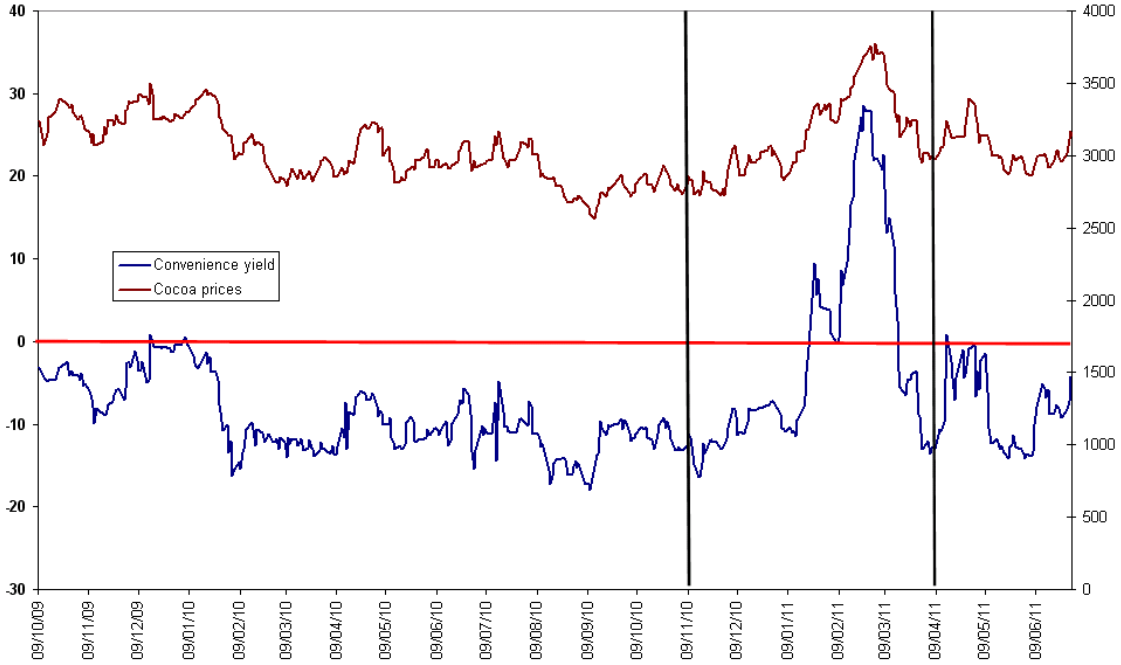


Figure 3: The Evolution of Cocoa Prices: 2001-2011

Bibliographie

Frunza, M. C., Guegan, D., Lassoudiere, A., (2010). Forecasting strategies for Carbon Allowances Prices: From classic APT to switching regimes. *International Review of Applied Financial Issues and Economics*, 2(23).

Une Analyse multi-variables robuste pour l'identification de typologies alimentaires
A robust multivariate analysis to identify dietary patterns

Habi Rachedi Fatiha[1], Rondeau Pascale, Marque Sébastien, Holmes Bridget Anna

Danone Research, Centre Daniel Carasso
RD 128, Avenue de la Vauve
F-91767 Palaiseau Cedex, France
[1] Email: *Fatiha.HABI-RACHEDI@danone .com*

Résumé

Les k-moyennes et la Classification Hiérarchique Ascendante (CHA), basée sur le critère de Ward de la variance minimale, ont été utilisés pour déterminer des typologies alimentaires sur un échantillon de 308 femmes, à partir de 27 groupes d'aliments prédéfinis. Trois paramètres statistiques : la pseudo-F statistique, le R-deux et le Critère Cubique de Classification (CCC)) nous ont permis de conclure que la méthode des k-moyennes est plus appropriée que la CHA. Ces paramètres nous ont permis également de déterminer le nombre (k) de clusters choisi à priori pour les k-moyennes.

Les catégories d'aliments très peu consommées comme les fruits de mer ou les fruits secs, génèrent des valeurs extrêmes. La méthode des k-moyennes étant très sensible aux valeurs extrêmes, les résultats obtenus dans un premier temps ne sont pas très satisfaisants. Pour palier à ce problème nous utilisons la méthode des k-moyennes sur des données corrigées par une technique proposée par Tukey en 1962, cette approche est nommée 'winsorized mean'.

Mots-clés : classification par les k-moyennes, classification hiérarchique, typologies alimentaires, boite à moustache, 'winsorized mean', test de Kruskal Wallis.

Abstract

The Ward's Agglomerative Hierarchical Clustering (AHC) and the k-means clustering method were undertaken in order to identify and characterize different dietary patterns in a sample of women living in the north of France (n=308). Three days of dietary data were collected for each individual. Each item of food and drink consumed was coded into one of 27 pre-defined food categories. The analysis is based on the mean of the 3 days. The k-means method was more appropriate according to three statistical parameters (pseudo F, R-squared and Cubic Clustering Criterion (CCC)), which also allowed us to select the number (k) of clusters established a priori for the k-means. Some clusters were characterized by extreme values of food groups consumed infrequently such as nuts and appetizers, shellfish and condiments and sauces. In order to avoid this effect we chose to cap food category variables at a given value using the winsorized approach which avoids the need to delete observations from the analysis. We used two techniques to determine the cap values: percentiles and box plots. The box plot approach was preferred since it provided us with more evenly sized clusters.

Four clusters of subjects with different dietary patterns were identified. For each cluster there was a negligible difference between the winsorized and the non-winsorized means for the food categories. A significant difference was observed in the mean age of subjects across the clusters using the Kruskal-Wallis test. Therefore we evaluated the effect of age on the clusters using the Van-Elteren (stratified Kruskal-Wallis) test. The use of k-winsorized means on dietary data enabled us to identify four dietary clusters based on reliable food categories with a good balance of subjects in each cluster.

Keywords: k-means clustering, hierarchical clustering, dietary patterns, box plot, winsorized mean, Kruskal-Wallis test.

1. Introduction

The objective of this analysis was to identify dietary patterns in a sample of 308 women aged between 18 and 60 years old recruited from one area in the North of France. The purpose of most pattern detection methods is to represent the variation in a data set in a manageable form by recognising classes or groups. There are basically two approaches that have been carried out in many types of studies with large data sets; Principal Component Analysis (PCA) and Clustering Analysis (CA). To date, little information is available to guide researchers in the choice of method to analyse dietary patterns and new approaches are not often explored. Dietary data were collected using three non-consecutive multiple pass 24-hour dietary recalls carried out over the telephone by trained dieticians. The 24-hour recall method involves the subject recalling all foods and drinks consumed the preceding day using a structured interview with specific neutral probes (Thompson and Byers, 1994). Data were collected on two weekdays and one weekend day (Sunday). Each item of food and drink consumed was linked to a food composition database and grouped into one of 27 pre-defined categories. The results presented here are based on mean intakes over the three days.

2. Dietary pattern analysis

The results of the PCA on the dietary data were not easy to interpret and more than three components were needed to have a good representation of the data. The CA method groups the subjects in a stepwise approach. Two types of CA were selected, the k-means (non hierarchical method) and Ward's Agglomerative hierarchical clustering (AHC based on Ward's Minimum-Variance). The grouping of subjects is done on the basis of similarities (dissimilarities) in eating behaviours, measured by distances (Euclidian) between the subject intakes.

Data were analysed using SAS® 9.2 and SAS® Enterprise Guide® 4.2 (SAS Institute Inc., Cary, NC, USA).

All food categories were standardized to a mean of 0 and a standard deviation of 1.

In the initial k-means algorithm, the k cluster centers are generated randomly, making it sensitive to starting conditions. An alternative way which greatly enhanced robust cluster recovery was developed by Milligan (Milligan, 1980). The FASTCLUS procedure in SAS is based on this modified algorithm.

There are three statistical parameters that indicate the measure of fit of the k-means or the Ward's Minimum-Variance methods: Pseudo-*F* statistic (PFS), Cubic Clustering Criterion (CCC) and all approximate expected R-squared (R^2). In general, the goal is to maximize each parameter.

The PFS parameter measures the separation among the clusters at the current level, CCC parameter tests the hypothesis that the data has been sampled from a uniform distribution on a hyper box and the R^2 parameter measures the variance proportion explained by the clusters (Sarle, 1983).

Larger positive values (more than 2) of the CCC show a larger difference from a uniform (no clusters) distribution.

The values of CCC for different numbers of clusters for Ward's AHC were all negative, indicating that the data distribution by clusters were close to uniform (no clusters) distribution.

In the k-means method, the number of clusters must be established a priori and therefore several solutions were compared with a varying number (N) of clusters (N from two to seven). The number of clusters was chosen based on the three statistical parameters described above considering also a good balance of subjects in each cluster.

N	PFS	R^2	CCC	min%	max%
2	16,88	0,05	-4,11	11%	89%
3	18,93	0,09	1,35	4%	60%
4	18,7	0,12	4,62	7%	44%
5	14,93	0,14	-0,9	3%	72%
6	15	0,16	2,8	1%	40%
7	15,59	0,18	8	1%	45%

min% and max% are the minimal and maximal percentages of subjects in the clusters.

Table 1: Statistical parameters of goodness of fit for k-means

Table 1 indicates that the CCC was good for k=4, 6 and 7 but the size of the clusters were not evenly balanced, with small (7% and 1%) and large (40%, 44%) clusters, the small ones were characterized by the extreme values of the food categories with a low percentage of consumers such as nuts and appetizers and shellfish.

2.1 Winsorized k-means

In order to avoid the effect of extreme values on the k-means clustering, the largest values for each food category were capped at a given value using the winsorized approach which avoids the need to delete observations from the analysis (Tukey, 1962, Mingxin, 2006).

There are two techniques to determine the cap values which we compared: percentiles and box plot with fences. The box plot with fences approach identifies extreme values in the tails of the distribution using quantities based on the inter quartile range. Values were capped at the upper inner fence: Q3 + 1.5*IQR where IQR= inter quartile range and Q3 the third quartile.

The Pseudo-F and the CCC parameters for k-means clustering with capping of the food categories using the 90, 95 and 99 percentiles (WPerct) and using the box plot (WBP) are shown in the following table:

	PFS		R^2		CCC		min%		max%	
Method N	WPerct	WBP	WPerct	WBP	WPerct	WBP	WPerct	WBP	WPerct	WBP
2	18,18	15,60	0,04	0,04	7,38	9,22	47	37	53	63
3	16,30	14,82	0,07	0,06	9,74	13,25	29	28	42	39
4	13,27	13,65	0,10	0,09	5,95	14,32	16	18	31	32
5	11,96	11,90	0,12	0,11	4,82	11,58	13	13	27	32
6	11,24	11,08	0,14	0,13	4,78	11,17	11	9	23	27
7	0,90	10,35	0,16	0,14	5,97	10,36	7	7	22	24

Table 2: Parameters of goodness of fit for winsorized k-means using percentiles and box plot

Table 2 indicates that the statistical parameters were better than those for k-means without correction (Table 1) and the clusters were more evenly sized according to the two techniques percentiles and box plot. Between the two correction techniques the Pseudo-F and the R^2 didn't differ greatly, however the CCC parameter is improved using the box plot technique. In addition, with the

box plot technique the correction is standardised across the food groups while the correction using percentiles can be done by the 90, 95 or the 99 percentiles according to the food category.

The optimum number of clusters (k) was selected as four (k=4) with a reasonable balance of subjects per cluster: Cluster 1, 58 subjects, Cluster 2, 94 subjects, Cluster 3, 100 subjects and Cluster 4, 56 subjects. For each cluster there was a negligible difference between the winsorized and the non-winsorized means for all food categories.

2.2 Characterisation of the clusters

2.2.1 Canonical Discriminate Analysis

Using canonical discriminate analysis (CDA), the food categories were transformed into three canonical variables (Can) which enables the visualization of the clusters and the associated food categories.

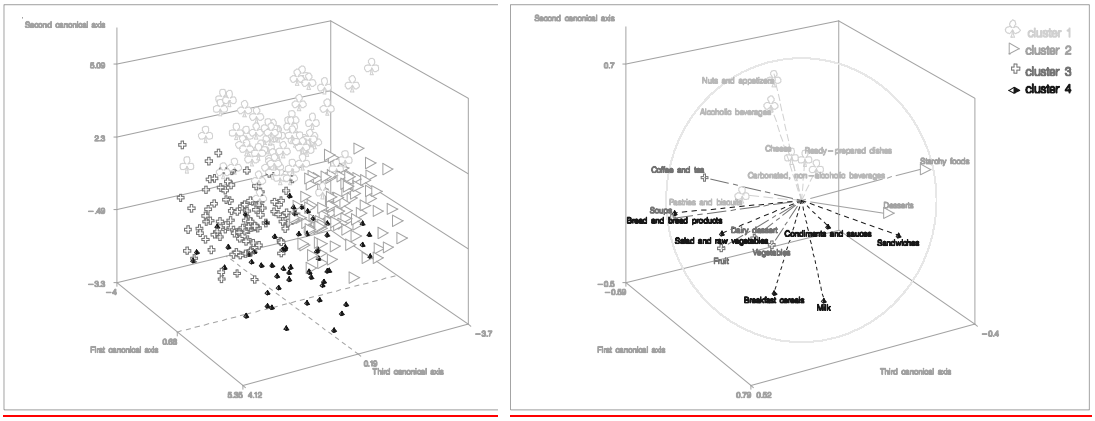


Figure 1: CDA displaying the division of the subjects (n=308) by the four clusters and the associate circle correlation

2.2.2 Tests

The dietary data were not normally distributed and therefore to test for differences in level of intake between the clusters the Kruskal-Wallis (KW) test was used. The Mann-Whitney test was used to test for differences between each pair of clusters. The Bonferroni correction was applied for the multiple comparisons. A significant difference in mean intakes across the four clusters was observed for 19 of the 27 food categories which appear on the circle correlation of CDA.

A significant difference was observed in the mean age of subjects (KW-test) across the four clusters. Despite this difference, adjusting for age with the Van-Elteren test (stratified KW-test) revealed only a few small differences in the food categories that characterized the clusters.

3. Conclusion

The classical method used to identify dietary patterns is PCA. Varraso et al (2011) demonstrated that PCA does not give stable results for small sample sizes (around 300); they proposed the factorial confirmatory analysis (FCA) to identify dietary patterns, this method is commonly used in social research. In this analysis we proposed two methods of cluster analysis to identify dietary patterns, a hierarchical (AHC) and not hierarchical clustering (k-means). We selected the k-means clustering for this sample recognizing a limitation of this method which is sensitivity to extreme values. The winsorized k-means is a solution to overcome this issue. The winsorized k-means is a robust method which is able to identify dietary patterns even in small sample sizes and without effect of age on almost all the clusters.

Bibliography

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ, p. 310:170.

Hu F (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol, 13, 3-9.

Hogg RV. (1974) Adaptive robust Procedures: A partial review and some suggestions for future applications and theory. J. Amer. Statist. Assoc. 69: 909-923.

Kim S (1992) The metrically trimmed mean as a robust estimator of location. Ann. Statist. 20.

Milligan GW & Cooper MC (1985) An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika, 50, 159-179.

Mingxin Wu (2006) Trimmed and winsorized estimators, A dissertation for the degree of Doctor of Philosophy, Michigan State University Probability and Statistics Department.

Sarle WS (1983) Cubic Clustering Criterion. SAS Technical Report A-108, Cary, NC. SAS Institute Inc.

Thompson F and Byers T (1994) Dietary assessment resource manual. J. Nutr. 124: 2245S-2317S.

Tukey J. W. (1962) The Future of Data Analysis, The Annals of Mathematical Statistics, 33, p. 18.

Varraso R, Garcia-Aymerich J, Monier F, Le Moual N, De Battle J, Miranda G, Pison C, Romieu I, Kauffmann F, Maccario J (2011). Aspects méthodologiques lies à l'estimation des typologies alimentaires en épidémiologie alimentaire. Nutrition clinique et métabolisme 25, S20-S52/Cahiers de nutrition et de diététique, 46, S20-S52.

Agrostat 2012. 12th European Symposium on Statistical Methods for the food industry

Cronobacter spp. exposure assessment after Pulsed electric field (PEF) treatment and storage of reconstituted powder infant formula milk (RPIFM).

M.C. Pina-Pérez, D. Rodrigo, A. Martínez*

Instituto Agroquímica y Tecnología de Alimentos. (IATA-CSIC). Avda Agustín Escardino, 7.C.P: 46980 Paterna (Valencia). SPAIN. e-mail: amartinez@iata.csic.es

Introduction. Non-thermal technologies to face *Cronobacter spp*. (González et al., 2006; Pina et al., 2007; Arroyo et al., 2010; Adekunte et al., 2010), are being recently considered alternatives to high temperature reconstitution (> 70ºC) of powder infant formula milk (PIFM), preserving the nutritional and quality value of the product which takes an important role on infant feed (FAO/WHO, 2006).

Objective. The present study assesses infants' exposure (N*f*) (healthy and inmunodepressed neonatal population) to *Cronobacter spp*. via the consumption of processed reconstituted PIFM with high-intensity pulsed electric fields (PEF).

Material and Methods. The baseline model includes H_0 value reported by FAO/WHO (2006) Lognormal [-3.84; 0.696] log cfu/g; PEF treatment at 10kV/cm-3000µs (Pina et al., 2007) and refrigerated storage at limit conditions 8ºC, 12h (FAO/WHO, 2006). The research work compares the effect of modification of input parameters (initial load (H_0); PEF treatment conditions; and temperature/time of RPIFM storage after treatment) on model outputs, describing four "what if" scenarios.

Monte Carlo simulation mathematical tool has been used to provide the outputs of the study: the most likely *Cronobacter spp.* concentration at the time of consumption (N*f*), based on inactivation (Pina et al., 2007) and growth models (Rosset, Noel and Morelli, 2007); and the neonatal daily probability of illness ($P_{inf}=1-e^{-rD}$) which is described using the single-hit model (Havelaar and Zwietering, 2004).

Results. The increase on PEF intensity from baseline to 40 kV/cm-360µs means the best scenario for preventing risk with a daily neonatal probability of illness defined by a Loglogistic distribution which 95th percentile is at 1.87 $x10^{-6}$ and 3.21 x 10^{-12} cfu/day for healthy and inmunodepressed neonates, respectively. When there is an increase on storage temperature from 8ºC to 37ºC, e.g. at enteral tube feeding, the estimated risk by consumption of 8 contaminated feedings per day is the highest $P_{inf} \approx 1$. These results point out the most influential factors on *Cronobacter spp* control under studied conditions (H_0> PEF intensity> storage temperature > storage time) and could be useful to define risk management strategies, mainly at hospitalary level.

Conclusions. PEF are presented as an effective non-thermal treatment to guarantee RPIFM microbiological safety even at relatively high H_0 (-1 log cfu/g) and inadequate storage conditions (8ºC, 24h)

References

(1)Adekunte A., Valdramidis V.P., Tiwari B.K., Slone N., Cullen P.J., O´Donell C.P., Scannell A. 2010. Int J Food Microbiol 142: 53-59.
(2)Arroyo, C., Cebrián G., Pagán R. y Condón S. 2010. Innovative F. Science Emerg Tech 11: 314-321
(3) FAO/WHO. 2006. Decisionalysis Risk Consultants, Inc. Ottawa, Ontario, Canada K1H6S3.
(4)González, S., G. J. Flick, F. M. Arritt, D. Holliman, and B. Meadows. 2006. *J. Food Prot.* 69:935–937.
(5)Havelaar A.H. and Zwietering.M. 2004. Trends in Food Science & Technology 15: 99-100
(6)Pina-Pérez,M.C., Rodrigo, D., Ferrer, C., RodrigoM., Martínez-López, A.2007 Int Dairy J 17(2007), 1441–1449.
(7)Rosset P, Noel V and Morelli E. 2007. Food Control 18: 1412-1418.

Modeling risk-benefit in a food chain: nutritional benefit *versus* microbial spoilage risk in canned green beans

AgroStat2012

Clémence Rigaux[a], Catherine M.G.C. Renard[b], Christophe Nguyen-the[b], Isabelle Albert[a], Frédéric Carlin[b]

[a]INRA, UR 1204 Met@risk, Méthodologies d'analyse de risque alimentaire , F-75005 Paris, France (e-mail : clemence.rigaux@paris.inra.fr)

[b]INRA, Université d'Avignon et des Pays de Vaucluse, UMR 408 Sécurité et Qualité des Produits d'Origine Végétale, F-84000 Avignon, France

Introduction

The same factors such as time, temperature, pH and O_2 concentration, affect both microbial changes and nutrient degradation in a food processing line. Microbial risk assessment is increasingly performed to evaluate the effect of processing fresh and processed foods on the microbial fate. A similar approach could be used to evaluate vitamin degradation during food processing operations. We therefore propose to model simultaneously the evolution of some bacteria and vitamin concentrations in a real food chain: a canned green bean processing chain. We aim to find a compromise on process parameters that optimizes nutritional benefit while keeping risk at an acceptable level.

Methodology

A risk and benefit second order Monte-Carlo simulation model was built, using literature data, data from a real green bean processing line, and expert opinions. Uncertainty and/or variability distributions were adjusted to model parameters using bayesian inference or maximum likelihood method, and performing a meta-analysis on heat resistance parameters.

Basic microbial and biochemical processes determining changes in bacterial and vitamin concentrations were defined at each step of the industrial process, as well as the most relevant chemical and physical environmental factors affecting those changes.

Results

The risk: Geobacillus Stearothermophilus

Non pathogenic thermophilic bacteria, responsible of spoilage

The canning process

The benefit: Vitamin C (AA and DHA)

Useful for immunity, cicatrization, antioxydant capacity

• Contamination	**Raw green beans**	• Initial concentrations of vitamin C
	Vegetable preparation (washing, trimming)	• Loss due to physiological evolution (green beans still "alive")
• Low inactivation due to high temperature (80-90°C) • Cross-contamination by water	**Blanching**	• Loss due to chemical reactions (destructured membranes) • Loss due to diffusion in the blanching water
• Possibility of cross-contamination by green bean debris	**Time interval (transfering)**	• Loss due to chemical reactions $AA \underset{}{\overset{+O2}{\rightleftharpoons}} DHA \rightarrow CGA$
• Cross-contamination via coverage liquid	**Addition of coverage liquid**	
• High inactivation due to high temperature processing (120-130°C)	**Sterilization**	• Loss due to chemical reactions
• Germination and possibility of growth, leading to food spoilage	**Storage**	• Loss due to chemical reactions • Loss due to diffusion in the juice

• **Inactivation models:**

Weibull primary model[1]: $\log_{10}\left(\frac{N_t}{N_0}\right) = -\left(\frac{t}{D}\right)^p$

Bigelow secondary model: $D = D_{ref}.10^{(T_{ref}-T)/z_T}.10^{(pH-pH_{ref})/z_{pH}}$

• **Growth models:**

Logistic primary model with delay λ=0: $N_t = \frac{N_{max}}{1+\left(\frac{N_{max}}{N_0}-1\right)\exp(-\mu_{max}.t)}$

Secondary model of Zwietering: $\mu_{max} = \mu_{opt}.\gamma(T).\gamma(pH)$

(N_t: contamination in cfu at time t, T: temperature, D: decimal reduction time, z: increase in temperature causing to a 10-fold reduction of D, p: shape parameter, μ_{max} and μ_{opt}: maximal and optimal growth rates)

• **Physiological evolution models:** $\frac{d[AA]}{dt} = -k$ and $\frac{d[DHA]}{dt} \approx 0$

• **Chemical reaction models[2]:** $\frac{d[AA]}{dt} = -k1.[AA].[O2]^{\beta}$

Arrhenius law : $k1,k3 = k_{ref}.\exp\left(-\frac{E_a}{R}\left(\frac{1}{T}-\frac{1}{T_{ref}}\right)\right)$ $\frac{d[DHA]}{dt} = k1.[AA].[O2]^{\beta} - k3.[DHA]$

• **Diffusion** at blanching : Fick law

(AA, DHA and CGA: Ascorbic, Deshydroascorbic and Ketoglutaric acids, O2: oxygen, []: concentrations, k,k1,k2: speed reaction coefficients, t: duration, T: temperature, Ea: activation energy, R: universal gaz constant)

Predicted evolution of *G.stearothermophilus* and vitamin C concentrations:

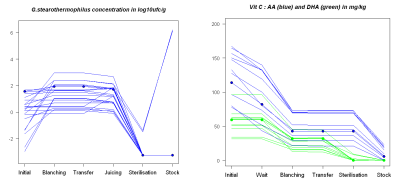


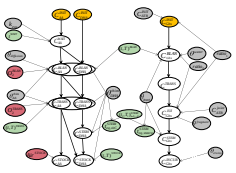
Figure 1. Several evolution scenarios in some green bean cans (obtained from different random variability and uncertainty values)

A simple sensitivity analysis shows that the most impacting factors seem to be:

• for *G.stearothermophilus:* heat resistance parameter D, green bean pH at sterilization, heat resistance parameter z_T, and cross contamination from blanching water.

• for vitamin C: initial concentration of vit C and proportion of AA, size of the residual headspace of air in the can, blanching and residence times, and kinetic rates of reaction $k_{ref,AA}$ and k.

Figure 2. Simultaneous risk-benefice model graph, confrontation to reality:



The only commands common to bacterial inactivation and to vitamin degradation are the blanching and sterilisation times and temperatures.

• Non stability predicted rate: ≈ 6%
→ rather high compared to industrial data

• Predicted vit C concentration: ≈ 12mg/kg
→ to be compared to ≈ 30mg/kg in litterature

Discussion and perspectives

➢ The model is still in progress, but only few parameters seem to have an important impact on both vitamin C and *G.stearothermophilus* final concentrations.

➢ The prediction of changes in vitamin concentrations and spoilage rate will be analysed and compared to measured data to validate the model. A more detailed sensivity analysis will permit to find the most impacting factors and possibly an optimization bivariate criteria.

References

(1) Mafart, P., Couvert, O., Gaillard, S., Leguerinel, I, 2002. On calculating sterility in thermal preservation methods: application of the Weibull frequency distribution model. International Journal of Food Microbiology, 72, 107-113.

(2) Penicaud, C., Etude et modélisation du couplage entre le transfert d'oxygène et les réactions d'oxydation dans les aliments au cours de leur conservation, PhD Thesis, University of Montpellier 2 Sciences et Techniques, France, nov 2009.

Acknowledgements: This project is supported by grants from the Agence Nationale de la Recherche (ANR, France) as part of the Ribenut project.

A Comparison of Independent Components Analysis with Principal Components Analysis Application to a Mid InfraRed data set

Delphine Jouan-Rimbaud Bouveresse[1,2], Douglas N. Rutledge[1,2]

1. INRA, UMR 1145 Ingénierie Procédés Aliments, 16 rue Claude Bernard, F-75005 Paris
2. Agroparistech, UMR 1145 Ingénierie procédés Aliments, 16 rue Claude Bernard, F-75005 Paris
delphine.bouveresse@agroparistech.fr - douglas.rutledge@agroparistech.fr

Independent Components Analysis (ICA)

✓ ICA aims to extract the pure signals and their proportions from a set of mixed signals.
✓ Each extracted signal is called an Independent Component (IC).
✓ All ICs are **statistically independent** from each other.

$$X = A\,S$$

where:
- **X** (*n*×*p*) is the original spectral data matrix (spectra on the rows)
- **A** (*n*×*k*) is the **mixing matrix** (the contribution of each pure signal to each row spectrum in **X**) → *Scores matrix*
- **S** (*k*×*p*) is the matrix of *k* **independent source signals** (on the rows), the independent components → *Loadings matrix*
- *k* is the number of calculated ICs

Principal Components Analysis (PCA)

✓ PCA aims to find the directions of greatest dispersion in the samples in the space of the variables.
✓ Each axis is called a Principal Component (PC).
✓All PCs are **orthogonal** to each other.

$$X = T\,P^{T}$$

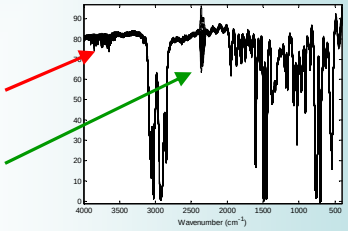
where:
- **X** (*n*×*p*) is the original spectral data matrix (spectra on the rows)
- **T** (*n*×*A*) is the **scores matrix** (coordinates of the samples projected onto the PCs)
- **P** (*p*×*A*) is the **loadings matrix** (direction of the PCs in the space of the original variables)
- *A* is the number of calculated PCs

The data

MIR spectra of a single Polystyrene film, recorded on 100 different spectrometers, between 4000 and 400 cm^{-1}[1].

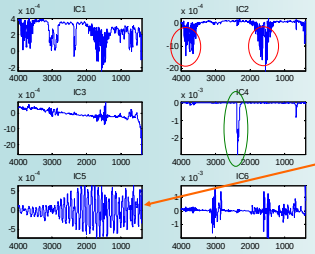
Peaks of H_2O due to varying atmospheric humidity levels

Peak of CO_2 due to varying atmospheric CO_2 levels



ICA Results

Loadings on IC1 to IC6 (*x*-axis = cm^{-1})

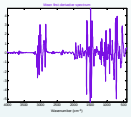


Red ovals: H_2O peaks Green ovals: CO_2 peak

IC2: H_2O spectrum

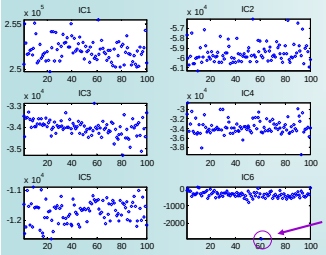
IC4: CO_2 spectrum

IC6 looks like a first-derivative spectrum



Oscillations like an interferogram, due to variations in the optical path

Scores on IC1 to IC6



Loadings vectors like a first derivative often reflect spectral shift in the data

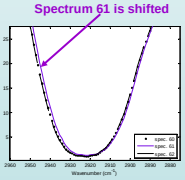
Investigation of the scores corresponding to IC6, PC5 and PC6 show that :

- IC6 accounts for sample 61

- PC5 and PC6 account for sample 61

The Figure below shows that spectrum 61 is shifted with respect to the others

Spectrum 61 is shifted

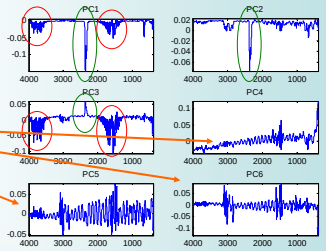


Sample 61

Sample 61

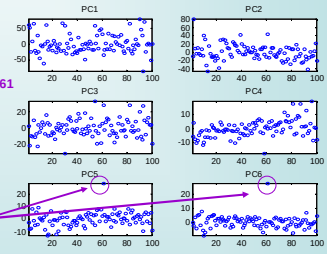
PCA Results

Loadings on PC1 to PC6 (*x*-axis = cm^{-1})



PC6 looks like a first-derivative spectrum. So does PC5 (behind the oscillations)

Scores on PC1 to PC6



Discussion and Conclusion

ICA and PCA are two methods to decompose experimental data into new calculated variables
The PCs are oriented so as to account for maximum dispersion in the data
The ICs are estimates of the pure source signals
The results show that 4 effects (H_2O level, CO_2 level, Variations in the optical path, Shifted spectrum N°61) are isolated in 4 distincts ICs
These four effects are also found by PCA, but are mixed together in different PCs (and each effect appears in several PCs).

Therefore, the ICA results are better for interpretation purpose.

[1] R. Aries, D. Lidiard, R. Spragg, *Spectroscopy* **5(3)** (1990), 41

Small Sample Size Capability Index for Assessing Validity of Analytical Methods

E. Rozet[1], B. Boulanger [2], E. Ziemons [1], R.D. Marini [1], Ph. Hubert [1]
Eric.Rozet@ulg.ac.be
[1] Analytical Chemistry Lab., University of Liège, Liège, Belgium.
[2] Arlenda s.a., Liège, Belgium.

INTRODUCTION & AIM

- The commonly used formulas to compute capability indices such as Cpk, will highly overestimate the true capability of analytical methods. Especially during methods validation or transfer, where there are only few experiments performed and, using in these situations the commonly applied capability indices to declare a method as valid or as transferable to a receiving laboratory will conduct to inadequate decisions.
- In this work, an improved capability index, namely Cpk-tol is proposed. Through Monte-Carlo simulations, they have been shown to greatly increase the estimation of analytical methods capability in particular in low sample size situations as encountered during methods validation or transfer.

MODIFIED CAPABILITY INDEX

The core problem when using capability indices in validation studies is the lack of sufficient data to estimate precisely the mean and standard deviation of the analytical method. In order to circumvent this and to take into account the uncertainty of the analytical method mean and standard deviation when computing Cpk, the use of tolerance interval should be preferred.

Additionally, the estimation of the mean and standard deviation of analytical methods should be made following the statistical model representing the way experiments have been performed. Method validation experiments or method transfer experiments are following a hierarchical or stratified sampling scheme that should be taken into account when computing analytical mean results and standard deviation and therefore to compute capability indices. For these stratified random sampling schemes commonly encountered during methods validations or transfers, a β-expectation tolerance intervals formula is given by Mee [1]:

$$[L,U]=[\hat{\mu}-k_E\hat{\sigma}_{IP};\hat{\mu}+k_E\hat{\sigma}_{IP}] \quad \text{(Eq.1)}$$

$$\text{where } k_E = t_{(df,(1+\beta)/2)}\sqrt{1+\frac{J\hat{R}+1}{JI(\hat{R}+1)}}$$

$$\text{with } df=\frac{(\hat{R}+1)^2}{\frac{\left(\hat{R}+\frac{1}{I}\right)^2}{J-1}+\frac{\left(1-\frac{1}{I}\right)}{JI}} \quad \text{and} \quad \hat{R}=\frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2}$$

t(df , γ) is the γ[th] percentile of a Student distribution with df degrees of freedom and $\hat{\mu}$ is the estimated mean of the results. The intermediate precision variance can be estimated using: $\hat{\sigma}_{IP}^2=\hat{\sigma}_B^2+\hat{\sigma}_W^2$. $\hat{\sigma}_B^2$ is the run-to-run or series-to-series variance and $\hat{\sigma}_W^2$ is the within-run or repeatability variance obtained with random one way ANalysis Of Variance (ANOVA) methodology [2]. J is the number of series performed and I the number of replicates per series.

The modified capability index proposed, Cpk-tol, is thus based on these tolerance intervals and is computed as it follows:

$$Cpk-tol=\min\left[\frac{USL-\hat{\mu}}{t_{(df,(1+\beta)/2)}\sqrt{1+\frac{J\hat{R}+1}{JI(\hat{R}+1)}}\hat{\sigma}_{IP}},\frac{\hat{\mu}-LSL}{t_{(df,(1+\beta)/2)}\sqrt{1+\frac{J\hat{R}+1}{JI(\hat{R}+1)}}\hat{\sigma}_{IP}}\right] \quad (Eq.2)$$

The Cpk index is computed with 3 at the denominator meaning that for a centred process the maximum fraction of non conforming result is about 2,700 dpm (precisely 2,699.796 dpm). In order to keep this same theoretical coverage of the distribution used with the Cpk index (i.e. $\pm 3\sigma$), the probability β of the Cpk-tol index is fixed to 0.9973.

SIMULATIONS

Independent validation results were generated from the random one-way ANOVA model described below:

$$X_{ij}=\delta+\phi_B+\varepsilon_W \quad \text{Eq. 3.}$$

where $X_{ij}$is the result of the j[th] measurement in series i, $\delta=\mu_{Lab}-\mu_T$ is the bias between the true (or reference or nominal) value of the result (μ_T) and the average value of the results of the laboratory (μ_{Lab}), $\phi_B$is the between series random effect supposed to be normally distributed $N(0,\sigma_B^2)$ and ε_W is the within-series (or repeatability) random error supposed to be independent and normally distributed $N(0,\sigma_W^2)$.

SIMULATIONS RESULTS

From the Figures 1a to 1f, it can be seen that the probability to exceed the true capability value (defined by the triangle in continuous line of Figures 1a-1f) by using the Cpk index is almost 50%, whatever the true capability value and whatever the sample size used in the method validation. Indeed the isoprobabilty curve (dashed line of Figures 1a-1f) that is almost exactly on the region which defines methods with known true Cpk value is the isoprobability curve of 50%.

By opposition when using Cpk-tol, Figures 1a to 1f show that the probability to exceed the true capability value is extremely low as the closest isoprobablity curve to the region defining methods with true capability indices of 1, 1.33 or 2 is the 10% one. Therefore there is only about 10% probability to declare a method capable when in reality it is not, i.e. the costumer risk is about 10% using such a capability index compared to the 50% risk observed for the classical Cpk index.

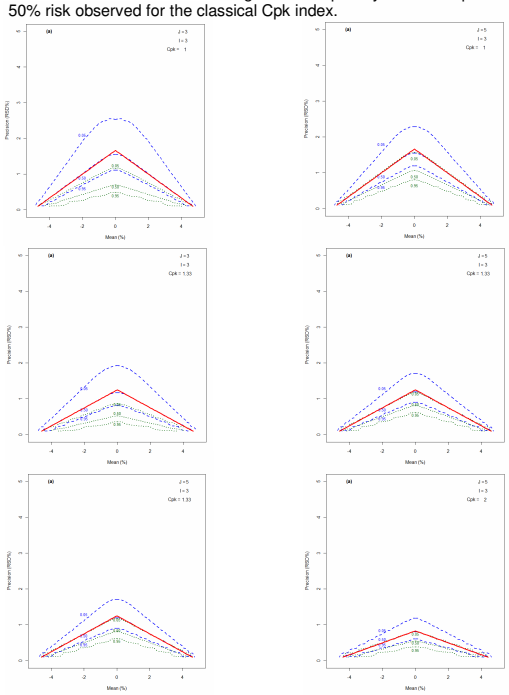


Fig. 1. Isoprobability contour measuring the probability that Cpk (dashed curves) and Cpk-tol (dotted curves) exceeds the true Cpk value of (a) 1, (c) 1.33 and (e) 2 with a design of 3 series and 3 repetitions per series or exceeds the true Cpk value of (b) 1, (d) 1.33 and (f) 2 with a design of 5 series and 3 repetitions per series.

While the Cpk-tol index controls well the costumer risk, the producer risk (i.e. the risk to conclude the method is not capable while it is truly capable) is relatively high. However this risk can be reduced by increasing the sample size of the method validation. This is shown by comparing Figures 1, 3, 5 obtained with 3 runs and 3 repetitions per run to Figures 2, 4 and 6 obtained with 5 runs and 3 repetitions per run for true Cpk values of 1, 1.33 and 2, respectively.

CONCLUSIONS

Finally, these simulations highlighted first the fact that using Cpk index to decide about the validity of analytical methods is highly controversial especially when using a method validation design of 3 runs and 3 replicates. Second, they showed that using Cpk-tol to make such a decision better controls the costumer risk, thus controls the risk for patients or public health risk, while the producer risk can be modulated by increasing sample size.

[1]. R.W. Mee, Technometrics 26 (1984) 251.
[2]. Searle S.R., Casella. G. and McCulloch C.E., Variance components (1992), Wiley.

Contact: Eric.Rozet@ulg.ac.be or:

ACKNOWLEDGMENTS

A research grant from the Belgium National Fund for Scientific Research (FRS-FNRS) to E. Rozet is gratefully acknowledged.

Reprints: http://hdl.handle.net/2268/104084/ or:

AN INNOVATIVE APPROACH TO SELECT THE PREDICTION MODEL IN THE DEVELOPMENT OF NIR SPECTROSCOPIC METHODS

E. Ziemons[1], J. Mantanus[1], E. Rozet[1,*], P. Lebrun[1], R. Klinkenberg[2], B. Streel[2], B. Evrard[3], Ph. Hubert[1]

[1] Laboratory of Analytical Chemistry, CIRM, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium.
[2] Galéphar Research Center M/F, Rue du Parc Industriel 39, 6900 Marche en Famenne, Belgium.
[3] Laboratory of Pharmaceutical Technology, CIRM, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium.
*E-mail: Eric.Rozet@ulg.ac.be

1. Introduction

FDA's Process Analytical Technology (PAT) aims at "improving the pharmaceutical development, manufacturing and quality assurance through innovation in product and process development, process analysis and process control" [1]. Taking into account its non-invasive, non-destructive character and fast data acquisition, near infrared spectroscopy is more and more integrated in the PAT system. However, implementation of a NIR quantitative method is performed using a time-consuming reference method and an iterative heuristic approach that will ultimately build a model allowing the prediction of the analyte of interest according to the product specifications.

2. Objectives

The aim of the present study was to develop an innovative approach based on the tolerance intervals and desirability indexes to select the most appropriate prediction model from a models plurality instead of using conventional criteria without objective decision rule such as R^2, RMSEC, RMSECV and RMSEP [2-3]. This new approach was performed on different steps of a real pharmaceutical manufacturing process: Active Pharmaceutical Ingredient (API) and moisture determination in pharmaceutical pellets [4-5].

3. Materials

Pilot batches of non-coated pharmaceutical pellets were manufactured at Galéphar Research Center M/F. The usual targeted formulation contains 57% w/w of API (T200999), this formulation will be further considered as the 100% active content formulation.Pilot batches containing 46 and 69% w/w of API were also manufactured (80 and 120% API formulations). Three independent batches were manufactured per formulation type. A confidential method previously developed and validated by Galéphar was used as reference method to determine the amount of API in batches of pellets.

Industrial batches of coated pharmaceutical pellets were manufactured at Galéphar Research Center M/F. Pellets samples were disposed on sieves in a cold room to reach the targeted moisture level. A thermogravimetric balance (HB43S Halogen balance, Mettler Toledo) was used as reference method to determine the amount of water in batches of pellets.

Samples were analysed by reflexion mode using a multipurpose analyzer Fourier transform near infrared spectrometer (MPA, Bruker Optics). Each spectrum was the average of 32 scans and the resolution was 8 cm^{-1} from 9000 to 4000 cm^{-1} (Figure 1).

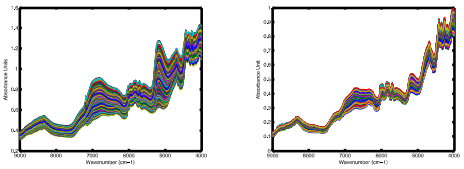


Figure 1. *NIR spectra of coated (moisture, left) and non-coated pharmaceutical (API, right) pellets for the calibration sets.*

4. Methodology

As shown in Table 1, sources of variability were introduced in the calibration set in order to meet the requirements of routine analyses and to guarantee the robustness of the NIR method: once a batch of pharmaceutical pellets was manufactured, the mixer-extruder and the sheronizer were succesively dismantled, cleaned and put together again before manufacture of a new batch. As different operators are likely to use the NIR method, the calibration protocol was designed for 2 operators. Different series of measurements of the same batches were perfomed to include day effect. In addition, samples were scanned under two different temperature conditions to include the spectral variation linked to temperature changes.

Table 1. *Calibration protocol ([a] API , [b] Moisture determination).*

Sources of variability	Calibration set [a]	Calibration set [b]
Analyte of interest (%)	80 - 100 - 120	0.5 - 1 - 2 - 4 - 5 - 10
Batches	9	3
Operators	2	
Series	4	
Temperature (° C)	25 - 35	5 - 25

Partial Least Squares (PLS) regresion using cross-validation (leave-one-out, contiguous block) was performed on the calibration set to build a prediction model.
Tolerance intervals and desirability indexes calculations were based on the cross-validation results from calibration sets. Dosing range, trueness, precision and Fitting Model Index (FMI) were used as desirability indexes. FMI is a global desirability function based on the 3 other indexes.

5. Results

Figure 2 and figure 3 display the desirability indexes and the accuracy profiles of a prediction model according to the PLS factor number for the determination of API and moisture, respectively.

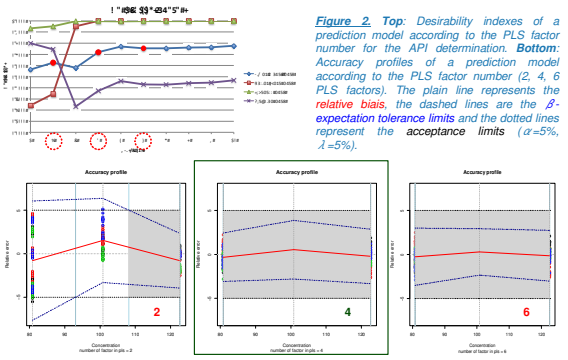


Figure 2. ***Top****: Desirability indexes of a prediction model according to the PLS factor number for the API determination.* ***Bottom****: Accuracy profiles of a prediction model according to the PLS factor number (2, 4, 6 PLS factors). The plain line represents the relative biais, the dashed lines are the β-expectation tolerance limits and the dotted lines represent the acceptance limits (α=5%, λ=5%).*

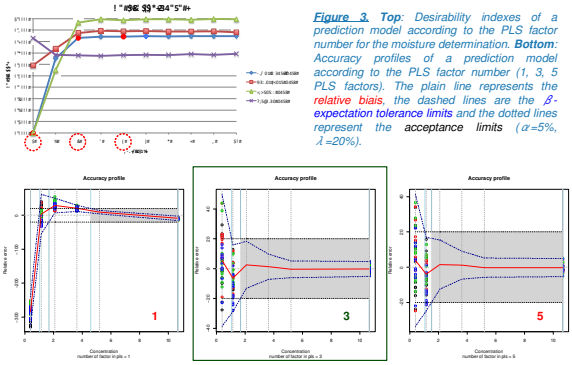


Figure 3. ***Top****: Desirability indexes of a prediction model according to the PLS factor number for the moisture determination.* ***Bottom****: Accuracy profiles of a prediction model according to the PLS factor number (1, 3, 5 PLS factors). The plain line represents the relative biais, the dashed lines are the β-expectation tolerance limits and the dotted lines represent the acceptance limits (α=5%, λ=20%).*

6. Conclusion

The innovative approach based on desirability indexes and tolerance intervals enables to select the most appropriate prediction model in full accordance with its very final goal, to quantify as accuratly as possible the analyte of interest. Desirability indexes, especially Fitting Model Index (FMI), increase significantly the objectivity of the decision process and reduce dramatically the development step and thus it eases the implementation of NIR quantitative methods on pharmaceutical manufacturing process.

References

[1] FDA, Guidance for Industry PAT, 2004.
[2] Hubert Ph. *et al.*, J. Pharm. Biomed. Anal., 36, 2007, 579-586.
[3] Rozet E. *et al.*, Anal. Chim. Acta, 591, 2007, 239-247.
[4] Mantanus J.*et al.*, Talanta, 2010, 1750-1757.
[5] Mantanus J. *et al.*, Anal. Chim. Acta, 2009, 186-192.

Check our publications on http://orbi.ulg.ac.be/

Les apports de la méthodologie Bagidis pour l'analyse de données métabonomiques : application à une étude spectroscopique de la Dégénérescence Maculaire Liée à l'Age

Advantages of the Bagidis methodology for metabonomics analyses: application to a spectroscopic study of Age-related Macular Degeneration

Catherine Timmermans[1], Pascal de Tullio[2], Vincent Lambert[3], Michel Frédérich[2], Réjane Rousseau[1] & Rainer von Sachs[1]

[1] *ISBA, Université catholique de Louvain. Voie du Roman Pays, 20, BE-1348 Louvain-la-Neuve.*
E-mail : *catherine.timmermans@uclouvain.be, rejane.rousseau@uclouvain.be, rvs@uclouvain.be*
[2] *CIRM, Université de Liège. Av. de l'Hôpital, 1, BE-4000 Liège.*
E-mail : *P.DeTullio@ulg.ac.be, M.Frederich@ulg.ac.be*
[3] *LBTD, Université de Liège et ophtalmologie, CHU de Liège. Av. de l'Hôpital, 1, BE-4000 Liège.*
E-mail : *vincent.lambert@chu.ulg.ac.be*

Résumé
La méthodologie Bagidis propose une mesure de distance entre spectres qui tient compte des variations horizontales et verticales affectant les pics spectraux, dans un cadre unifié. Cette méthode repose sur une décomposition des spectres dans une base d'ondelettes de Haar asymétriques. Ses atouts pour l'étude de spectres ^{1}H RMN en métabonomique sont illustrés ici dans le cadre d'une étude d'une maladie oculaire, la dégénérescence maculaire liée à l'âge. Une analyse visuelle, un modèle de détection de la maladie et une recherche de biomarqueurs sont proposés et comparés avec des méthodes reconnues.
Mots-clés : spectroscopie, métabolomique, non-alignement, ondelette, distance, classification

Abstract
The Bagidis methodology proposes a distance measure between spectra, that takes into account, in a unified framework, both horizontal shifts and amplitudes variations that might affect spectral peaks. The method relies on the expansion of the spectra in unbalanced Haar wavelet bases. Its opportunity for investigating ^{1}H NMR spectra in metabonomics is illustrated here in the framework of a study of an eye disease: age-related macular degeneration. Visual analysis, disease detection model and search for biomarkers are proposed here and compared with known methods.
Keywords : spectroscopy, metabolomics, misalignement, wavelet, distance, classification

1 Introduction

Metabonomics and metabolomics studies are analyses which aim at the simultaneous detection of every small weight molecule present in a biofluid, an organ or an organism (see *Nicholson and*

Lindon, 2008, for instance, for an introduction to this field). Those "small weight molecules", the molecular weight of which being typically less than 1500 daltons, are referred as *metabolites*. Numerous biological processes affect the concentrations of metabolites. Compounds of which the concentration is specifically modified by a given process are called *biomarkers* for that process. They can be seen as the *fingerprint* of the biochemical reactions underlying the given process. Identifying biomarkers leads to a better understanding of biological processes, and might help at designing efficient tools for its detection or prediction.

Metabolomics studies become more and more frequent in various scientific area (*Lindon et al, 2007*, gathers some example applications): detection of origin in the food industry, cultivars discrimination in agronomy, toxicological studies in environmental and pharmaceutical sciences, screening of drug candidates in pharmaceutical sciences, diagnostic tool in medicine, etc. The present work is concerned about metabolomics data investigation for biomedical purposes: we aim at discriminating blood serum samples between patients suffering from Age-related Macular Degeneration (AMD) and healthy patients. AMD is an ocular disease, that is a leading cause of vision loss in western countries amongst people aged fifty or older (see *Noël et al, 2007* , for instance). However, behind this specific application, the methodology we describe has a larger scope and might advantageously be applied on various metabolomics datasets.

Metabolomics datasets often consist in nuclear magnetic resonance spectra (an overview of this technique can be found in *Lindon et al, 2007*). From a statistical point of view, those spectra are curves with *sharp local patterns* ("spectral peaks"). Not only their amplitudes but also their locations and shapes are affected by noise, this noise arising from the biological variability of the samples but also from unavoidable changes of the experimental conditions of spectra acquisition. However, most multivariate statistical methods rely on the good alignment of the peaks to be compared (see *Timmermans and von Sachs, 2010*, for a discussion). Otherwise, false differences might be detected between the spectra. Realignment techniques, such as *dynamic time warping* have thus been developed, which can be applied as a preamble to the statistical analysis. Those realignment techniques are however imperfect.

In this context, the Bagidis methodology (*Timmermans and von Sachs, 2010*) aims at explicitly and simultaneously taking into account both amplitudes variations and horizontal shifts that might affect the patterns in a curve. This methodology relies on the definition of a semi-distance based upon the expansion of the curves in unbalanced Haar wavelet bases. For each curve, an unbalanced Haar wavelet basis is selected so as to hierarchically encode the patterns the curve is made of: the main patterns are supported by the first basis vectors, while subsequent basis vectors support less important ones. Every basis vector is associated to a specific level change in the curve. Such wavelet bases are associated to each of the spectra, using a sliding window to focus on successive smaller spectral zones. The distance between two spectra is measured as a weighted sum of hierarchically computed differences in both the locations and the amplitudes of the pattern from one spectra to another. Visualization tools, classification procedure and statistical tests can be used, that take into account the Bagidis semi-distance.

Given this, we investigate the AMD dataset as follows: we blindly discriminate blood serum samples from healthy and diseased patients; we build a nonparametric model for predicting the AMD health status from blood serum; we select statistically discriminative spectral peaks with a aim to identify AMD biomarkers; we discuss whether statistically discriminative spectral peaks are related to systematic amplitude changes or horizontal variations, or both simultaneously. At each step of our analysis, we discusses how Bagidis compares to a recently published statistical analysis of the AMD dataset (*Rousseau, 2011*) and show how this methodology can be used as

an useful complement to statistical tools usually used in metabolomics (*Rousseau et al, 2008*).

This paper is organized as follows. Section 2 gives an overview of the BAGIDIS methodology. Section 3 describes the AMD dataset. Section 4 discusses the statistical analysis of the AMD dataset. Section 5 concludes.

2 An overview of the Bagidis methodology

The acronym BAGIDIS stands for *BAses GIving DIStances*, as *basis expansion* is at the core of the methodology, the latest being centered on the introduction of a new *distance measure*. The BAGIDIS methodology has been introduced in *Timmermans and von Sachs, 2010*. Further investigation of its use in nonparametric functional statistics (*Ferraty and Vieu, 2006*) is provided in *Timmermans et al, 2011*. Key ideas are as follows.

As a first step of the procedure, each curve is decomposed in a set of short series using a sliding window, so that each windowed series should not contain two significant patterns of the same amplitude. The length of the window is problem-dependent an is denoted Dt. Each windowed segment x of each of the curves in the dataset is expanded in a particular wavelet basis, which is referred to as the *Best Suited Unbalanced Haar Wavelet Basis* (BSUHWB). We denote the expansion of x in this basis as $x = \sum_{k=0}^{Dt-1} d_k \psi_k$, where the coefficients d_k (hereafter the *detail* coefficients) are the projections of x on the corresponding basis vectors ψ_k. The BSUHWB basis is obtained using the *Bottom-Up Unbalanced Haar Wavelet Transform* (BUUHWT) proposed by *Fryzlewicz, 2007*.

An interesting property of the Unbalanced Haar wavelet bases expansions, is that the set of points $\{y_k\}_{k=1\ldots Dt-1} = \{(b_k, d_k)\}_{k=1\ldots Dt-1}$ determines the *shape* of x uniquely, the complete determination of x requiring the additional coefficient d_0 that encodes the mean level of the series. Furthermore, the BUUHWT induces an interesting property of hierarchy in the resulting BSUHWB expansion. The idea is that the ranking of the basis vectors of the BSUHWB reflects the decreasing importance of the patterns they encode, for the description of the global shape of x. The notion of *hierarchy* that we refer to is the hierarchy induced by the BUUHWT algorithm itself: by construction, x is encoded in its BSUHWB as the sum of a constant mean level (rank $k = 0$) and a linear combination of level changes, the few first (small rank indexes) encoding the most striking features of x, while the last ones (large rank indexes) are less important. In such a way, the *Bottom-Up Unbalanced Haar Wavelet Transform* allows for an automatic and unique hierarchical description of each of the segment into a segment-adapted orthonormal basis. The hierarchy makes the resulting bases comparable to each other, although different. Consequently, we propose to compare the segments through a weighted p-norm between their mapping $\{y_k\}$ into the location-amplitude space of their breakpoints and details coefficients:

$$d^{\text{BAGIDIS}}(x^{(1)}, x^{(2)}) = \sum_{k=1}^{Dt-1} w_k \left\| y_k^{(1)} - y_k^{(2)} \right\|_{\lambda p} = \sum_{k=1}^{Dt-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1-\lambda) \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}$$

with $p = 1, 2, \ldots, \infty$, with $\lambda \in [0; 1]$, and where w_k is a well suited weight function. As such, this semi-distance takes advantage of the hierarchy of the well adapted unbalanced Haar wavelet bases: breakpoints and details of similar rank k in the hierarchical description of each segment are compared to each other, and the resulting differences can be weighted according to that rank. As the breakpoints point to level changes in the segments, the term $\left| b_k^{(1)} - b_k^{(2)} \right|$ can be interpreted as a measure of the difference of location of the features, along the horizontal axis.

Being a difference of the projections of the segments onto wavelets that encode level changes, the term $\left|d_k^{(1)} - d_k^{(2)}\right|$ can be interpreted as a measure of the differences of the amplitudes of the features, along the vertical axis. Such a dissimilarity d^{BAGIDIS} is shown to be a semi-distance. It is computed for each windowed segment separately, for each pair of spectra of the dataset. The semi-distance between two spectra is then defined as the average value of the distances between its windowed segments.

In a prediction setting, weights w_k should ideally be 1 at rank k if that rank carries information for discriminating the series, and 0 otherwise. This is easily obtained using a cross-validation procedure. When no prediction criterion is at hand, or in order to get a first idea of how the dissimilarities do behave, we suggest in *Timmermans and von Sachs, 2010,* to *a priori* use the weight function $w_k = \frac{\log(Dt+1-k)}{\log(Dt+1)}$. This allows to associate a large weight to the comparison of features encoded at the first rank of the hierarchy, and a decreasing weight to the smaller features at the end of the hierarchy, which is empirically what we expect. The parameter λ actually defines a scaling in the *breakpoints-details* plane, and hence in the original units of the problem. Setting λ at its extreme values 0 or 1 allows to investigate the contributions of the breakpoints differences and details differences separately. In a prediction setting, λ can easily be optimized using cross-validation. Besides, the presence of this parameter allows the semi-distance to be robust with respect to scaling effects: if λ is optimized according to a given criteria (such as the mean square error of a prediction model), the relative dissimilarities between the series of a dataset will remain the same, whatever the scales of measurements along the horizontal and vertical axes, so that the predictive qualities of the model will not be affected by such a change in the units of measurements.

3 The AMD dataset

AMD is an ocular pathology, that can lead to rapid vision loss. It affects central fine vision, needed for reading, driving and face recognition, for instance. This disease exists in two distinct forms, one of which arising from an uncontrolled formation of new blood vessels (*angiogenesis*) under the *macula*, a part of the retina at the rear of the eye. The misknowledge of AMD anthology motivates the search for biomarkers in blood serum samples through a metabolomics approach (*Noël et al, 2007*). The AMD database was originally collected for a study lead by *de Tullio, Frédérich and Lambert* (Université de Liège). It consists in 200 blood samples, 100 of which arising from AMD patients and the other 100 arising from non-AMD patients ("control" patients), the AMD health status being diagnosed by an ophthalmologist. All AMD patients are aged over sixty and are followed by an ophthalmologist at *Centre Hospitalier Universitaire* in Liège, Belgium. Control patients are aged-matched patients in the same hospital, without any sign of ocular disease and not having a known history of AMD. The database also contains some additional general and clinical information. A complete description of this database can be found in *Rousseau* (2011).

A one-dimensional ^{1}H NMR spectrum was acquired from each blood sample, using a 500 MHz Bruker Avance spectrometer. A CPMG sequence with water pre-saturation was applied to attenuate broad signals arising from protein and water. Due to spectral acquisition problem, 6 AMD samples and 1 control sample were removed from the study. The resulting product of an ^{1}H NMR spectrum acquisition is a time signal called *Free Induction Decay* (FID). In order to chemically interpret the signal, each FID is converted in a spectral signal using a Fourier transform. Before and after this Fourier transform, several other pre-treatment of the

signals are also needed for the data to be statistically exploitable. In this study, we used the automatic pre-treatement procedure for metabolomics data which is advised and validated by *Rousseau, 2011*. It includes first order phase correction, suppression of the solvent, apodization by a scale function, apodization by an exponential function, Fourier transform, zero-order phase correction, setting to zero of negative values, warping, conversion in ppm scale, spectral window selection, bucketing, removal of undesired regions, spectral zone aggregation and normalization. More details on the acquisition procedure and on the pre-treatments can be found in *Rousseau, 2011*.

As a result of this procedure, the AMD dataset contains 193 spectra, of which 94 comes from AMD patients. Each spectra is a curve of 600 consecutive intensity measures in a spectral range going from 10 to 0.2 ppm.

4 Statistical Analyses

Except if mentioned otherwise, we make use of the Bagidis semi-distance with parameters $Dt = 25$, $p = 2$, $\lambda = 0.5$ and the default value of w_k. Results we obtain are compared with the recent results obtained by *Rousseau, 2011*. We see that additional insight into the data is gained by using the Bagidis methodology.

4.1 Visual analysis

Figure 1, *top left*, provides with the projections of the spectra on the first plane of a principal component analysis (PCA). This representation is to be compared with a *multidimensional scaling* (MDS) representation of the dataset, based upon the Bagidis semi-distance (with $\lambda = 0.5$) in Figure 1, *top right*. Multidimensional scaling is a projection technique that aims at preserving given distances between the observations in the dataset, so that the proximities of the data in the plane of projection can be interpreted -up to a certain degree- as "real" proximities of the data according to the chosen distance. MDS used jointly with the Euclidean distance corresponds to a PCA. In both representations, points are colored in different values according to their AMD health status.

We see that using Bagidis allows for a nearly optimal discrimination between AMD and non-AMD serum spectra, the distinction being essentially encoded by one single axis, while PCA detects an effect of the AMD health status but does not achieve such a clear discrimination. Furthermore, four outliers were detected in the data by *Rousseau, 2011* using visual inspection of this PCA representation, and a PCA for group-centered data. Those spectra are marked by triangles instead of points in Figure 1. They were removed of the dataset in *Rousseau, 2011*. We do not observe such a aberrant behavior for those spectra when using Bagidis. This might indicate that a problematic warping, resulting in misalignment, may be the cause of the aberrant behavior observed in the PCA. In this study, we do not discard those spectra from the database.

Figures 1 *bottom left and right* are MDS representations, obtained using a balance parameter λ fixed as 0 and 1 respectively. This allows to diagnose the effect of detail differences and breakpoint differences in the distance separately. Although some information on the AMD health status is clearly contained in the detail differences, we observe the major role of breakpoints location for discriminating the spectra. This might indicate a systematic peak appearance, shape modification of a peak, horizontal shift, or change of sign in the difference of amplitudes of neighbor peaks.

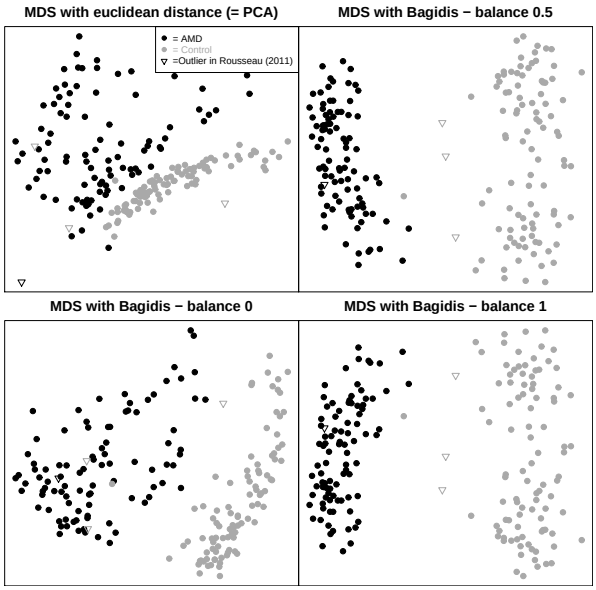


Figure 1: PCA representation, as compared with MDS representation using BAGIDIS with balance parameter $\lambda = 0.5\,, 0$ and 1 respectively. Points are colored according to the AMD health status of the corresponding projected spectrum.

4.2 AMD detection

We aim at predicting from the spectrum if a patient is affected by AMD or not. A training set of 150 spectra is randomly selected and a functional nonparametric discrimination model is adjusted (*Ferraty and Vieu, 2007*). This model is a k-nn predictor relying on a matrix of semi-distances between the spectra, with k being cross-validated. We consider the adjustment of such a model using BAGIDIS and compared its performances with those obtained using the same model with several other semi-distances: the Euclidean distance, the PCA-based distance, a derivative-based semi-distance, a semi-distance that realigns before computing an Euclidean distance (see *Ferraty and Vieu, 2007* or *Timmermans et al, 2011* for definitions). In each case, the number of misclassification observed on the remaining 43 spectra is recorded. This test for the prediction of the AMD health status from the spectra is repeated 80 times, with different randomly selected training sets. Results are summarized in Table 1, for BAGIDIS and its best competitor, being a PCA-based semi-distance with at least 6 components. We observe that the non-optimized BAGIDIS obtains *no error* 10% more often than the PCA-based semi-distance. Furthermore, we can optimize the weights and the λ parameter of the BAGIDIS semi-distance using a cross-validation procedure within the training set, and the resulting model is tested on the remaining 43 series. This test is repeated 18 times on different randomly selected training sets, and no prediction error occurs. At each repetition, only 1 non-zero weight is selected. We observe no prediction error in every case, indicating a risk of misclassification that is probably smaller than 0.05. This indicates a very good capacity of discriminating the serum spectra from AMD and healthy patients.

	Occurrences of 0 error out of 43 predictions	Occurrences of 1 error out of 43 predictions
PCA-based semi-distance with $q \geq 6$	40 times out of 80 **50%**	40 times out of 80 **50%**
Non-optimized BAGIDIS semi-distance with prior weights and $\lambda = 0.5$	48 times out of 80 **60%**	32 times out of 80 **40%**
Optimized BAGIDIS semi-distance (1 non zero weight is selected)	18 times out of 18 **100%**	0 times out of 18 **0%**

Table 1: Summary results for the prediction of the AMD health status from the spectra.

4.3 Search for biomarkers

As a last step of the analysis, we aim at identifying AMD biomarkers in the spectra. Six advanced statistical methods for the discovery of of metabolomics spectral biomarkers from ^{1}H NMR spectra have been identified in *Rousseau et al, 2008*: multiple hypothesis testing (MHT), supervised principal component analysis (s-PCA), supervised independent component analysis (s-ICA), discriminant partial least squares (PLS-DA), linear logistic regression (LLR) and classification and regression tree (CART). A description of those methods can be found in *Rousseau et al, 2008*, as well as an assessment of their relative performances: recommendation is given to use s-PCA with caution due to its low general efficiency; use of CART is discouraged due to its noise sensitivity; the other four methods are diagnosed promising. All those methods have been applied to the AMD database (with outliers excluded) by *Rousseau, 2011*. For each method, the 20 most significant biomarkers have been identified. Results are presented in Figure 2. Here, we compare those results to the ones we obtain using *double geometrical t-tests* (*Timmermans and von Sachs, 2010*) based upon the BAGIDIS semi-distance.

The idea of *double geometrical t-tests* is as follows. We first restrict the dataset to sliding segments of the data located in a given range of ppm values, and compute the BAGIDIS distance matrix between those segments. Then, we test for the equality of the means of the distances between two AMD segments (*intra-group distances*) and the distances between one AMD and one non-AMD segment (*inter-group distances*). We also test for the equality of the means of the distances between non-AMD segments (*intra-group distances*) and between one AMD and one non-AMD segment (*inter-group distances*). In both case, the alternative is that intra-group distances are lower than inter-group distances. These tests are performed using Welch t-tests, assuming independence and normal distribution of the distances about their group mean. Combining the results of both tests allows to deduce the relative positions of AMD segments and non-AMD segments. Only if both t-tests significantly reject their null hypotheses are the two groups statistically different in mean. In this way, we detect if the selected sliding segment is significantly discriminant with respect to the AMD health status, by requiring a significance $\alpha = 1e - 10$ for both t-tests. This test is actually performed for each sliding segment of the dataset. A Bonferroni correction is thus applied to each p-value to account for the 576 simultaneous comparisons.

Detected differences are differences in the windowed spectral segments, and not at a specific spectral location. This is a difference with competitor methods. A given spectral location contributes therefore to a number of segments equals to Dt. For each spectral location, the number of significantly discriminant segment at which it contributes is computed and reported in Figure 3 for different parametrization of the BAGIDIS semi-distance. This number of significances

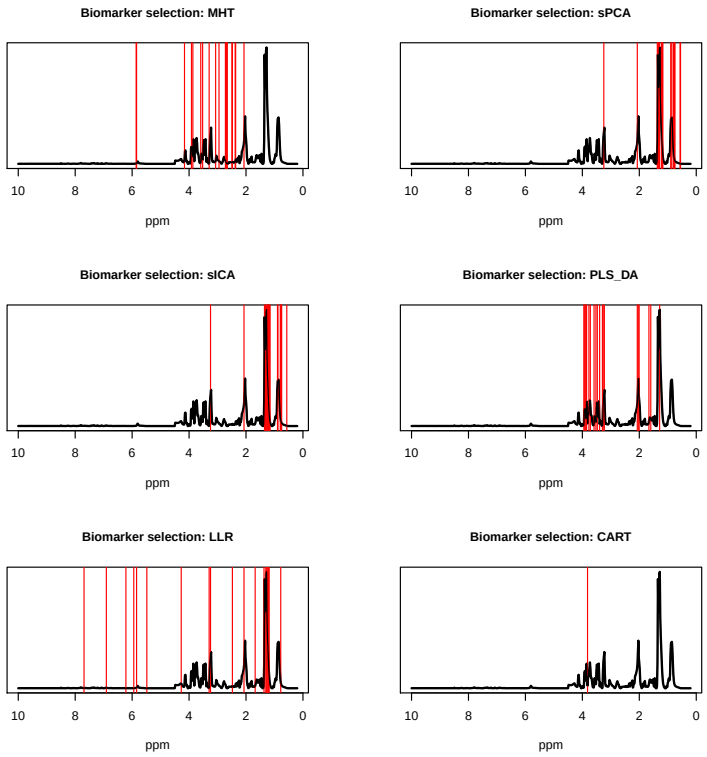


Figure 2: Search for AMD biomarkers using MHT, s-PCA, s-ICA, PLS-DA, LLR and classification and CART. For each method the 20 most significant biomarkers identified by *Rousseau, 2011* are marked by a vertical line. A typical spectrum of the AMD database is superimposed to ease the interpretation of biomarker detection.

can be used to search for biomarkers. The higher the number of significances, the higher the indication that the related spectral location might be a biomarker. A number of significances equals to Dt ($Dt = 25$ here) for a given spectral location indicates that each segment where this location contributes is significant with respect to the AMD health status. This clearly indicates for a biomarker.

Spectral zones from 3.99 to 3.06 ppm, as well as 2.48 and 2.27 ppm are strongly identified as biomarkers in Figure 3 (*top, left*). Those spectral zones are also detected by MHT and PLS-DA. A contribution in this zone, located at 3.24 is also detected by LLR, s-ICA and s-PCA, while CART has its only detection at 3.82. A highly significant detection of Bagidis also takes place in the spectral zone 7.24 to 6.89 ppm, which also contains a significant detection of LLR at 6.90 ppm. Some detection, although slightly less significant, is also found in the spectral zone 0.64 to 0.39, which is also detected by s-PCA, s-ICA, and, at one single location, by LLR. A detection at 2.07 also occur, which is also identified by all competitor methods except for CART. Possible, less significant, biomarkers are pointed out around 8.47 (no detection by other methods), 5.75 (also with LLR and MHT), 4.27 (also with LLR), 4.13 (also with MHT), 1.67 (also with LLR), 1.59 (also with PLS-DA), 1.36 (also with LLR and s-ICA), 1.22 (also with

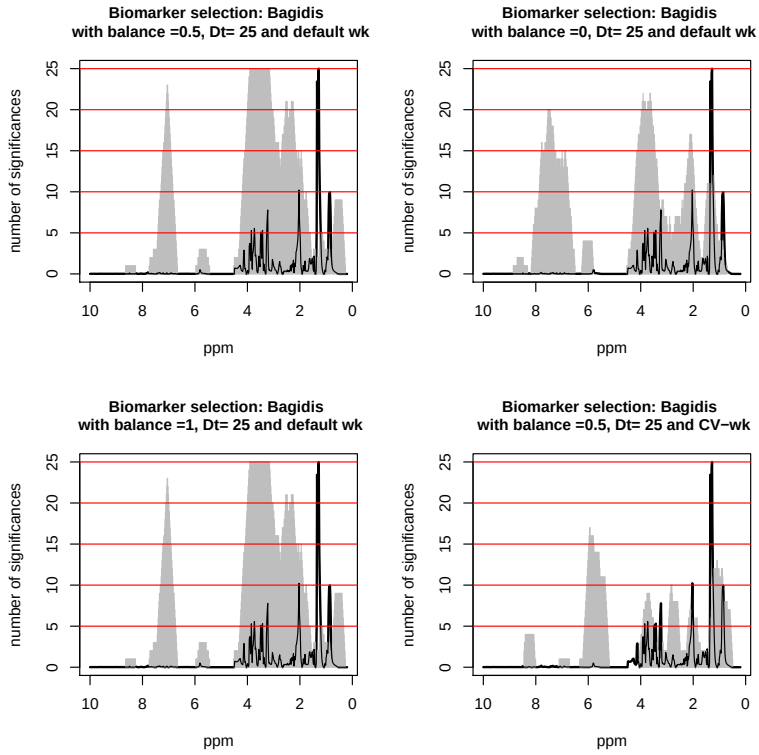


Figure 3: Search for AMD biomarkers using Bagidis with different parametrizations. For each spectral location, the number of detection of a spectral segment significantly discriminative for AMD which covers this location is provided. A typical spectrum of the AMD database is superimposed to ease the interpretation of biomarker detection.

s-PCA, s-ICA and LLR) and 0.90 ppm (also with s-ICA and s-PCA). Setting the λ balance parameter to 1 (breakpoints only) in the Bagidis semi-distance does not significantly modify these results. This is in accordance with the visual analysis in Subsection 4.1. Setting $\lambda = 0$ (amplitudes only) suppresses the detection of the spectral zone at ppms lower than 0.64, while the detection of the spectral zone at 1.22 ppm becomes more clear. The relative contributions of the spectral peaks around 5.75 and 2.07 ppm increase. This helps gaining an insight in the way the detected spectral zones do differ in AMD and non-AMD spectra. Finally, Figure 3 (*bottom, right*) identify spectral zones of significant differences between AMD and non-AMD spectra when w_k is set to its cross-validated value, as obtained in Subsection 4.2. This allows for finding a discriminant, sufficient but not exhaustive, set of biomarkers for AMD, as those biomarkers are the only one which contributes to the AMD detection model calibrated in Subsection 4.2.

We summarize this analysis by observing that Bagidis detects in one single study nearly all the spectral zones which were detected as biomarkers by at least one of the competitor methods. This emphasizes its consistency and its large scope of detection, which might be valuable for reducing the number of different statistical tools needed in a metabolomics study. Very few detection occur that are not detected by at least one other method, which might be an indication

that the method does not increase false detection rate relative to the combined use of competitor methods. From a biological point of view, this study has identified some biochemical pathways that could be implied in the anthology of AMD. Some additional biological experiments are required in order to validate these results.

5 Conclusion

This metabolomics study of the AMD dataset using the Bagidis methodology has been shown to be a useful complement to recent statistical analyses in the same field (*Rousseau, 2011*). It provides more informative visual discrimination of AMD and non-AMD blood samples which does not highlight outliers with respect to the semi-distance used. It allows for building a detection model for the AMD health status whose performances are shown to be really good. Finally, it allows for detecting - in a single analysis - a large set of biomarkers, this detection otherwise requiring the combination of six advanced statistical methods for biomarker search.

This analysis was performed using the *R software for statistical computing* and the library *Bagidis* which will be publicly available soon.

Bibliography

Ferraty F. & Vieu P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer.

Fryzlewicz P. (2007). Unbalanced Haar Technique for Non Parametric Function Estimation. Journal of the American Statistical Association, 102, pp. 1318-1327.

Lindon J.C., Nicholson J.K. & Holmes E. (2007). The Handbook of Metabonomics and Metabolomics. Elsevier.

Nicholson, J.K. & Lindon J.C. (2008). System biology: Metabonomics. Q & A. Nature, 455, pp. 1054-1056.

Noël A., Jost M., Lambert V., Lecomte J. & Rakic J.-M. (2007) Anti-angiogenesis therapy of exudative age-related macular degeneration: current progress and emerging concepts. Trends in Molecular Medicine, 13, 345-352.

Rousseau R., Govaerts B., Verleysen M. & Boulanger B. (2008), Comparison of some chemometric tools for metabonomics biomarkers identification. Chemometric and intelligent laboratory systems, 91, pp. 54-66.

Rousseau R. (2011). Statistical contribution to the analysis of metabonomics data in ^{1}H NMR spectroscopy. PhD thesis. Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium.

Timmermans C. & von Sachs R. (2010). Bagidis, a new method for statistical analysis of differences between curves with sharp discontinuities. Submitted.
URL: `http://www.stat.ucl.ac.be/ISpub/dp/2010/DP1030.pdf`

Timmermans C., Delsol L. & von Sachs R. (2011). Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features. Submitted.
URL: `http://dial.academielouvain.be/handle/boreal:81200`

Méthodes de régression avec "*p* grand" de prédire composants chimiques

Regression methods with "*p* large" to predict chemical components

Carmen Ybarra-Moncada[1], Gricelda Vázquez-Carrillo[2], Aldo Rosales-Nolasco, Nancy Toriz-Robles, Antonio Carbajal-Linares and Oswaldo Rubio-Covarrubias

[1] *Universidad Autónoma Chapingo, km 38.5 carr. México - Texcoco. CP 56230, Chapingo, México*
ycydrive@gmail.com

[2] *Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP). Km 13.5 Carretera Los Reyes-Texcoco Coatlinchán, Edo. De México. CP 56250.*
gricelda_vazquez@yahoo.com

Abstract

In study we illustrate the use of LS-SVM after initial principal components analysis (PCA) and partial least squares analysis (PLSA) as a data pretreatment on a chemometric data set. The case study is based on NIR measurements of samples of potato flour. The application of these regression procedures is on the evaluation of soluble sugars (SS), starch and protein by means of NIR spectroscopic data of potato flour made from potato clones destined for industrial use. The performance of LS-SVM was not convincing as the method performed poorly. However, after initial PCA and PLSA as pretreatment methods, the MSPE values for LS-SVM were about the same as for PLSR models. PLSR prediction models come out best in terms of MSPE in the responses predicted (SS, starch and protein).
Keywords : potato, prediction, PCA, PLSR , LS-SVM

1. Introduction

This study was motivated by chemometric modeling problems in which near infra red (NIR) spectroscopic data are used to estimate quality attributes of biological products. Chemometrics provides methodologies to produce not only rapid and accurate predictions but also non-invasive and non-destructive quality assessments. This technology represents a huge advantage for Mexican farmers in order to participate in a fair trade on the basis of the real attributes of their potato harvest.
Standard predictive chemometric modeling approaches include principal components regression (PCR) and partial least squares regression (PLSR). More recently least squares support vector machines (LS-SVM) have been applied by scientists in a number of disciplines to regression problems with predictors larger than samples ($p>n$). In order to implement LS-SVM it is necessary to choose the values of its two hyper-parameters. Poor choices may lead to serious under-fitting or over-fitting.

It is desirable to understand the statistical properties of new methods, such as LS-SVM, if they are to be used routinely in applications. Although much research has been devoted to LS-SVM as a regression and classification computing tool, little information is available on either its statistical

properties or its application in NIR spectroscopy to biological products. Owing to the availability in a short time of hundreds of spectral predictors, the problems to solve are essentially: variable selection, collinearity, non-linearity, outliers, over-fitting and under-fitting. In order to deal with some of these issues LS-SVM (with and without data pretreatment), PCR and PLSR models were built. The models' performances were assessed with mean squared prediction error (MSPE) on a test set. The objectives of the study was to apply the methodology for the analysis of chemometric data, where the many very highly correlated explanatory variables constitute major problems. Thus PCR, and PLSR, the most common multivariate calibration methods in chemometrics, are considered here. Likewise LS-SVM is also considered. These techniques are applied to real biological data with the purpose of predicting quality attributes of potato clones.

2. Materials and methods

2.1 Data set

The information is based on NIR measurements of potato flour samples made from 93 clones of potato. The spectra for each are quantifed in absorbance over the range 400 to 2500 nanometers (nm) with 4 nm resolution, giving 525 highly correlated explanatory variables xj ; j = 1, 2, …, 525; as illustrated in Figure 1.

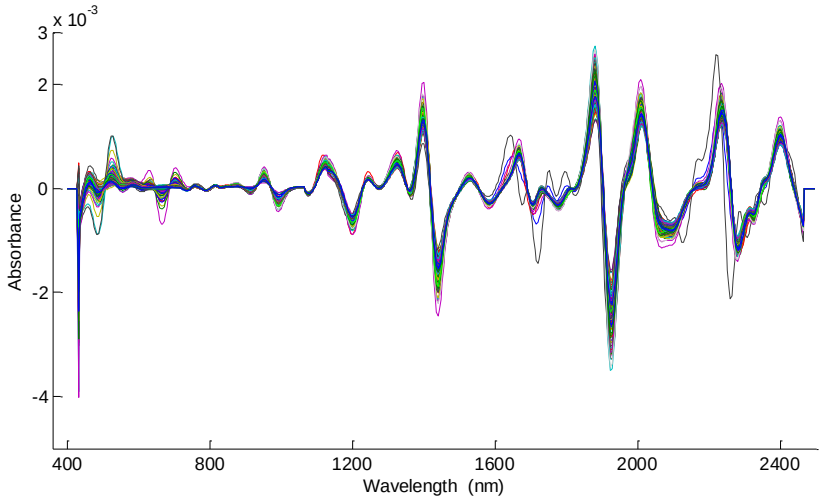


Figure 1 : NIR spectra of 220 potato flour samples from 400 to 2500 nm.

Here we combine LS-SVM with PCA and PLSA, to see whether LS-SVM has a possible role in augmenting standard prediction methods. The approach used is as follows: It involves splitting the data into three subsets. Training set 1 is used to select and estimate appropriate PCs. These PCs, after standardising to have equal variability, are used as the new X-variables in subsequent analyses. Training set 2 with the revised X-variables is then used to tune the hyperparameters and estimate the parameters of the LS-SVM. Finally, the testing set is used to test the fitted model in terms of MSPE. The chosen data partition kept the same 25% testing set for all the analyses. This allows a direct comparison between the results for LS-SVM, PCR, PLSR and LS-SVM_combined for the case study.

2.2 Regression models

The collinearity problem is addressed by regularized methods, which trade bias for variance, reducing the least squares estimates in weight or value. PCR and PLSR are useful regularized methods when the explanatory variables are large in number and highly correlated, as with spectroscopy data.

Principal Components. As Johnson and Wichern (2002) indicate, the general aims of PCA are data reduction and interpretation by explaining the variance-covariance structure of explanatory variables, through replacing them with uncorrelated linear combinations (k_j) made from the original explanatory variables. Once the selection of PCs has been concluded, the next stage is to use the scores matrix K containing exclusively a subset of k_js to carry out a multiple linear regression by regressing $\dot{y}$ on the first ζ PCs, $k_1,\ldots, k_\zeta$. The regression model can be expressed as:

$$\dot{y} = \mathrm{K}\beta_{PC} + \varepsilon$$

where β_{PC} denotes the regression coeficients whose estimator is calculated by normal least squares.

Partial least squares. Like PCA, the PLS approach chooses ζ PLS components k_j , where $\zeta \leq \cdot \min(p; n-1)$. It is expected that the ζ PLS components selected should be highly correlated with **y.** By contrast with PCA, the PLS approach sequentially searches for linear combinations by performing a simultaneous decomposition of X and y with the constraint that these linear combinations (new explanatory variables) explain as much as possible of the covariance between X and y (Frank, 1987). The second stage is properly the PLSR, where $\dot{y}$ is regressed on the ζ PLS components, which is fited by ordinary least squares and it can be expresed as:

$$\dot{y} = \mathrm{K}q + \varepsilon$$

where K is the scores matrix and ***q*** vector corresponds to loadings.

Least squares support vector machine. This method can be thougth of as a penalised regression method and needs to be trained. It has two hyperparameters γ and σ that must be carefully choosen. According to Suykens *et al*. (2002) LS-SVM is an element of a new generation of machine learning algorithms that are closely associated to regularized networks and kernel procedures.

According to Suykens and Vandewalle (1999) and Roobaert (2002) the basic idea of LS-SVM is twofold: initially to provide a non-linear function approximation by transforming the X-matrix into a high dimensional X-space by using a specific function to transform the original data X by associating elements of X with one or more elements of a new space. In this new and high dimensional X-space the regression model is fitted.

The following penalised regression problem must be solved, given γ and σ:

$$\min J(\beta, e) = \frac{1}{2}\beta^T\beta + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2$$

Subject to the constraints: $y_i = \varphi(x_i)\beta + \beta_0 + e_i$, $i = 1,\ldots,n.$

Then a function kernel is defined as:

$$K = (k_{ij}) y_i \quad \text{with} \quad k_{ij} = \varphi(x_i)^T \varphi(x_j)$$

and

$$k_{ij} = \exp\left(\frac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2} \right)$$

Therefore it must be solved for the intercept term and the Lagrange multipliers α:

$$\begin{bmatrix} 0 \\ y \end{bmatrix} = \begin{bmatrix} 0 & 1^T \\ 1 & K + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \alpha \end{bmatrix}$$

The regression model can be expressed as:

$$\dot{y} = \beta_0 + \sum_{i=1}^{n} k_{ij} \alpha_i .$$

3. Results and discussion

So disappointing was the performance of LS-SVM (Table 1) to model the case study data.
As shown by Table 1 the MSPE values of LS-SVM method are often larger than those of PCR and PLSR for the constituents of potato flour, in the main because of difficulty in tuning the hyperparameters.

After an initial reduction in the number of explanatory variables by PC and PLS, the results suggest that the LS-SVM$_{PC}$ and LS-SVM$_{PLS}$ approaches really improve the performance in prediction of LS-SVM method. In all three constituents LS-SVM$_{PLS}$ does better than LS-SVM$_{PC}$.

Method	Constituent of potato flour		
	MSPE Soluble sugar (p = 314)	MSPE Starch (p= 425)	MSPE Protein (p = 388)
PCR	1.63	I26.41	1.92
PLSR	0.68	28.03	0.98
LS-SVM	1.65	28.41	0.51
LS-SVM$_{PC}$	0.86	27.69	1.52
LS-SVM$_{PLS}$	0.65	27.62	0.62

Table 1: Mean squared prediction error (MSPE) values of PCR, PLSR, LSSVM, LS-SVM$_{PC}$ and LS-SVM$_{PLS}$ prediction methods for soluble sugar, starch and protein in potato flour.

Figure 2 shows predicted against measured response of starch, sugar and protein of PCR, PLSR and LS-SVM methods. As shown, a poor performance of the prediction models to

predict starch and soluble sugar in potato flour was obtained, whereas for protein the models improved their performance.
The LS-SVM methods performed poorly in the case study. Large values of the tuned σ^2 hyperparameter were obtained by the tuning method here applied, which led to LS-SVM estimates essentially averaging the training data (over-smoothing/under-fitting). Consequently large MSPE values were obtained.for starch and soluble sugar.

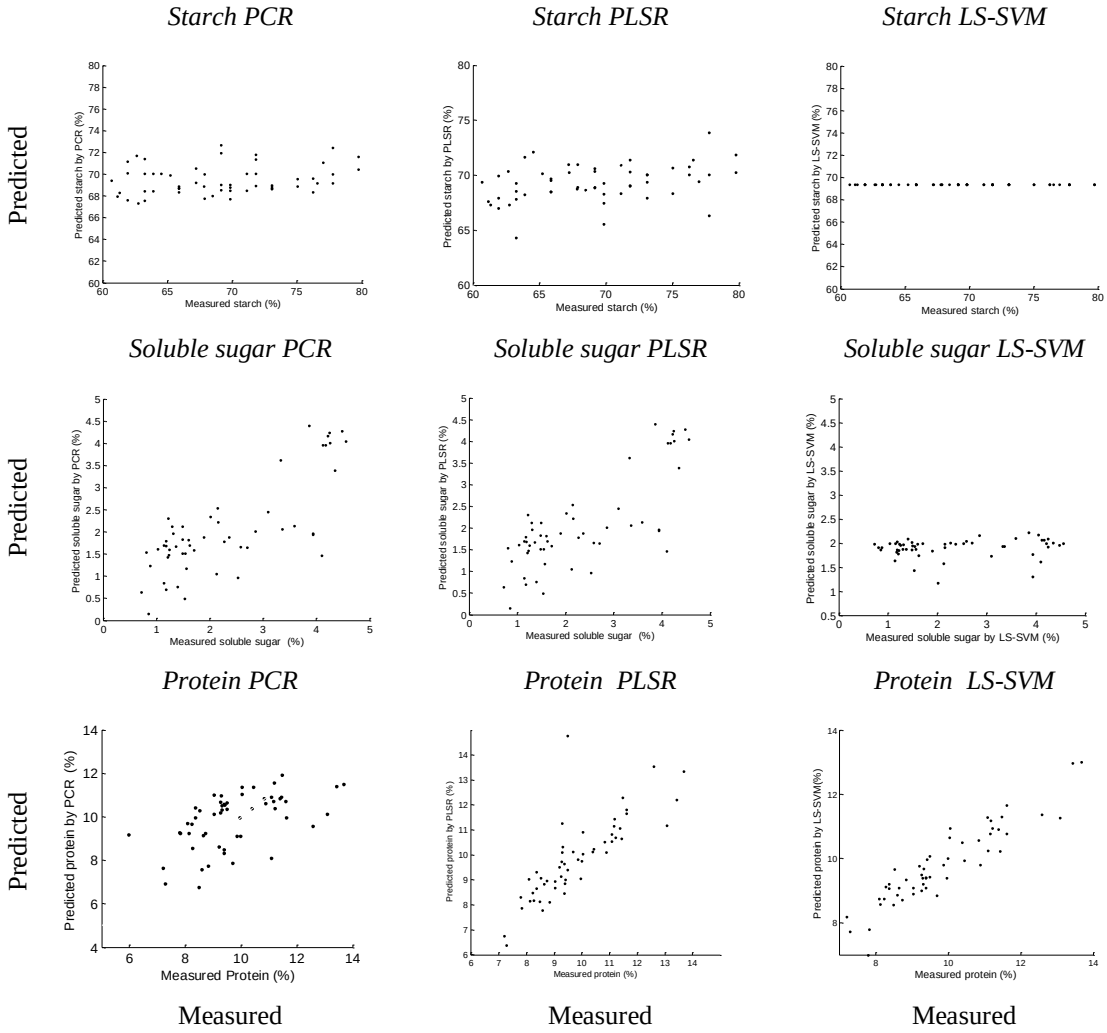


Figure 2. Predicted against measured response starch, soluble sugar and protein values for the PCR, PLSR and LS-SVM models.

Figure 3 shows predicted against measured response of starch, sugar and protein of LS-SVM_{PC} and LS-SVM_{PLS} methods. These results suggest that this approach really improves the performance in prediction of LS-SVM method.

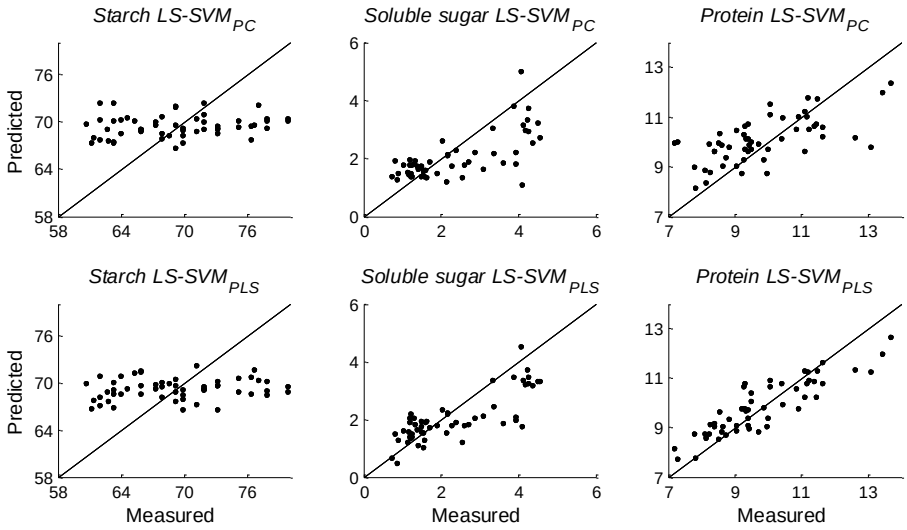


Figure 3. Predicted against measured response of starch, sugar and protein values for the LS-SVM$_{PC}$ and LS-SVM$_{PLS}$ models.

PLSR was generally found to be superior to the other regression methods in terms of MSPE. The tuning method to estimate the hyperparameters was not satisfactory. Consequently LS-SVM did poorly overall relative to the other methods. Finally a method that combined LS-SVM with PC and PLS did better than LS-SVM.

References

Frank, I. E. (1987). Intermediate least squares regression method, *Chemometrics and Intelligent Laboratory* 1, 233-242.

Johnson, R. A. & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis, Pearson Education.

Roobaert, D. (2002). DirectSVM: A simple support vector machine perceptron. The Journal of VLSI Signal Processing 32, 147-156.

Suykens, J., VanGestel, T., DeBrabanter, J., DeMoor, B. & Vandewalle, J. (2002). Least Squares Support Vector Machines, World Scientific.

Suykens, J. & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural Processing Letters. 9, 293-300.

Segmentation d'un panel de consommateurs et identification des qualités sensorielles importantes dans chaque segment: une étude de cas portant des pommes.

Consumer segmentation and elicitation of the key drivers: a case study with apples.

Morgane Blasi[1], Marion Brémond[1]& Mathilde Charles[2]

[1] *UNAM University, ONIRIS, USC "Sensometrics and Chemometrics Laboratory", Nantes, F-44322, France.*
E-mail : *morgane.blasi@oniris-nantes.fr*

[2] *Groupe ESA, Laboratoire GRAPPE, Angers*

Résumé

Aujourd'hui les industries portent de plus en plus d'intérêt aux préférences des consommateurs afin de mieux cibler leurs attentes face aux produits. Elles sont également sensibles à l'existence de groupes de préférences, en fonction de l'importance donnée aux qualités sensorielles. Elles utilisent pour cela une méthode bien connue : la cartographie externe des préférences, bien que d'autres alternatives existent. Une comparaison de différentes méthodes est réalisée, en prenant en compte différents critères de choix tels que la qualité des modèles obtenus, la facilité de mise en œuvre ainsi que la construction d'un graphique synthétique et parlant pour les industriels. Des méthodes sont sélectionnées et comparées à la cartographie externe en s'appuyant sur un jeu de données portant sur 31 variétés de pommes. Celles-ci ont été évaluées par 15 experts à l'ESA d'Angers, qui les ont notées en utilisant 15 descripteurs sensoriels. Parallèlement, l'appréciation globale des pommes a été évaluée par un panel de 224 consommateurs pour lesquels les « drivers » les plus importants des préférences sont identifiés.

Abstract

Nowadays, companies are more and more interested in understanding consumers' preferences towards products. They are also concerned by finding segments of consumers together with their preferences key drivers. For this purpose, they use the well known preference mapping method. Alternatively, other existing methods may prove to be efficient. A comparison of different methodologies is realized, by taking into account of different criteria such as the quality of the fitted models, the easiness of implementation and interpretation of the results. Methods are chosen and compared to prefmap by using a data set on 31 varieties of apples. These apples were assessed by 15 experts at ESA of Angers, on the basis of using 15 attributes. At the same time, a hedonic evaluation by 224 consumers allowed us to identify the most important drivers of preferences.

Investigation des relations entre une épreuve de citation libre et une épreuve de tri libre.

Investigating the relationships between a free listing task and a free sorting task.

Thibault Charpentier, Nicolas Jobard, Gaël Jouanlanne

UNAM University, ONIRIS, USC "Sensometrics and Chemometrics Laboratory", Nantes, F-44322, France.
E-mail *: thibault.charpentier@oniris-nantes.fr*

Résumé

Cerner la perception d'une gamme de produits par des sujets est un enjeu essentiel pour le développement de produits. Le tri libre et la citation libre sont deux méthodes d'évaluation des perceptions qui ont été étudiées indépendamment. Pour le tri libre, on demande à des sujets de regrouper des produits en fonction de leurs ressemblances et leurs différences et de caractériser chaque groupe par des verbatims. En ce qui concerne la citation libre, les sujets sont amenés à lister le maximum d'éléments par rapport à un sujet donné. L'objectif de la présente étude est d'explorer les liens entre ces deux procédures d'évaluation. Pour cela, 81 consommateurs ont participé à ces deux épreuves. Dans un premier temps, il leur était demandé de citer un maximum de fruits. Dans un deuxième temps, une liste de 20 fruits leur était présentée afin qu'ils les classent en différents groupes qu'ils ont par la suite décrits. L'étude de cette base de données permettra de mettre en évidence les recoupements des informations et les complémentarités entre ces deux méthodes.

Abstract

The assessment of the perception of a set of products by a panel of subjects is of paramount importance in product development. Free sorting task and free citation task are two procedures which have been advocated as tools to assess product perception. In free sorting task, the subjects are instructed to sort the products in groups, considering that products in the same group are perceived as similar. Moreover, the subjects are asked to describe each group using their own words. In free citation, the subjects are asked to cite all they know about a given domain or product. The aim of the present study is to investigate the relationships between the two procedures of assessment. For this purpose, an experiment was conducted. A panel of 81 subjects was asked to sort as many fruit as possible. Thereafter, they were asked to sort a list of 20 fruits. The data from each procedure are analyzed separately and in common.

Classification de variables qualitatives autour de composantes latentes. Application à des données d'une enquête consommateurs.

Clustering of categorical variables around latent components. Application to consumer survey data.

Justine Chevalier, Marlène Herbreteau & Agnès Tariel

UNAM University, ONIRIS, USC "Sensometrics and Chemometrics Laboratory", Nantes, F-44322, France.
E-mail : *justine.chevalier@oniris-nantes.fr*

Résumé
La classification de variables vise à former des groupes de variables homogènes permettant ainsi de réduire l'information initiale. Contrairement au cas quantitatif, relativement peu de travaux ont porté sur la classification de variables qualitatives alors que cette problématique s'avère pertinente dans le cadre des études consommateurs, faisant généralement appel à un grand nombre de variables qualitatives sous la forme de questions fermées. Parmi les approches proposées, certaines d'entre elles (Abdallah & Saporta 1998, Qannari & al. 1998, Derquenne 2006) prennent en compte différents critères d'association entre variables qualitatives alors que celle adoptée par Saracco & al. (2010) vise à déterminer simultanément les groupes de variables et les variables latentes associées, par le choix de la maximisation de leur rapport de corrélation. De manière analogue à cette approche ou encore à celle de Vigneau & al. (2006) dans le cas quantitatif, nous cherchons à déterminer des composantes latentes résumant au mieux des groupes de variables qualitatives. Pour se faire, différents critères d'association sont étudiés tels que le critère du Rand, le critère de Tschuprow ou encore celui de Janson Vegelius sur la base des matrices relationnelles associées aux variables. Par la suite, nous proposons de déterminer les composantes latentes traduisant l'homogénéité du groupe de variables à partir des critères étudiés. L'approche est finalement appliquée à un jeu de données issu d'une enquête internet relative au comportement alimentaire de consommateurs français.

Abstract
The aim of this paper is to study the clustering of categorical variables around latent components. Unlike the case where variables are quantitative ones, little attention has been paid on the clustering of qualitative variables. Among the different approaches proposed, we can cite several ones based on the use of association criteria while a latter method (Saracco & al. 2010) aims at determining both groups and latent variables on the basis of the criterion of the correlation ratio. Several association criteria are discussed in the first part such as Rand, Tschuprow and Janson Vegelius criteria to measure the proximity of two categorical variables. In a second part, an approach similar to CLV one (Vigneau & al. 2006) is adopted, where the categorical variables are grouped together around latent components, on the basis of the previous association criteria. This approach is applied in order to analyze a consumer survey about food behavior.

Distance entre partitions. Application à la classification de sujets à l'issue d'une épreuve de tri libre.

Distance between partitions. Application to the clustering of subjects after a free sorting task.

Marie Lafontaine, Adèle Seibert & Bérangère Shrum

UNAM University, ONIRIS, USC "Sensometrics and Chemometrics Laboratory", Nantes, F-44322, France.
E-mail : *marie.lafontaine@oniris-nantes.fr*

Résumé

L'épreuve de tri libre est une technique d'évaluation de la perception de produits de plus en plus fréquemment employée. L'analyse des données recueillies (partitions individuelles) est le plus souvent réalisée sur la base de la matrice de dissimilarités globales du panel, supposant l'homogénéité de la perception par les sujets. Ce travail vise à étudier les distances entre sujets après une épreuve de tri. L'objectif est de construire une méthode de classification des perceptions individuelles ainsi que de fournir des outils permettant de tester les différences de perceptions entre groupes de sujets.
Les techniques présentées seront appliquées à une étude sur la catégorisation de marques de voitures par un panel de 80 sujets. On étudiera en particulier l'influence de l'âge et du genre des sujets sur leurs catégorisations.
Mots-clés : Catégorisation, tri libre, partition, analyse sensorielle.

Abstract

The free sorting task is a popular technique for the evaluation of the perception of products. The analysis of resulting data (individual partitions) is commonly based on the aggregated matrix of dissimilarities, assuming the homogeneity of the panel. This communication presents the analysis of the distances between subjects after a free sorting task. The aim is to develop a technique for clustering the partitions given by the subjects and to propose tools for testing differences of perception between groups of subjects.
These techniques will be applied to a categorization of car brands by 80 subjects. The influence of age and gender onto the perception will be tested.
Keywords : Categorization, free sorting, partition, sensory analysis.

Prétraitement automatisé de la ligne de base en spectroscopie Raman pour l'analyse chimiométrique.

Automated baseline preprocessing in Raman spectroscopy for chemiometric analysis.

Philippe Sistat[1], Lidwine Grosmaire[2], Stefano Deabate[1] & Patrice Huguet[1]

[1] *Institut Européen des Membranes, UMR5635 ENSCM/UM2/CNRS, Université de Montpellier 2, CC047, Place Eugéne Bataillon, 34095 Montpellier cédex 5.*
E-mail : *philippe.sistat@iemm.univ-montp2.fr*

[2] *UMR Qualisud, Laboratoire Physique Moléculaire et Structurale, UFR des Sciences Pharmaceutiques et Biologiques,15 avenue C. Flahault BP 14491, 34093 Montpellier Cedex 5*
E-mail : *lidwine.grosmaire@univ-montp1.fr*

Résumé
Un algorithme permettant une détermination automatique de la ligne de base en spectroscopie Raman est présenté ici. L'algorithme s'attache principalement à rechercher des points d'intérêts à travers lesquels faire passer une ligne de base optimale.
Mots-clés : ligne de base, spectroscopie Raman, fluorescence.

Abstract
An algorithm for automated baseline subtraction is presented in this work. The algorithm is mainly devoted to the search of point of interest which the baseline must pass through.
Keywords : Baseline, Raman spectroscopy, fluorescence background

1 Introduction

Experimenter often needs to remove baseline in Raman spectra. When confronted to huge dataset it becomes necessary to use an automated script to do this job. When the datasets all have the same origin, it can be useful to find a way to choose the baseline point for fitting with a great repetability. For this to be achieved, the information or parameters given to the program must be global for the whole procedure. The algorithm described in this work has been developed in the context of preprocessing Raman spectra for visualizing concentration profile of molecular ions and solvent molecule within ion-exchange membranes during the processing of food industry solutions. Indeed, ionic polymer are known to generate a high fluorescence background. Baseline subtraction of Raman spectra may be carried out by hand or automatically [Liland 2010; Peng 2011; Zhao 2007; Rowlands 2010]. The proposed method relies essentially on geometric consideration. It combines both basic ideas from rolling ball algorithm [Kneen 1996; Liland 2011] and curve fitting method [Dood 2002]. The algorithm seems to be robust and have been tested on different media (aqueous solutions, ion-exchange polymer in the presence of salt and organic molecule) before applying any further multivariate analysis in order to obtain the concentration profiles.

2 The algorithm

Let's consider one spectrum as a given set of points (ν_k, I_k) where k is the index varying from 1 to the total number of points reserved for the whole spectrum $n_{\tilde{\nu}}$.

The basic idea is to plot the spectrum in the (x, y) plane after the following transformations being defined:

- $x = f(\tilde{\nu})$;
- $y = g(I)$

where f and g are increasing monotonic functions. This allows to define in a univocal way the reciprocal functions f^{-1} et g^{-1}, and in the same time to preserve the point order in the spectrum. Thus, we obtain a finite set of points (x_k, y_k) sharing the same index k previously defined.

In this plane, let's imagine a wheel with a radius R that will try to roll under the curve defined by the transformed spectrum. The path followed by the center of the R radius wheel define a curve situated at a vertical distance from the spectrum (x_k, y_k) at least equal to R. The notion is quite similar to the trochoide one. If we translate vertically this curve by $+R$, we define a curve always located under the spectrum at a vertical distance at least equal to 0. This curve could roughly define a baseline.

The encoutered peaks must be positives for the method to works. Hence, we must eliminate the occuring negative artefacts on spectra. For this to be achieved, a sliding median filter behaves very well. Linear filters such as sliding average or gaussian filter are less interesting because they can easily be biased by outliers.

Spectra are usually provided under the form of a set of n_s spectra and stored in a $(n_s \times n_{\tilde{\nu}})$ matrix. Indeed, in a set of spectra corresponding to the same experiment the wavenumbers are taken on the same sample grid indexed by k with $k = 1 \dots n_{\tilde{\nu}}$. The wavenumber values are refered as $\tilde{\nu}_k$.

In order to perform the baseline shift we must be in the (x, y) plane, for each x_k abcissa belonging to a wavenumber of the spectrum we carry out the following test:

- For a given index i such as $x_k - R < x_i < x_k + R$, we seek $\overline{y}_k$ such as

$$\overline{y}_k = \min_i (y_i - \sqrt{R^2 - (x_i - x_k)^2}) \tag{1}$$

  The circle with radius R and center $(x_k, \overline{y}_k)$ touch the spectrum at point coordinate (x_j, y_j) where j is the value of index i for which the expression $(y_k - \sqrt{R^2 - (x_i - x_k)^2})$ is minimal.

  This contact point is often the same for differents k indexes. It can be seen as an obstacle over which stumble our R radius circle while rolling. It's its instantaneous rotation center in a kinematic analogy.

The search for the lowest circle with a center at the abcissa x_k in contact with the curve is depicted on the figure 1.

The square of the abcissa differences can be computed once and stored in memory in a symetric matrix $(a_{i,j} = (x_i - x_k)^2)$ before running the algorithm. Only half of the matrix need to be filled owing to the symmetry property.

Two important kind of points appears:

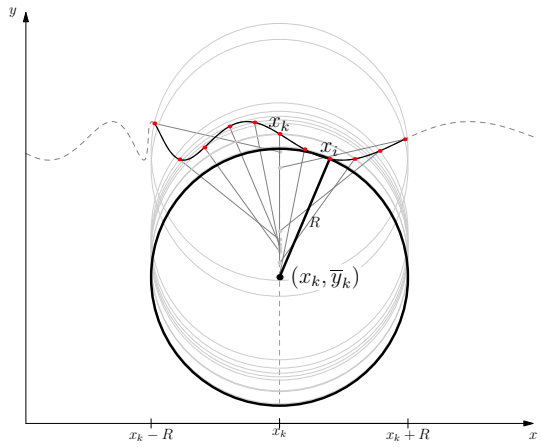


Figure 1: Contact circle.

- the points (x_k, y_k) that are rest points for several circles with successive indexes. By definition, when several circles share the same contact point, their centers are all located at the same distance R from this contact point and it follows that the curve described by the corresponding points (roughly the baseline) is an arc of a circle. Hence, the first approximation for the baseline build in this way is a serie of circle arc with radius R. A list of the successive point belonging to the spectrum and being the successive contact point can be build. The set point is noted $\{A_r\}$ with $0 < r \leq n_r \leq n_{\tilde{\nu}}$. The given list can be used as node for fitting a smooth baseline, like a classical polynomial curve for instance.

- the other kind of points is constituted by the junction nodes between the successive arc. The suite of points is then noted $\{B_s\}$ with $0 < s \leq n_s \leq n_{\tilde{\nu}}$.

Once the points $\{A_r\}$ being determined a smooth curve can be fitted through theses points or a spline can be forced to theses points. The obtained curve then defined the baseline. The other set of points $\{B_r\}$ can possibly help as initial estimates to automatically find the wavenumbers domain where peaks appears.

2.1 The parameters

The function f and g together with the fixed value of the radius R of the rolling ball are the main parameters for the algorithm. The function f acting on the wavenumbers allows a better distribution of the points on the abcissa. This is equivalent to use a ball with a varying radius. Thus we can modulate the interval between successive point and for instance reduce the width of a broad band like the water one in order not to be trapped in such a domain. The function f can be defined once for fitting the baseline of a given set of spectra sharing the same nature. Hence, the node points will all be computed in the same way.

3 Conclusion

The algorithm described here allowed us to preprocess large Raman datasets for ionic membrane content visualization. Since the parameters used to extract the baseline can be shared by all the spectra obtained the intervention of the experimenter is only done once for hundred of spectra. The method has been proven robust and permits the automated post-processing of data by any chemometric method.

4 Acknowledgements

This work was funded by the ANR-USAR grant PROMEMSEL N° 023617 "*Physico-chimie et transfert de solutés organiques à travers des membranes en présence de sel*".

Bibliographie

Kneen, M.A. & Annegarn, H.J. (1996). Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds. *Nuclear Instruments and Methods in Physics Research B*, 109/110, 209-213.

Liland, K.H. & Mevik B.-H. (2011). R package "baseline". R package, url:http://cran.r-project.org.

Liland, K.H., Almøy, T. & Mevik, B.-H., (2010). Optimal choice of baseline correction for multivariate calibration of spectra. *Applied Spectroscopy*, 64(9), 1007-1016.

Rowlands, C. & Elliott S. (2010). Automated algorithm for baseline subtraction in spectra. *Journal of Raman Spectroscopy*, 42, 363-369.

Peng, J., Peng, S., Xie, Q. & Wei, J. (2011). Baseline correction combined partial least squares algorithm and it application in on-line Fourier transform infrared quantitative analysis. *Analytica Chimica Acta*, 690, 162-168.

Zhao, J., Lui, H., McLean, D.I. & Zeng, H. (2007). Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Applied Spectroscopy*, 61(11) 1225-1232.

Dodd, J.G. & DeNoyer L.K. (2002). Handbook of vibrational spectroscopy, Vol. 3. *Sample characterization and spectral data processing* Ed. Chalmers, J.M. & Griffiths P.R., Chichester: John Wiley & Sons Ltd.

RMT RÉSEAU MIXTE TECHNOLOGIQUE
EXPERTISE POUR LA DÉTERMINATION DE LA DURÉE DE VIE MICROBIOLOGIQUE DES ALIMENTS

Taking into account varibility and uncertainty in models for assessing the microbiological shelf-life in foods

Application to Sym'Previus probabilistic module

EL Jabri M.*, Pinon A., Ellouze M., Stahl V., Denis C., Thuault D., Guillier L., Augustin J.C.
*Correspondant : cellule opérationnelle Sym'Previus, mohammed.eljabri@adria.tm.fr

Introduction

The integration of variability and uncertainty, in models for shelf-life evaluation of food products, represents the main element for the reliability of results.

The aim of this research is to quantify the impact of both variability and uncertainty in the microbiological shelf-life estimation of food products.

Different growth kinetics characterizing the same food were modeled. Non Linear Mixed Effect Models were used based on the Stochastic version of Expectation-Maximisation Algorithm (SAEM) to characterize variability and uncertainty associated to growth parameter estimation. These estimates were determined by Monolix software. They were then used for simulation, using 2D Monte Carlo method, in Sym'Previus probabilistic module to estimate with precision the probability to exceed a critical value at the end of food shelf-life.

Probabilistic approach

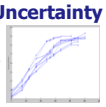
While zero-risk doens't exist, it became necessary to quantify the probability to exceed a criterion level at the end of food shelf-life, for example 100 CFU/g for *Listeria monocytogenes*.
The Probabilistic approach allows to estimate the probability to exceed this criterion with a confidence interval.

Sym'Previus, decision making tool

Microbiological and food variability ⇨ **Food-related behavior** ⇨ **Initial contamination** ⇨ **Storage conditions**

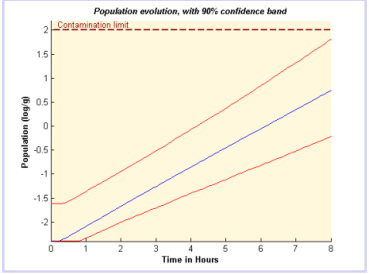
Probabilistic Simulation

Variability and Uncertainty

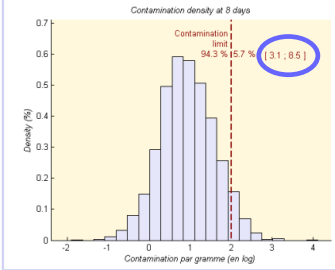
Growth parameters

μ_{max} : Growth rate
lag : Lag time
N_{max} : Maximal contamination

Initial contamination
n : Number of analysed samples
r : Number of positive samples
m: m-gram sample

Storage conditions and product charasteristics
Temperature, pH, water activity



Growth simulation in Contaminated Sales Units (CSU)



Probability to exceed a criterion at the end of shelf-life in CSU

Conclusion

Quantify the probability to exceed a criterion for different input data, taking into account both variability and uncertainty, allows to evaluate the impact of each input factors and contributes to identify the key factors on which efforts should be made to make the product safer (intial contamination, aw, pH...). Sym'Previus is a decision making tool that allows a reliable evaluation of contamination throughout the shelf-life of food.

RMT "expertise for determining the microbiological shelf-life of foods"

This work was carried out as part of RMT for the determination of microbiological shelf-life of foods in consultation with the Scientific and Technical Committee of Sym'Previus (STC). This project has received funding from ACTIA (Paris) to support the activities developed by RMT.

Scientist Interest Group (S.I.G) : Sym 'Previus

Center Network ACTIA : Actilait, ADRIA Développement / ADRIA NORMANDIE / Aérial / IFIP / Institut Pasteur de Lille
Public Laboratories : INAPG / INRA / ENVA / ANSES
Industrial membres of UNIR association : Bel / Bongrain / Danone / Pernod Ricard
Authorities : *Ministries of Agriculture and Research*