Système d'aide à la conduite des procédés viticoles et oenologiques, basé sur l'utilisation d'un capteur piéton de suivi de maturité

V. Geraudie

T H È S E

pour l'obtention du titre de

Docteur du Centre International d'Etudes Supérieures en Sciences Agronomiques de Montpellier

Ecole Doctorale : Sciences des Procédés Sciences des Aliments

Système d'aide à la conduite des procédés viticoles et œnologiques, basé sur l'utilisation d'un capteur piéton de suivi de maturité

présentée et soutenue publiquement le 2 décembre 2009 par

Vincent GERAUDIE

JURY

Président :	Pr. Alain CARBONNEAU,	Montpellier SupAgro.
Rapporteurs :	Pr. Jose Emilio GUERRERO, Pr. Gilbert GRENIER,	Univ. de Cordoue (Esp.) ENITA de Bordeaux.
Directeur de thèse :	Pr. Jean-Michel ROGER,	Cemagref Montpellier.
Examinateurs :	Dr. Sylvie MARCHESSEAU, Pr. Francis SEVILA,	Polytech'Montpellier. INRA Midi-Pyrénées.
Invités :	Remi NIERO, Dr. Bruno TISSEYRE,	PELLENC SA. Montpellier SupAgro.

Thèse cofinancée par l'ANRT et la société PELLENC SA. dans le cadre d'un contrat CIFRE.

A André,

Toi qui me faisait peur étant petit...
Je te voyais seulement plein d'étoiles.
Toi que j'apprécie tant étant plus grand...
Je te vois seulement comme mon grand-père.

Aux "sherpas" des laboratoires.

Un sommet de plus à gravir sur le chemin de la recherche. Ils travaillent dans l'ombre et d'arrache-pied des jours durant pour faire fonctionner les expérimentations des uns et des autres, avant de s'effacer, au dernier moment, et laisser leur commanditaire planter seul le drapeau en haut de la montagne.

Remerciements

Les pages de ce mémoire reflètent l'écho de nombreuses rencontres que j'ai eu la chance de faire durant ces trois années (et un ε) de thèse. Ces quelques lignes sont l'occasion pour moi de remercier toutes les personnes qui, de près ou de loin, ont suivi (et supporté) ma vie de thésard.

Je tiens tout d'abord à remercier profondément le Pr. Jean-Michel ROGER qui m'a encadré et guidé tout au long de mes travaux. Grâce à toi j'ai beaucoup appris, aussi bien sur le plan scientifique qu'humain. Je suis sûr que sans ces quelques frayeurs tu te serais ennuyé car tu es un homme de défi. La preuve en est, tu m'as pris comme thésard.

Je suis très sensible à l'honneur que m'ont fait le Pr. Jose Emilio GUERRERO et le Pr. Gilbert GRENIER en acceptant d'être rapporteurs de ce travail. Je remercie également les membres de mon jury ; Dr. Sylvie MARCHESSEAU, le Pr. Alain CARBONNEAU, le Pr. Francis SEVILA et M. Remi NIERO.

J'exprime toute ma reconnaissance aux membres à mon comité de thèse : Véronique BELLON-MAUREL, Hernan OJEDA et Bruno TISSEYRE. Vous m'avez toujours laissé votre porte ouverte. Nos discussions et l'intérêt que vous avez toujours porté à mes travaux m'ont poussé à approfondir mes axes de recherche et à me perfectionner. De même, je remercie Serge GUILLAUME pour m'avoir guidé quand j'étais dans le flou, tout comme Bernard PALLAGOS pour avoir été sans biais et robuste face de mes questions. Je suis également reconnaissant à Gérard LEROY pour la réalisation des nombreuses pièces nécessaires à mes expérimentations.

Je remercie de même la société Pellenc, ainsi que M. Jean-Marc GIALLIS et Jean-Louis FERRANDIS, pour m'avoir confié ces travaux et de me faire confiance pour la suite.

Les campagnes de mesures qui ont fourni les données utilisées dans ces travaux ont été menées dans différents centres. Toute ma reconnaissance va donc également à l'ensemble des personnes qui ont participé à l'obtention de ces données, spécialement aux équipes de l'INRA de Pech-Rouge, de l'IFV, de Moët & Chandon et du CIVC. Je tiens particulièrement à remercier l'ensemble de l'équipe de Pellenc Australia (Louise, tu peux me sortir une planche, I'm back !) et du domaine Bremerton pour l'accueil chaleureux qu'ils m'ont toujours réservé au cours de mes hivers au chaud... Euh... Au cours des campagnes de mesures hémisphère sud.

Encore merci à tous les Cemagrefiens et Pellenciens que j'ai côtoyé durant ces années.

Un grand merci plus personnel...

A Ma Truffe, que dire après trois années de co-bureau... De toute façon on se voit pour l'apéro du Jeudi et sans nuisance sonore à 20h20, c'est promis ! Allez, Force et honneur

Ma Truffe ! A Coach, tu n'es pas un ingénieur de l'espace mais j'espère que tu amèneras tes enfants sur la lune. A Alexia, mais où trouves-tu toutes ses idées photos et autres ? Sans oublier le reste de l'équipe Potager-Soirée-Jeux ; Brigitte, Hazael, Michaël, La Harpe et Elodie. Orianne, prends soin de l'hôtel au Sanglier. Elvira, notre soleil espagnol, sans toi je chercherai encore un rapporteur. A tous ceux des aquariums d'à côté, bonne chance à ceux qui démarrent et bon courage à ceux qui voient la fin !

Une pensée particulière pour mes compagnons de galère estudiantine, merci pour votre amitié. A JB et tes "c'est oui ou c'est non ! ? !". S'il y a une bonne gragedoullerie derrière, tu connais la réponse ! Mais surtout, si j'ai pu dire oui à Valérie, c'est grâce à toi. A La Fourmi, l'expert du venin, le tyran des TP et le prophète du 7A. Au Binôme, le roi des farceurs, le seigneur des clubs et le maître de la chauve-souris. Je peux bien prendre 5 min, non pas pour une 'tite clope, mais pour vous remercier tous les deux. Qu'auraient été ces années ISIM sans vous ? Sans oublier Chaussette, Toto, L'Ours, mes fillots...

A Renaud et Caro. De camps d'été en camps d'été, de week-end en week-end, des amitiés se sont tissées,... Vous connaissez la suite.

Merci à toute ma famille et tout spécialement à mes parents, pour la curiosité qu'ils m'ont transmise, pour toutes les chances qu'ils m'ont offertes ainsi que pour leurs encouragements et leur confiance sans faille.

Enfin, merci à ma Patate voyageuse. Le plus fabuleux des voyages est d'être avec toi.

N'est sutipde que la stupidité.
Un morbac n'est pas un cafard.

Table des matières

Liste des acronymes

ESM	**E**rreur **S**tandard **M**oyenne
SEC	Erreur standard d'étalonnage ou ***S**tandard **E**rror of **C**alibration*
SECV	Erreur standard de validation croisée ou ***S**tandard **E**rror of **C**ross **V**alidation*
SEP	Erreur standard de prédiction ou ***S**tandard **E**rror of **P**rediction*
SIF	**S**ystème d'**I**nférence **F**loue
Vis-NIR	Visible proche-infrarouge ou ***VIS**ible and **N**ear **I**nfra**R**ed*

Liste des symboles

$f(.)$ $\mathcal{M}(.)$	Fonction ou modèle théorique non-linéaire
θ_p	Vecteur des p paramètres du modèle
y_i	Variable : expliquée, dépendante, de réponse, d'observation...
x_i	Variable : indépendante, explicative, régresseur...
ε_i	Erreur aléatoire, représente la variance inexpliquée par le modèle
σ_ε^2	Variance des erreurs
$\widehat{y_i}$	Valeur prédite par le modèle
$\widehat{\theta_p}$	Valeur estimée des p paramètres du modèle
$\widehat{y_i^*}$	Valeur simulée à partir des observations prédites par le modèle
$\widehat{\theta_p^*}$	Valeur simulée à partir des p paramètres estimés du modèle
r_i	Résidu $(y_i - \widehat{y_i})$
$\mathbf{W}$	Lettres majuscules grasses employées pour désigner une matrice
R^2	Coefficient de détermination
$\mu_A(x)$	Degré d'appartenance au sous ensemble flou A
$\pi(.)$	Possibilité
$L(.)$	Fonction de vraisemblance
$p(.)$	Probabilité

Résumé - L'UMR - ITAP du Cemagref de Montpellier et la société PELLENC S.A. développent conjointement un système de mesure par spectroscopie proche infrarouge permettant de suivre, de manière non destructive, certains critères de maturité du raisin (brevet en cours de dépôt). Ces critères sont : les teneurs en sucre, en anthocyanes et en eau ainsi que l'acidité totale.

L'objectif de ces travaux est de prédire la courbe d'évolution "la plus probable" de ces critères de maturité, à partir des premières mesures recueillies grâce à ce capteur, tout en tenant compte des connaissances et hypothèses faites par le viticulteur. Grâce à ces prédictions, le système permet d'approcher la date des vendanges en fonction des hypothèses et des objectifs du viticulteur.

Chaque courbe est modélisée par une fonction explicite pourvue de paramètres spécifiques. Les fonctions retenues ont un nombre de paramètres intentionnellement réduit pour des raisons d'interprétabilité. La méthode développée dans le cadre de cette thèse a donc pour but d'estimer les paramètres de ces fonctions, à partir des jeux de mesures et des connaissances et hypothèses du viticulteur, ces deux sources d'informations étant entachées d'incertitudes.

Prises indépendamment, ces deux sources d'informations permettent d'obtenir la valeur la plus probable des paramètres ; les mesures en utilisant les techniques d'ajustement de modèle, les connaissances et hypothèses en utilisant celles des systèmes à base de règles. Néanmoins, la précision avec laquelle les paramètres sont estimés à partir des mesures est fonction de leur nombre et surtout de leur répartition sur la période de maturation. Les courbes de maturation ne sont que partiellement connues alors que les fonctions sont construites pour décrire toute cette période. L'imprécision des paramètres estimés peut donc être considérable. De la même manière, l'estimation des paramètres à partir des connaissances et hypothèses du viticulteur représente des conditions standards qu'il faut pouvoir ajuster aux conditions réelles. L'utilisation d'un système basé sur la fusion de ces deux techniques d'estimation des paramètres (ajustement à partir des mesures et ajustement à partir des connaissances et hypothèses) permet de combler les faiblesses de chacune : le manque de données de suivi est complété par des informations issues des connaissances et hypothèses du viticulteur.

Les principaux verrous scientifiques levés au cours du développement de cette méthode sont : l'estimation de paramètres en utilisant simultanément différentes sources d'informations (les mesures et les connaissances et hypothèses du viticulteur), la prise en compte des incertitudes entachant ces informations et leur propagation sur la *courbe résultat*.

Les applications directes de ces travaux se font sur le raisin mais les méthodes développées au cours de cette thèse sont facilement transposables à d'autres domaines.

Mots-clés - Suivi de maturité, raisin, aide à la décision, spectroscopie visible proche infra-rouge.

Abstract - ITAP Research Unit of Cemagref and PELLENC S.A. company have jointly developed a near infrared spectroscopy measurement system which aims at characterizing and monitoring various ripeness criteria of wine grapes before harvest (patent register work-in-progress). These criteria are : sugar content, anthocyanin concentration, water content and total acidity.

The main objective of this work is to predict the "most probable" curve of these maturity criteria by using two kinds of information : measures collected with this sensor and the expertise and assumptions made by the winegrower. Through these predictions, the system estimates an optimal harvest date according to the winegrower's target.

Each curve is modelized by an explicit function with specific parameters. For interpretation reasons, the selected functions have an intentionally reduced number of parameters. The method developed in this thesis is therefore intended to estimate these parameter functions by using sets of measurements and the winegrower's knowledge and assumptions. These two sources of information contain uncertainty.

Taken independently, these two information sources are able to generate the most probable values for the parameters ; sensor measurements by using fitting methods and winegrower's knowledge and assumptions by using rule systems. Nevertheless, for the sensor measurement case, the precision in the parameter estimation depends on the number of measurements and their distribution over the ripeness period. Ripeness curves are partially known while functions are built to describe all this period. The estimated parameter uncertainties can be considerable. In the same way, estimate parameters from winegrower's knowledge and assumptions symbolise standard conditions which must be adjusted with actual conditions. A system based on the fusion of these two parameter estimation techniques (fitting from measures and adjustment from knowledge and assumptions) enables one to manage and avoid the weaknesses of such techniques : the lack of data is supplemented by information from winegrower's knowledge and assumptions.

Scientific problems that rose during the development of this method were : (i) parameter estimation using simultaneously two sources of information (measures and winegrower's knowledge and assumptions), (ii) uncertainties in this information and their impact on the result curve.

Direct applications of thesis work shall be carried out on the wine grape but the developed methods are easily transferable to other fields.

Keywords - Ripness monitoring, grape berry, decision aid, near-infrared spectroscopy.

Vincent GERAUDIE.
- Pellenc SA - Département Recherche et Développement - Route de Cavaillon BP 47 - 84122 Pertuis Cedex (France).
- CEMAGREF Montpellier - UMR Information et Technologies pour les Agro-Procédés - 351 rue Jean François Breton - 34196 Montpellier Cedex 5 (France).

Chapitre 1

Introduction

Sommaire

1.1 Contexte social

La viticulture française doit faire face depuis une vingtaine d'années à de nombreuses difficultés : apparition des vins du nouveau monde sur le marché de l'exportation, stagnation de la consommation de vin dans l'Union Européenne, abandon progressif par les consommateurs des vins dits de masse non identifiés vers des vins de qualité [23]. Afin de faire face à ces évolutions du marché et ainsi assurer les débouchés nécessaires à ses productions, le secteur viticole a engagé une profonde mutation.

Cette mutation se traduit en partie par un changement d'objectif des exploitations viticoles. Le principal objectif n'est plus d'assurer un volume de production mais une *Qualité* de production afin d'accroître sa compétitivité. Cette volonté d'optimisation de la *Qualité*, dans un cadre global d'amélioration de la performance, constitue encore à l'heure actuelle une priorité pour la viticulture française. A titre d'exemple, en Languedoc-Roussillion, certains cépages traditionnels dit gros producteurs comme l'aramon, sont en voie de disparition, alors que les cépages aromatiques de type Grenache, Syrah, Merlot, qui représentaient 27% de la superficie du vignole en 1988 en représentent actuellement 54%. De plus, l'amélioration de la qualité passe naturellement par l'obtention de raisins sains, de haute qualité, c'est-à-dire présentant une maturité optimale au regard des impératifs des procedés d'élaboration propres à chaque type de vin.

La qualité du raisin est une notion multicritère, qui s'appréhende par la maturité technologique (sucre, acidité, pH), la maturité aromatique et la maturité phénolique. Dans ce contexte, un besoin d'informations fiables et rapidement disponibles sur la mesure et le suivi de la maturité est apparu.

1.2 Maturation et maturité du raisin

La maturation du raisin peut être définie comme la période qui va de la véraison à la récolte. Il est classique de résumer l'ensemble des processus biochimiques de la maturation par la transformation du raisin vert, dur et acide, en un fruit coloré, souple, sucré et riche en arômes [80, 39]. A la différence de la véraison qui constitue un évènement parfaitement défini, la maturité du raisin ne constitue pas un stade physiologique manifeste [34]. En effet, selon les objectifs choisis et les critères d'appréciation retenus, il est possible de décrire différentes maturités [34, 5].

1.2.1 La maturation du raisin

La croissance de la baie de raisin comporte deux phases de croissance successives, séparées par une phase de ralentissement [13, 80]. Il est possible d'ajouter à ce schéma de croissance une phase de décroissance à la fin de la maturation correspondant à la surmaturation [9]. (Voir fig. 1.1).

La *phase I* ou *phase de croissance herbacée* se caractérise par une intense activité métabolique. Une importante multiplication cellulaire, une respiration intense ainsi qu'une accumulation rapide d'acide ont lieu. La chlorophylle est le pigment prédominant. Durant la *phase II* ou *plateau herbacé*, il se produit un ralentissement de la croissance

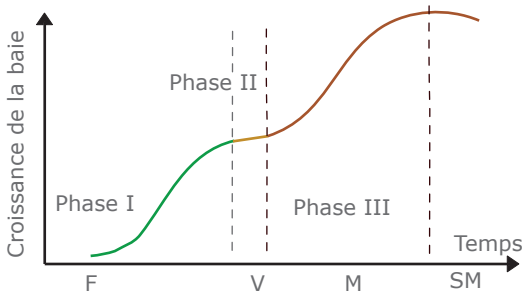


Figure 1.1 – Etapes de la croissance de raisin. F) Floraison V) Véraison M) Maturation SM) Surmaturation.

accompagné d'un profond bouleversement dans le métabolisme des baies. Les baies vont rapidement perdre leur chlorophylle et commencer à se pigmenter. La véraison marque la fin de cette période. La *Phase III* correspond à la maturation proprement dite. La croissance cellulaire reprend et s'accompagne de diverses modifications physiologiques. La pulpe accumule des sucres libres et les teneurs en acides malique et tartrique diminuent. La pellicule se charge en produits secondaires d'une importance œnologique majeure : les composés phénoliques et les substances aromatiques. A la fin de la maturation, lors de la surmaturation, une diminution de flux entrant d'eau engendre une perte de volume et de poids de la baie. La figure 1.2 présente de manière schématisée, les différentes parties d'une baie de raisin citées précédement.

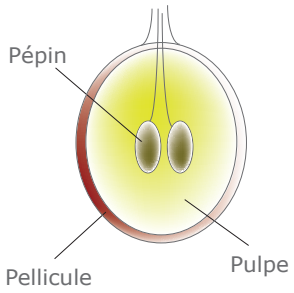


Figure 1.2 – Schéma d'une coupe transversale d'une baie de raisin.

Pris indépendamment, les principaux constituants de la baie, évoluent de manière non simultanée au cours de la maturation. De plus, la manière d'aborder la mesure de ces constituants, en quantité (Voir fig. 1.3.a) ou en concentration (Voir fig. 1.3.b), aboutit à des cinétiques très différentes.

1.2.2 La maturité du raisin

Il est possible de distinguer différentes maturités [34, 5] : celle de la pulpe, correspondant à une teneur optimale en sucre et en acidité dite *maturité technologique* ou *industrielle* ; celle de la pellicule, stade auquel les composés phénoliques et les substances aromatiques ont atteint un optimum dite *maturité aromatique* et *maturité phénolique*. Le biologiste va quant à lui définir la *maturité physiologique*. C'est-à-dire le moment où les pépins sont aptes à germer. Cette dernière est sans importance œnologique ou technologique.

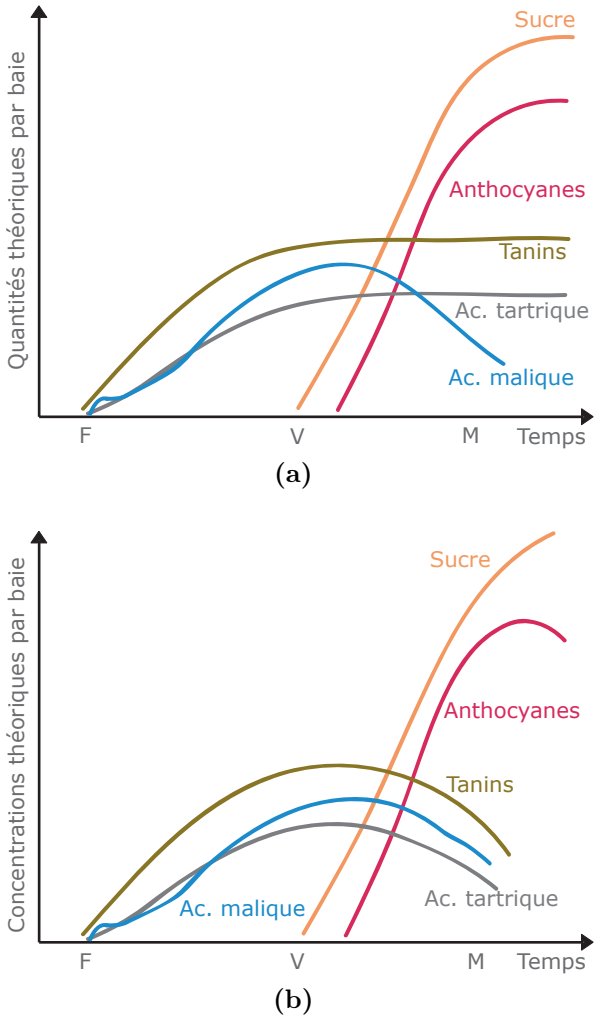


Figure 1.3 – Evolution des principaux constituants par baie a) mesure en quantité par baie b) mesure en concentration par baie.

La maturité technologique peut être approchée par différents indices de maturité (p. ex. rapport du sucre sur l'acidité de la pulpe). La maturité phénolique tient compte de la teneur globale en molécules de cette famille : les anthocyanes, les polyphénols et les tanins mais aussi leur aptitude à l'extraction [80]. Ainsi, la maturité phénolique peut se définir comme le niveau de maturité permettant d'obtenir simultanément un potentiel important en composés phénoliques et leur bonne capacité à diffuser dans le vin futur.

La *maturité œnologique* est définie comme celle qui permettra d'obtenir le meilleur vin possible, pour la situation considérée et le type de vin recherché [34, 5]. Cette maturité sera donc un compromis entre les différentes maturités présentées ci-dessus. Par exemple, une importante teneur en sucre, une faible acidité, une couleur riche et une palette aromatique complète sont les principaux critères pour déclencher les vendanges d'un raisin destiné à produire un vin rouge. Mais l'accroissement trop rapide du taux de sucre peut imposer un ramassage prématuré, alors que les autres éléments ne sont pas arrivés à un parfait degré de maturité. A l'inverse, il peut être préférable de prolonger la maturation lorsque des conditions climatiques défavorables n'ont pas permis une cinétique normale de tous les critères de maturité. Ce compromis, qui dépend de nombreux paramètres, a pour objectif principal le maintien du potentiel qualitatif du raisin.

1.2.3 Comment aborder la maturation et la maturité ?

La multiplicité des critères de qualité du raisin (ou de la baie) impose de l'aborder comme un système complexe où sont transportés, synthétisés, transformés ou stockés les éléments essentiels (eau, sucres, acides organiques, éléments minéraux et autres métabolites secondaires). Ce système est d'autant plus complexe que ses constituants peuvent varier de façon coordonnée ou indépendante et que la vigne est soumise à un environnement changeant durant la phase de maturation : température, rayonnement incident, état hydrique du sol, etc. En effet, la composition de la baie en sucres (glucose, fructose, saccharose), en acides (acide malique et tartrique) et en composés phénoliques (tanins, anthocyanes, etc.) est très dépendante de l'état hydrique de la vigne, du microclimat des grappes et de l'architecture de la végétation [8, 20, 26, 75].

Le viticulteur peut donc légitimement se poser de nombreuses questions sur la manière d'aborder la maturation et la maturité de ses parcelles [9] :

- La maturation se déroule-t-elle normalement ?
- La maturité phénolique est-elle atteinte ?
- La maturité phénolique est atteinte mais la concentration en sucre augmente ; faut-il vendanger ?
- La vendange est-elle hétérogène ?
- Quand vendanger par rapport à un style de vin donné ?

Seule l'étude de la dynamique des critères de maturité, obtenue à partir d'analyses faites sur le raisin, permet de répondre à ces questions. Il est donc très important de suivre dans le temps l'évolution du raisin au niveau de sa croissance (poids et volume) et de sa biochimie (composition) [9].

Le terme *suivi*, dans la suite du manuscrit, fait indirectement appel aux notions de *temporel*, *dynamique* ou *cinétique*.

1.3 Suivi de la maturation

1.3.1 Mise en place du suivi de la maturation

Le suivi de la maturation doit obéir à certaines règles afin de fournir une information fiable. De plus, un suivi de maturation efficace est un compromis entre une série d'analyses complètes - donc longues et coûteuses - et une série d'analyses sommaires (sucre et acidité).

Le suivi doit commencer suffisamment tôt pour obtenir la dynamique d'évolution des critères de maturité [80]. En fonction du niveau d'information souhaité, ce suivi sera fait de la nouaison à la maturité-vendange ou durant la phase de maturation seulement [9]. Habituellement, le viticulteur débute le suivi analytique 20 à 25 jours après la mi-véraison et suit précisément quelques parcelles représentatives de son domaine. Il procède ensuite à un suivi plus approfondi 30 à 35 jours après la mi-véraison [6]. Une fois les parcelles de référence arrivées à maturité, les contrôles sont étendus aux parcelles ayant une maturité équivalente. Ainsi le programme prévisionnel des vendanges peut être réalisé.

Le suivi de la maturation se heurte à la grande variabilité de la constitution des baies d'une même grappe ou d'une même parcelle à un moment donné [4, 5, 9]. En effet, le raisin d'une même parcelle à un même moment est très hétérogène. Néanmoins, il est possible de caractériser l'état de la parcelle en suivant des protocoles d'échantillonnage bien définis. Ainsi, les échantillons collectés fournissent le même bulletin d'analyse que la parcelle entièrement récoltée aurait fourni. Il existe plusieurs méthodes de prélèvement (méthode des 200 baies, portion de grappe ou grappe entière) [10, 80].

1.3.2 Suivi de la croissance du raisin

Le poids moyen des baies augmente régulièrement de la véraison à la maturité. A maturité le poids reste stable avant de légèrement diminuer par perte d'eau [80, 13].

Le suivi du volume permet d'apprécier l'homogénéité des baies d'une parcelle. Il permet également d'apprécier le chargement en sucre et en métabolites secondaires des baies [9].

1.3.3 Suivi de la composition du raisin

La teneur en sucre [1] augmente lentement puis de manière plus importante au cours de la maturation en fonction des conditions climatiques. Cette teneur va ensuite se stabiliser avant de croître lentement sous l'effet d'une perte en eau de la baie au cours de la surmaturation [80]. De même, le suivi de la quantité ou du chargement en sucre des

1. La teneur s'exprime dans une unité relative faisant appel à la notion de concentration (P. ex. $g.l^{-1}$, $^{\circ}$ Brix, etc)

baies [2], qui permet de s'affranchir de la quantité d'eau dans la baie, est un indicateur du fonctionnement de la vigne. En effet, il permet de savoir si la vigne est active et accumule des sucres ou bien les concentre [9]. Des études ont mis en évidence qu'à partir de la cinétique de chargement en sucre, il est ainsi possible d'approcher une date de vendange en fonction du style de vin recherché [7]. La charge ou quantité de sucre par baie est accessible à partir des mesures de volume de la baie et de la teneur en sucre.

Le suivi des acides apporte d'importantes informations pour définir la maturité technologique [9]. Après avoir atteint son maximum après la véraison, l'acidité totale diminue jusqu'à la fin de la maturation. En fonction du type de vin recherché, l'acidité est un critère déterminant pour déclencher les vendanges. Si cette dernière est trop faible, le vin risque d'être *plat*. A l'inverse, si elle est trop importante, le vin sera *désagréable* [80]. Cette évolution est différente en fonction des acides. La quantité d'acide malique diminue ; sa consommation par respiration et par transformation est plus importante que son afflux. La quantité d'acide tartrique reste quant à elle sensiblement constante ; les pertes par respiration sont équilibrées par son afflux et sa synthèse [39].

Dans l'ensemble, les composés secondaires (tanins, anthocyanes et polyphénols) augmentent au cours de la maturité en passant par un maximum à maturité avant diminuer. Néanmoins, en fonction de leur localisation et de leur nature, leur dynamique d'évolution ainsi que leur extractibilité sont différentes [80]. Les anthocyanes et les tanins augmentent dans la pellicule alors que les tanins diminuent dans les pépins. Il en est de même pour leur extractibilité. La dynamique de chargement des anthocyanes et des polyphénols permet de savoir si leur biosynthèse est active, stimulée ou inhibée [9]. Leurs suivis, en particulier celui des anthocyanes, sont donc un bon indicateur de la maturation du raisin.

Plusieurs indices de suivi de maturité sont également proposés dans la littérature. Parmi ces indices il est possible de citer : le rapport acidité tartrique sur acidité malique, le rapport de l'indice réfractométrique sur acidité totale, le rapport glucose sur fructose, etc. [5, 39].

Le plus connu est l'*indice de maturation* défini par le rapport du sucre (en $g.l^{-1}$) sur l'acidité (en $g.l^{-1}$ de H_2SO_4) de la pulpe [39]. A partir des données historiques, pour un même cépage dans des situations similaires, il est possible de dire que plus cet indice est élevé, plus le vin sera de qualité. Cet indice simple doit néanmoins être utilisé avec précaution. Par exemple, il ne permet pas de comparer les cépages entre eux [34]. Une excellente vendange doit avoir des valeurs comprises entre 35 à 40 dans certaines régions alors que ces valeurs sont différentes ailleurs. De plus, cet indice ne tient pas compte de la pellicule (arômes, polyphénols).

Ces indices sont de bons outils d'aide à la décision si leurs limites sont prises en compte. En effet, aucun ne prend en compte l'ensemble des éléments permettant de définir la maturité œnologique optimale.

2. La quantité ou le chargement s'exprime dans une unité absolue faisant appel à la notion de masse (P. ex. g par baie)

1.3.4 Utilisation des suivis de maturation

Un suivi temporel et multi-critères des baies a une grande importance en viticulture. Il fournit de nombreuses informations au viticulteur. Ces informations, associées à ses propres connaissances, lui permettent :

- de contrôler le bon déroulement de la maturation, de gérer au mieux la conduite de ses parcelles [51],
- d'appréhender le devenir qualitatif de son futur millésime [55]
- de planifier le plus efficacement possible son chantier de récolte en fonction d'une maturité "cible" [46],
- d'identifier la/les causes d'un éventuel dysfonctionnement en recoupant les différentes dynamiques de chargement présentées ci-dessus [9].

De plus, tous les vins prétendant à une Appellation d'Origine Contrôlée doivent être conformes aux préconisations de l'INAO. Les critères du cahier des charges portent notamment sur le rendement maximum, le degré alcoolique minimal, etc. Il est donc primordial pour le viticulteur de déclencher les vendanges au bon moment.

Toutefois, ces dynamiques de maturation ne sont pas constantes entre cépages ou pour un même cépage d'une année sur l'autre ou d'une région à l'autre. En effet, l'environnement (climat, sol, parasites, etc.) et la conduite de la parcelle (architecture de la plante, espacement, fumure, irrigation, etc.) ont une grande influence sur l'évolution de ces critères de maturité.

1.4 Principaux paramètres d'influence climatique sur la maturation

De nombreux paramètres climatiques ont un effet sur les voies métaboliques. Cette influence se traduit alors par une modification de la composition du raisin. A partir des historiques de données météorologiques, l'expérience montre qu'il est possible de dégager des tendances générales.

1.4.1 Effet de la lumière

La lumière a différents effets sur le raisin. D'une part, elle fournit l'énergie nécessaire à la photosynthèse et stimule plusieurs processus métaboliques et d'autre part, elle a un effet radiatif entraînant un échauffement du raisin et de son air environnant [34, 39]. Cette dualité des effets de la lumière rend difficile les études tentant de mettre en évidence les effets individuels de la lumière et de la température.

La disponibilité en rayonnement du milieu ne constitue pas, en général, un facteur limitant sur le fonctionnement de la vigne. Néanmoins dans la pratique, certains systèmes de conduite peuvent entraîner une perte de cette énergie lumineuse. Les grappes à l'ombre sont toujours moins sucrées et plus acides que celles au soleil [34]. Ce constat peut également s'étendre au niveau des années : les années de grande insolation donnent des raisins riches en sucre et pauvres en acide [39].

1.4.2 Effet de la température

La température est l'un des paramètres les plus influents sur la maturation. Elle a une influence sur l'ensemble du cycle de la vigne : débourrement, floraison, maturation et qualité finale du raisin [34].

Lors de la phase herbacée, des températures trop élevées sont défavorables à la multiplication cellulaire [9]. L'optimum de température varie en fonction de la littérature : entre 20 °C et 25 °C [34], 24 °C pour les cépage septentrionaux et 28 °C pour les cépages méridionaux [39].

Durant la maturation, des températures trop importantes favoriseront l'accumulation des sucres vers d'autres parties de la plante. La température va également influencer la dégradation de l'acide malique. En effet, l'activité de l'enzyme malique augmente régulièrement entre 10 °C et 46 °C [34].

De fortes températures auraient donc un effet négatif sur la qualité globale des raisins. Pour certains auteurs, cet effet peut être contrebalancé par une importante amplitude thermique entre la journée et la nuit. De telles amplitudes seraient l'origine de la qualité des vins blancs de climats dits "froids" quand ils sont comparés aux vins blancs de climats dits "chauds" (ces derniers ne conservant pas les acides organiques) [51]. Néanmoins, la température nocturne compte plus que l'amplitude thermique. En effet, il est possible d'avoir des climats dits "chauds" avec une amplitude importante (p. ex. 30 °C durant la nuit et 45 °C durant la journée) mais la qualité du raisin obtenue avec de telles conditions sera mauvaise. L'indice de fraîcheur de nuit permet d'appréhender ce problème [94].

La température joue également un rôle déterminant dans la synthèse des anthocyanes et de certains composés aromatiques. D'importantes températures diurnes ne sont pas favorables à l'obtention d'un raisin riche en anthocyanes [51, 91]. De faibles températures nocturnes ne permettent pas cependant de contrebalancer les effets négatifs de températures diurnes importantes [91].

Différents indices comme le produit héliothermique de Branas (1946), les degrés-jours de Winkler (1962) ou bien encore l'indice héliothermique modifié de Huglin (1978), pour ne citer que les plus connus, ont été développés pour appréhender le potentiel d'un climat donné pour une bonne maturation du raisin [5, 9].

1.4.3 Contrainte hydrique

Une contrainte hydrique trop faible, due à un excès de précipitation ou d'irrigation, provoque une diminution de la qualité du raisin en modifiant considérablement sa composition [34].

Avant la véraison, une contrainte hydrique trop forte peut avoir des conséquences irréversibles sur le grossissement des baies même si cette contrainte diminue après la véraison [9, 34]. Cette contrainte forte n'affecte pas la division cellulaire mais diminue le volume des cellules [76].

Durant la maturation, une contrainte hydrique modérée sera favorable à l'obtention d'un raisin de qualité : teneurs en sucre et en anthocyanes élevées associés à une faible acidité [9]. Mais une contrainte hydrique trop forte ou trop faible empêche de parvenir à un état de maturité satisfaisante [26].

Une forte pluie au voisinage de la maturité peut provoquer une importante absorption d'eau à travers la pellicule qui aura comme conséquence un éclatement des baies [34].

Sur la base des recherches et des informations empiriques actuellement disponibles, l'état hydrique optimal pour la vigne par rapport au cycle végétatif et à l'intensité de la contrainte hydrique a pu être évalué [74]. La figure 1.4 présente ces résultats.

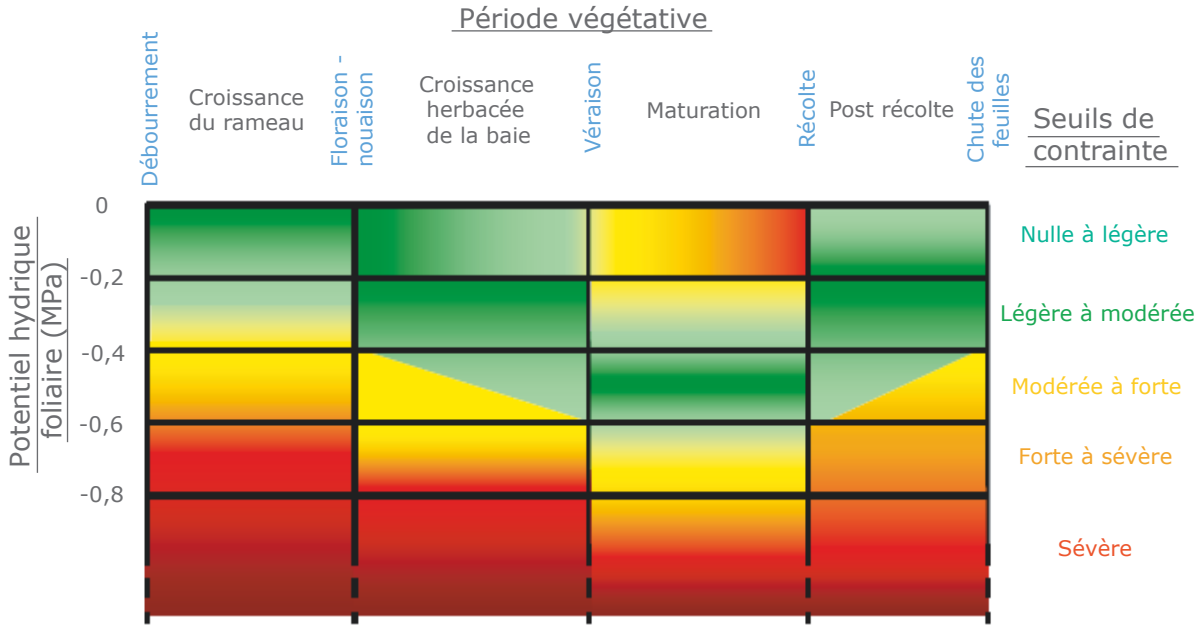


Figure 1.4 – Etats hydriques favorables (zones vertes), déconseillés (zones jaunes) et défavorables (zones rouges) [74].

1.4.4 Interactions entre ces facteurs

Au cours des journées d'été, le microclimat des feuilles devient défavorable à la photosynthèse par le cumul d'une énergie lumineuse importante combinée à de fortes températures. Si la contrainte hydrique est trop forte, la transpiration ne peut plus jouer son rôle de régulateur thermique. La température des feuilles augmentera donc encore, ce qui aura pour conséquence la diminution, voir l'arrêt de la photosynthèse [39].

Des essais sur de la Syrah ont montré que le chargement en sucre est fonction de la contrainte hydrique et du rayonnement lumineux [99].

Néanmoins, il est important de rappeler que les facteurs ayant une influence sur la qualité du raisin ne sont pas seulement climatiques. L'évolution des différents critères de maturité, notamment pour les teneurs en sucre, anthocyanes et l'acidité totale, provient de l'interaction subtile et complexe de l'environnement, du mode de conduite et de la

vigne en elle-même [8, 20, 75]. De plus amples renseignements peuvent être trouvés dans de nombreux ouvrages spécialisés [5, 9, 34, 39].

1.5 Méthodes d'analyse du suivi de la maturation

Aujourd'hui, afin d'accéder aux concentrations ou aux quantités des critères de maturité, différentes analyses doivent être réalisées au cours de la maturation. En fonction des informations souhaitées ou de l'état d'avancement de la maturation, ces analyses seront plus ou moins nombreuses (Voir chap. 1.3.1).

1.5.1 Les méthodes d'analyse dites "classiques"

Traditionnellement, le moût, c'est-à-dire le jus de pulpe, issu des baies échantillonnées est obtenu par différent procédés en fonction des analyses réalisées (p. ex. pressurage simple pour un contrôle sommaire, pressurage plus centrifugeuse à fruit pour un contrôle des composés phénoliques) [5].

La détermination de la teneur en sucre, qui est essentielle, s'effectue le plus souvent par une mesure physique indirecte telle que la densimétrie ou la réfractométrie. Le résultat est exprimé en diverses unités en fonction des instruments utilisés et/ou des habitudes du viticulteur [34, 39]. A titre d'exemple :

- En France, la densité relative apparente à 20 °C sera mesurée avec un mustimètre. Elle fournit le poids en gramme d'un litre de moût. La Teneur en Alcool Probable ou Puissance (TAP) est également utilisée. Elle est calculée à partir de la teneur en sucre exprimée $g.l^{-1}$. Le degré Brix ou degré Balling fournit le poids en sucre du moût en gramme pour cent gramme de moût. Il s'agit en réalité du pourcentage de matière sèche soluble (MSS) dans le moût, mesuré avec un réfractomètre ou un densimètre. Cette mesure n'est valable qu'au dessus de 15 °Brix. En effet, en début de maturité, lorsque la teneur en sucre est faible, l'approximation teneur en sucre égale MSS est fausse.
- En Suisse et en Allemagne, le degré œchsle est utilisé. Il correspond aux deux derniers chiffres de la densité.
- En Australie, le degré Baumé sera préféré. Il est mesuré avec un densimètre étalonné à l'aide d'une solution de sel marin.

Différentes formules permettent de passer d'une unité à l'autre :

- 1 % vol. alcool = 16,83 $g.l^{-1}$ de sucre
- Sucre ($g.l^{-1}$) = (Densité relative - 1) × 2000 + 16
- °Baumé = 144,32(1 - 1/Densité relative).

L'acidité totale est obtenue par titration. Elle s'exprime en $g.l^{-1}$ d'acide sulfurique (H_2SO_4) ou d'acide tartrique ($C_4H_6O_6$) [34]. Les deux principaux acides également dosés dans le moût sont l'acide tartrique et l'acide malique et secondairement l'acide citrique. L'acidité peut également s'exprimer en milliéquivalents (meq). 1 meq correspond à 75 $mg.l^{-1}$ pour l'acide tartrique, 67 $mg.l^{-1}$ pour l'acide malique et 64 $mg.l^{-1}$ pour l'acide citrique [39]. Les dosages doivent être rapidement réalisés après le prélèvement. L'acide

tartrique peut précipiter si les échantillons sont conservés au froid ou sur la paillasse [9].

Les techniques permettant d'accéder à l'évolution des composés phénoliques au cours de la maturation sont lourdes et fastidieuses [34]. La baie dans son ensemble (pulpe et pellicule) est utilisée pour ces analyses. L'hétérogénéité des baies en matière de composés phénoliques et leur localisation spécifique dans la pellicule entraînent des difficultés pour apprécier leur état de maturité phénolique. Néanmoins, les méthodes d'analyse par spectrométrie sont fiables et ont démontré leur utilité [9]. Une méthode rapide a été mis au point par l'Institut Français de la Vigne et du Vin (IFV) pour approcher le Potentiel polyphénolique de la vendange (broyage des baies suivi d'une extraction dans une solution d'éthanol et d'acide chlorhydrique et mesure de l'absorbance à 280 et 520 nm).

A l'heure actuelle, aucune méthode simple ne permet d'obtenir rapidement un indice de maturation pour les substances aromatiques et leur aptitude à l'extraction. Leurs dosages sont très couteux et restent du domaine de la recherche ou de laboratoires particulièrement équipés [34, 9].

La dégustation ou *analyse sensorielle* de la baie de raisin dans son ensemble (pulpe, pellicule et pépins) est un bon critère de jugement pour suivre la maturité. Les caractères aromatiques de la pulpe et de la pellicule sont particulièrement intéressants [34]. Même s'il s'agit d'une méthode empirique et subjective, avec de l'habitude, la dégustation permet d'apprécier l'évolution des critères de maturité. L'appréciation de la maturité ou de l'équilibre global entre les critères de maturité est plus délicate à obtenir [9].

Différents organismes publient des protocoles d'analyse comme l'Institut Français de la Vigne et du Vin (IFV) ou bien encore l'Organisation internationale de la vigne et du vin (OIV). Le Recueil des méthodes internationales d'analyse des vins et des moûts a été publié pour la première fois en 1962 par l'OIV. Il a été réédité de nombreuses fois en intégrant les textes complémentaires établis par la sous-commission des méthodes d'analyse et d'appréciation des vins. Ce Recueil joue un grand rôle pour l'harmonisation des méthodes d'analyse. Plusieurs pays ont introduit dans leur réglementation propre ces définitions et ces méthodes.

Toutes ces analyses sont réalisées à l'aide de méthodes dites "classiques" dont les principaux défauts sont :

- d'être des méthodes destructives,
- d'obtenir les résultats après un certain délai,
- d'être généralement longues à mettre en œuvre,
- de nécessiter l'emploi de réactifs coûteux et souvent polluants,
- de ne pouvoir être appliquées que par des opérateurs qualifiés.

Ce sont toutefois celles qui sont utilisées traditionnellement par la profession.

1.5.2 Les méthodes d'analyse par spectroscopie visible - proche infrarouge (Vis-NIR)

La spectroscopie optique, quelle que soit la gamme de longueur d'onde utilisée (UV, visible, infra rouge, etc.), a pour objectif d'extraire des informations sur la matière à

partir de son interaction avec la lumière. En pratique, l'analyse peut être qualitative (identification d'un composé à partir de sa signature spectrale), ou quantitative (dosage d'une substance).

La spectroscopie Vis-NIR utilise la gamme des longueurs d'onde comprise entre 450 et 2500 nanomètres. Cette technique d'analyse permet de connaître, de manière rapide et non destructive, la composition chimique de nombreux produits. En laboratoire, la mesure est déjà performante pour un grand nombre de produits alimentaires. Son application à la mesure de la qualité des fruits connait ces dernières années un important développement [73].

Les possibilités offertes par cette technique ont conduit au développement d'outils mobiles permettant des mesures au champ. Les études visant à prédire la composition des fruits, directement au champ, dans le but d'évaluer leur qualité avant la récolte permettent de souligner les capacités prometteuses de la spectroscopie Vis-NIR portable à "bas coût" [98].

Concepts de base de la spectroscopie Vis-NIR

La gamme de longueur d'onde utilisée en spectroscopie Vis-NIR fait partie du domaine de la spectroscopie vibrationnelle. Il est possible de représenter, de manière très simplifiée, une molécule comme un ensemble d'atomes reliés entre eux par des ressorts. Ces liaisons sont le résultat d'un équilibre des forces de liaisons. D'une part, il existe au sein d'une molécule une répulsion entre les noyaux chargés positivement et entre les nuages d'électrons chargés négativement. D'autre part, il existe une attraction entre le noyau des atomes et les électrons. Chaque ressort vibre a une fréquence qui dépend du groupe chimique impliqué dans la liaison. L'énergie d'un rayon lumineux incident ne peut être absorbée que si la fréquence de la lumière ν [3] est identique à la fréquence de la liaison chimique.

Les spectres sont acquis à l'aide de spectrophotomètres. Ils sont formés d'au moins 4 éléments :

- une source lumineuse,
- un système permettant de présenter l'échantillon,
- un système de séparation de lumière polychromatique en fonction des longueurs d'onde,
- un système de mesure photosensible.

L'échantillon est irradié et le rayonnement transmis, rétro-diffusé ou réfléchi (Voir fig. 1.5) est "chargé" en information. En effet, après avoir pénétré dans le produit, les caractéristiques spectrales du rayonnement incident sont modifiées en fonction des processus de diffusion et d'absorption. Ces changements sont fonction des longueurs d'onde, de la composition chimique du produit, ainsi que des propriétés de diffusion de la lumière du produit.

3. λ (la longueur d'onde) = c (la célérité) / ν (la fréquence de l'onde)

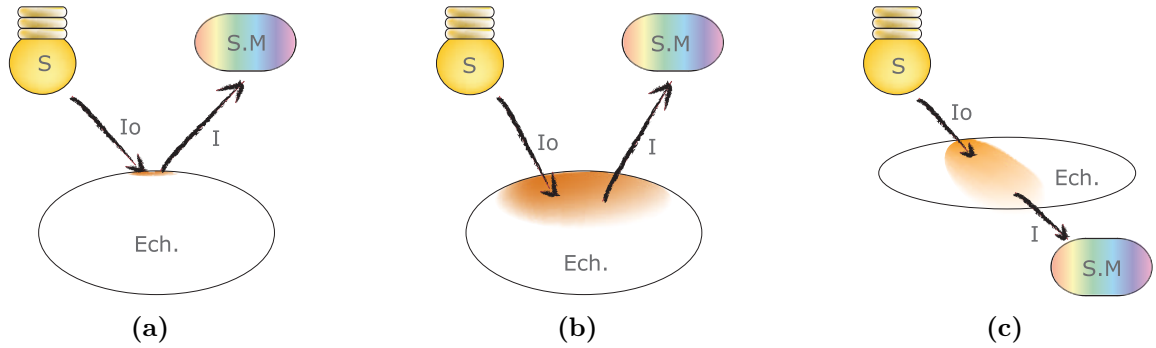


Figure 1.5 – Disposition de la source (S) de l'échantillon (Ech) et du Système de mesure (SM) pour obtenir un spectre en : a) réflection b) rétrodiffusion c) transmission.

L'absorption est modélisée par la loi de Beer-Lambert (Voir éq. 1.1). Cette loi relie l'intensité du rayonnement transmis, $I(\lambda)$, à celle du rayonnement incident, $I_0(\lambda)$, en fonction la longueur du chemin optique, l exprimé en cm, et de la concentration, c exprimée en $mol.l^{-1}$, du composé présent dans la solution :

$$I = I_0 e^{-k(\lambda)\ l\ c} \tag{1.1}$$

où $k(\lambda)$ est le Coefficient d'extinction, fonction de la longueur d'onde λ ($l.mol^{-1}.cm^{-1}$).

Le spectre présente des pics d'absorption centrés sur des longueurs d'onde propres au composé. Ceci forme un spectre caractéristique du composé chimique ayant interagi avec le rayonnement incident. Dans des conditions idéales, en utilisant une gamme de longueurs d'onde appropriée et en maitrisant le chemin optique, il est possible de déterminer la concentration c à partir du spectre d'absorbance, théoriquement proportionnel au spectre caractéristique :

$$A(\lambda) = -log_{10}\frac{I(\lambda)}{I_0(\lambda)} = k(\lambda)\ l\ c \tag{1.2}$$

où I/I_0 est la transmittance et $A(\lambda)$ est l'absorbance à la longueur d'onde λ (sans unité). Dans ce cas idéal, il n'est pas nécessaire d'utiliser un spectrophotomètre. La mesure de l'absorption à une longueur d'onde correctement choisie permet de connaitre la concentration. Cette mesure simple est celle de la densité optique (DO).

Si plusieurs composés sont présents, les absorbances se combinent. L'absorbance mesurée peut alors être considérée comme la somme des absorbances dues à l'ensemble des composés. Il convient également d'ajouter un spectre d'absorption dû à des phénomènes physiques comme la diffusion dans le milieu. Enfin, il faut tenir compte des interactions d'ordre 1 entre les solutés. Une telle démarche analytique devient complexe. En effet, elle nécessite la connaissance des spectres d'interaction. Tous ces modèles sont donc des approximations.

Les spectres Vis-NIR contiennent donc des informations pertinentes sur les caractéristiques chimiques des échantillons mesurés mais ces dernières sont "cachées" dans les spectres. Des méthodes de chimiométrie doivent alors être mises en œuvre pour extraire correctement ces informations. La chimiométrie peut être définie comme étant :

"*la discipline qui utilise les mathématiques, les statistiques et la logique formelle (a) pour concevoir ou choisir des procédures expérimentales optimales ; (b) pour fournir le maximum d'informations chimiques d'intérêt en analysant les données chimiques, (c) et pour obtenir des connaissances sur les systèmes chimiques*" [63]. Cette discipline est apparue à la suite du développement des instruments de mesure chimique rapides utilisés pour les analyses de routine. En effet, ces instruments fournissent des mesures indirectes représentées au sein d'importantes bases de données spectrales qui nécessitent d'être interprétées afin d'extraire l'information chimique souhaitée.

Concepts de base de la chimiométrie

Différents prétraitements des spectres permettent une meilleure extraction de l'information. Par exemple, il est possible de réaliser :

- la moyenne de plusieurs spectres lors de l'acquisition afin de réduire le bruit thermique du détecteur. Le nombre de spectres moyennés dépend de la rapidité d'acquisition d'un spectre. En effet il ne doit pas affecter le débit des mesures. Ce nombre dépend également du rapport signal sur bruit du spectrophotomètre. Plus ce rapport est faible, plus le nombre de spectres à moyenner doit augmenter.
- la normalisation des spectres afin d'améliorer la linéarité de la relation entre l'absorbance et la concentration. Ce prétraitement est donc intéressant lorsque une technique de régression linéaire est utilisée par la suite. La normalisation consiste à diviser par exemple chacune des longueurs d'onde par la racine de la somme au carré de l'ensemble des longueurs d'onde.
- la dérivation des spectres afin de réduire la ligne de base et de séparer plus clairement les bandes d'absorption. La dérivée d'ordre deux est souvent préférée car elle permet d'enlever les effets additifs du type $a\ \lambda\ +\ b$.
- le calcul de l'absorbance, même si normalement la pénétration de la lumière dans un tissu biologique est beaucoup plus complexe.

Une fois les spectres prétraités, une technique de régression multivariée est utilisée afin d'établir une relation entre la matrice de n spectres de N longueurs d'onde et le vecteur de n mesures de référence que l'on cherche à prédire. A titre d'exemple , la *régression linéaire multiple* (MLR) permet d'estimer la valeur des coefficients à affecter à chacune des longueurs d'onde du spectre afin d'obtenir la teneur en sucre.

Il est ensuite nécessaire d'estimer les performances du modèle obtenu. Pour cela, des techniques de *validation* ou de *test* sont utilisées. La qualité d'un modèle, issu d'un certain type de modélisation, ne peut pas être estimée avec les jeux de données qui ont servi à sa construction. L'erreur qui serait calculée dans ces conditions serait fausse car le modèle est spécialiste de son ensemble d'apprentissage. Cette erreur ne nous dirait absolument rien des capacités qu'aurait le modèle sur de nouveaux jeux données. Beaucoup de modèles peuvent conduire à d'excellents résultats sur les données de construction ou apprentissage et des résultats catastrophiques sur de nouvelles données.

Diverses techniques de simulation ont donc été développées pour estimer les performances d'une modélisation (p. ex. une MLR). L'une d'elles est la *Validation Croisée*. En utilisant une validation croisée, plusieurs modèles (de même architecture) sont construits sur des sous-ensembles disjoints, à partir de l'ensemble des données disponibles. Chacun

des modèles créés est alors testé sur la partie des données qui n'a pas été utilisée pour sa construction. Les résultats sont ensuite combinés pour fournir une estimation des performances du type de modélisation testée (p. ex le calcul du $SECV$). Le *Leave One Out* est un cas particulier de validation croisée. Chacun des ensembles d'apprentissage est constitué de l'ensemble complet des données dont un seul individu a été retiré. Si la base de mesure comprend n mesures, $n-1$ mesures sont utilisées pour l'apprentissage et la mesure restante est utilisée pour le test. L'opération est ainsi répétée n fois.

L'un des critère de performance du modèle est sa précision. Elle est calculée à partir de l'erreur standard (moyenne quadratique de l'erreur), définie comme :

$$SE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \tag{1.3}$$

Où :
- y_i la valeur mesurée
- $\widehat{y_i}$ la valeur prédite par le modèle
- n le nombre de mesures

Si les prédictions ($\widehat{y_i}$) sont obtenues par validation croisée, la précision du modèle sera exprimée par l'erreur standard de validation croisée : le $SECV$. Si les prédictions sont faites sur un ensemble de test, c'est-à-dire à partir d'un échantillon de mesure différent de celui qui a servi à créer le modèle, la précision du modèle sera exprimée par l'erreur standard de prédiction : le SEP, et il peut y avoir un biais[4]. Si le biais est retiré aux y_i, avant de calculer l'erreur standard, la précision du modèle sera exprimée par l'erreur standard de prédiction corrigée : le $SEPc$. Ces derniers sont exprimés dans l'unité de la méthode chimique de référence.

Dans le cas d'une MLR, le coefficient de détermination, R^2, qui correspond au carré du coefficient de corrélation , R, est un autre critère de qualité de la modélisation :

$$R = \frac{Cov(\widehat{y_i}, y_i)}{\sigma_{\widehat{y_i}} \cdot \sigma_{y_i}} \tag{1.4}$$

Où :
- $Cov(\widehat{y_i}, y_i)$ la covariance des valeurs mesurées et prédites
- σ_{y_i} l'écart-type des valeurs mesurées
- $\sigma_{\widehat{y_i}}$ l'écart-type des valeurs prédites par le modèle
- n le nombre de mesure.

Le Coefficient de détermination estime la variance expliquée par la régression. Ce dernier est sans unité.

La spectroscopie Vis-NIR en viticulture

La spectroscopie est un outil de mesure prometteur et performant pour suivre certains critères de maturité du raisin. Au laboratoire, des modèles robustes, basés sur des

4. Le biais est la moyenne algébrique de l'erreur : $biais = \Sigma(y_i - \widehat{y_i})/n$

spectres obtenus en transition, permettent de prédire le taux de sucre avec de faibles erreurs de validation [37, 53]. Les modèles obtenus pour les raisins noirs semblent généralement de meilleure qualité que ceux obtenus pour les raisins blancs. Des essais ont également démontré la possibilité d'évaluer les propriétés de texture des baies sur du Cabernet Franc [59]. Enfin, des essais de mesure au champ, avec des prototypes portables, ont également montré leur potentiel. Ces résultats témoignent de l'intérêt de suivre directement au champ l'évolution de la teneur en sucre, le pH ou la teneur en anthocyanes. Ces essais ont été menés sur du Cabernet Sauvignon, Merlot, Pinot Noir, ou Chardonnay [58, 68]. Le tableau 1.1 présente certains de ces résultats.

Référence	Cépages	λ (nm)	Mode	Critère	SEP
Jaren *et al* (2001)	G. V.	800-2500	R	Sucre ($^\circ$Brix)	1.04 à 1.59
Herrera *et al* (2003)	CS. Ca. Ch.	650-1100	RT	Sucre ($^\circ$Brix)	1.34 à 2.96
Dambergs *et al* (2003)	CS. S. Me. G.	400-2500	R	Sucre ($^\circ$Brix)	1
				pH	0.11
				Antho. (mg/g)	0.05
Chauchard *et al* (2004)	Cg. Mo. UB.	680-1100	R	Acidité ($g.L^{-1}$)	1.28
Arana *et al* (2005)	V. Ch.	500-800	R	Sucre ($^\circ$Brix)	1.27
Cozzolino *et al* (2005)	CS. S. Me.	400-1100	R	pH	0.045
				Antho. (mg/g)	0.06

Table 1.1 – Exemple de mesures de critère de maturité en spectroscopie proche infrarouge [22]. G. Grenache, V. Viura, CS. Cabernet Sauvignon, Ca. Carmenère, Ch. Chardonnay, S. Syrah, Me. Merlot, Cg. Carignan, Mo. Mourvèdre, UB. Ugni-Blanc.

Notons que les essais tendent à montrer que les spectres obtenus en transmission fournissent de meilleurs résultats que ceux obtenus en rétrodiffusion [47].

Les critiques pouvant être formulés à l'encontre de la spectroscopie Vis-NIR sont :
- la nécessité de passer par une importante phase d'étalonnage. En effet, il faut collecter une grande quantité d'échantillons qui doivent être analysés avec les méthodes de référence. De plus, ces échantillons doivent être représentatifs de l'ensemble du raisin qui sera par la suite mesuré. Cette phase d'étalonnage représente une partie longue et onéreuse lors de la mise au point de ce type de mesure.
- l'impossibilité de mesurer les substances minérales. En effet, le rayonnement infrarouge est seulement absorbé par les composés organiques.
- le faible seuil de sensibilité de cette méthode de mesure.

Néanmoins, cette technique présente de nombreux avantages :
- elle est rapide : quelques millièmes de seconde permettent de recueillir le spectre d'absorption d'une grappe pour prédire sa composition. Les méthodes d'analyses biochimiques classiques demandent plusieurs heures, en particulier pour mesurer la teneur en anthocyanes. De plus, elle ne nécessite pas de préparation de l'échantillon.

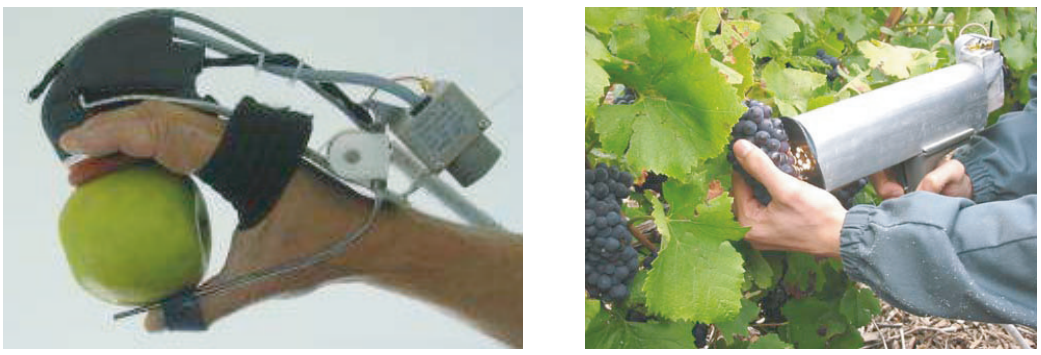
– elle est non destructive : le raisin est directement mesuré sur le cep et laissé intact après la mesure. Cette propriété est particulièrement intéressante car elle permet de réaliser ultérieurement des mesures sur une même grappe.
– elle est peu onéreuse : hormis l'investissement initial dans l'appareil et l'éventuelle constitution d'une base de données pour l'étalonnage. Une fois le spectrophotomètre correctement étalonné, de nombreux critères permettant d'évaluer la maturité peuvent être obtenus à partir d'une même mesure sans coût additionnel.

1.6 Projet Spectron™

1.6.1 Genèse du projet et ses acteurs

Au sein de l'UMR-ITAP (Information et Technologies pour les Agroprocédés) du Cemagref de Montpellier, l'équipe IODE (Image, Optique et Décision) étudie et met au point des systèmes de perception et de décision pour les produits, équipements et agroprocédés. Les systèmes de perception, développés par cette équipe, mettent en œuvre des capteurs optiques basés sur la spectrométrie et la vision artificielle. Comme nous venons de le voir, la spectrométrie Vis-NIR est une méthode d'analyse puissante et rapide pour "interroger" la matière de façon non destructive.

L'UMR-ITAP a coordonné de 1998 à 2001 le projet européen de recherche *Glove* (FAIR PL97-3399) (Voir fig 1.6.a). Ce projet avait pour but de démontrer la faisabilité d'un gant instrumenté permettant de mesurer au champ, de manière non destructive, la qualité des fruits (teneur en sucre, maturité et calibre). Fort de ce succès, l'ITV (Institut Interprofessionnel du Vin) a souhaité appliquer les résultats du projet *Glove* sur le raisin. Pour mener à bien ce nouveau défi, un programme de recherche Cemagref/ITV a été créé (financement ACTA). L'objectif était de développer et d'expérimenter sur le terrain un nouvel ensemble de mesure portable adapté au raisin, le *Tromblon* (Voir fig. 1.6.b). Ce premier pré-prototype a permis de valider le concept d'un appareil portable, capable de mesurer directement sur la grappe, sans la détruire, différents critères de maturité du raisin (sucre, acidité totale).



(a) **(b)**

Figure 1.6 – Prototypes développés au Cemagref. a) Glove b) Tromblon.

La Société PELLENC S.A., spécialiste de la mécanisation en viticulture, conçoit et

met au point des produits pertinents pour les viticulteurs afin d'être toujours en parfaite adéquation avec les demandes mouvantes de ce marché. Ce résultat est principalement dû à son service de Recherche et de Développement, qui reste à l'écoute des besoins de ses clients. A ce titre, plus de 250 brevets ont été déposés de par le monde par PELLENC S.A.

Forts de leur expertise, l'UMR - ITAP et la société PELLENC S.A. ont décidé de développer conjointement un système, portatif et autonome, de mesure par spectroscopie Vis-NIR permettant de suivre, de manière non destructive, l'évolution temporelle de certains critères de maturité du raisin. Ce système est baptisé ***Spectron***™. Ces critères sont : les teneurs en sucre, en anthocyanes et en eau ainsi que l'acidité totale. De plus, un outil d'aide à la décision utilisant les mesures recueillies grâce au Spectron™ devra permettre d'approcher la date des vendanges en fonction des règles et des objectifs du viticulteur.

1.6.2 Etat des lieux du projet Spectron™ en début et fin de thèse

Développement du capteur

Pour récupérer un signal de qualité, nécessaire à l'obtention de mesures fiables, un important travail d'optimisation du capteur a été mené. L'un des défis était d'augmenter la quantité de signal récupérée (utilisation de grande surface captante et de sources lumineuses puissantes) tout en minimisant le bruit. Pour cela, une attention particulière a été portée à la conception électronique et la conception de l'architecture de la partie optique du capteur. Cette dernière a été conçue dans le but de limiter l'entrée directe de la lumière provenant de la réflection spéculaire, source importante de bruit en spectroscopie Vis-NIR. De plus, la conception a également été optimisée pour obtenir un système de mesure réellement portable. Tous les composants : batterie, spectrophotomètre, sources lumineuses et microcontrôleur sont contenus dans un même boitier. Au final, le système pèse moins de 800 g et est pleinement adapté à une utilisation au champ (c.-à-d. autonome, résistant aux chocs et aux éclaboussures).

Les figures 1.7 présentent l'évolution de ce capteur ; d'un système contenu dans un sac-à-dos et nécessitant un ordinateur pour fonctionner à un système plus ergonomique tenu par une main.

(a) **(b)** **(c)**

Figure 1.7 – Evolution des prototypes du Spectron™. a) 2006 b) 2007 c) 2008-09.

Campagnes de mesures

Différentes campagnes de mesures ont été conduites au cours de la thèse pour d'une part créer une base de données spectrales et d'autre part, tester en grandeur nature les prototypes. Ces campagnes se sont déroulées en différents lieux : en Languedoc Roussillon (domaine de l'INRA et IFV), en Champagne (domaine privé et CIVC) et en Australie (domaine privé). Différents cépages ont été testés : Syrah, Merlot, Cabernet, Chardonnay, Pinot Noir, Meunier, etc.

Durant les campagnes de mesures, des lots de grappes étaient quotidiennement collectés de la véraison à quelques jours après les vendanges. Les lots étaient prélevés aléatoirement dans les vignes. Ils étaient ensuite ramenés au laboratoire pour être mesurés avec le Spectron™ puis analysés avec les méthodes destructives de référence. Cette base de données spectrales a permis d'étalonner les modèles de prédiction des critères de maturité mais aussi de tester différents prétraitements sur les spectres nécessaires à l'établissement de ces modèles. La figure 1.8 et le tableau 1.2 présentent quelques résultats d'étalonnage.

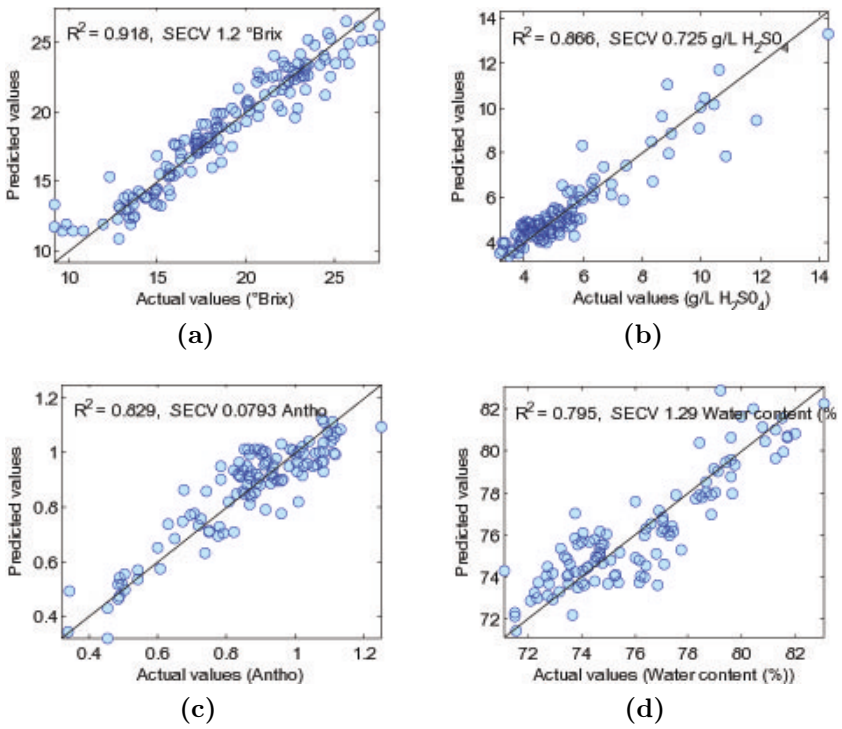
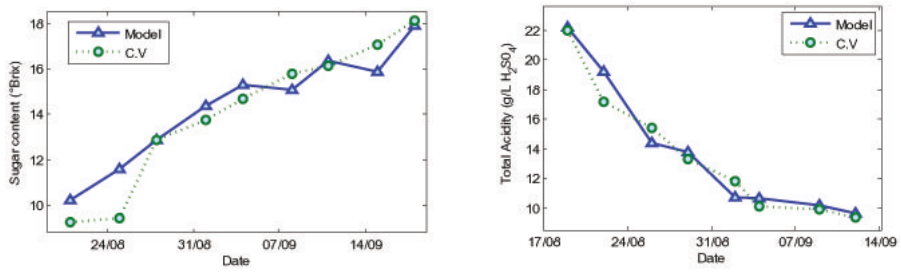


Figure 1.8 – Tests d'étalonnages pour de la Syrah. Valeurs prédites après validation croisée en fonction des valeurs réelles. a) teneur en sucre b) acidité totale c) teneur en anthocyanes d) teneur en eau.

En plus de cette phase d'étalonnage, des séries de mesures de validation étaient réalisées dans les vignes. Entre 150 et 200 grappes étaient directement mesurées sur le cep (c.-à-d. non prélevées) en suivant le même chemin de suivi de maturité que celui du viticulteur. Cette autre base de données spectrales a permis de tester les modèles établis en confrontant les prédictions obtenues avec le Spectron™ aux analyses dites "classiques" faites par le viticulteur. La figure 1.9 présente ces comparaisons pour le suivi de la teneur en sucre et de l'acidité totale sur du Pinot Noir en Champagne.

Variétés	SECV	R^2
Syrah (Australie)	0.92	0.93
Cabernet (Australie)	1.42	0.90
Pinot Noir (Champagne)	1.07	0.88
Meunier (Champagne)	1.11	0.83
cépage rouge (Modèle générique)	1.12	0.95
Chardonnay (Champagne)	1.20	0.83

Table 1.2 – Critères de qualité des étalonnages obtenus sur différents cépages pour la teneur en sucre (° Brix).



(a) **(b)**

Figure 1.9 – Suivi de la teneur en sucre (a) et de l'acidité (b) sur du Pinot Noir (Champagne - 2008). Model : valeurs prédites. C.V : valeurs chimiques de références.

Les résultats de ces campagnes de mesures ont été présentés au symposium international "Frutic Chile 2009" [41].

Le projet Spectron™ a obtenu la médaille d'or du Palmarès de l'Innovation Sitevi 2009.

1.7 Problème opérationnel

Comme expliqué précédemment, les cinétiques de maturité ont une grande importance en viticulture. Ils fournissent de nombreuses informations au viticulteur qu'il associe à ses connaissances pour gérer au mieux ses parcelles (Voir chapitre 1.3.4). Néanmoins, si le viticulteur souhaite réaliser une projection dans le temps de ces cinétiques, il doit se fier en bonne partie à sa logique et à son intuition. En effet, il n'existe pas à l'heure actuelle de ressource permettant de l'aider dans son raisonnement.

Le Spectron™, quant à lui, permet de fournir des mesures qui sont :
- répétées dans le temps,
- en temps réel,
- et multivariées,

- teneur en sucre ($°Brix$, TAP, etc.),
- acidité totale ($g.l^{-1}H_2SO_4$),
- teneur en anthocyanes (mg/l),
- teneur en eau (%).

Partant de ce constat, l'objectif de ce travail de thèse vise à développer un outil capable de tracer la courbe d'évolution *la plus probable* de certains critères de maturité, à partir des premiers points de suivi, tout en intégrant les connaissances du viticulteur. En effet, la construction de ces *cinétiques prédictives* rendrait moins subjective l'interprétation des données de suivi et permettrait ainsi d'aboutir à des réponses optimales quant aux questions du viticulteur sur la gestion de ses parcelles.

Néanmoins, les données dont dispose le viticulteur sont de nature très différente. Ses connaissances peuvent être des connaissances académiques, son expérience ou bien encore des hypothèses (p. ex. les prévisions météorologiques). Les informations issues du suivi peuvent être la mesure d'un critère de maturité (p. ex. les prédictions de teneur en sucre obtenues grâce au Spectron^{TM}) ou bien des informations factuelles (p. ex. des observations sur l'état sanitaire de la parcelle).

Cet outil devra donc être capable de gérer simultanément plusieurs sources d'informations de natures différentes :
- données du Spectron^{TM},
- connaissances propres,
- hypothèses,
- et informations factuelles recueillies par le viticulteur.

Le terme *connaissances expertes*, dans la suite du manuscrit, fera appel à l'ensemble des données dont dispose le viticulteur en dehors des mesures analytiques de suivi, c'est-à-dire les *connaissances académiques*, l'*expérience* ou bien encore les *hypothèses*.

Afin de permettre une bonne coopération connaissances expertes - mesures de suivi, un tel outil a pour obligation d'être compatible avec le *langage* du viticulteur. Il doit pour cela être facilement interprétable. Cette volonté oblige à utiliser des modèles permettant de traduire la réalité de manière satisfaisante.

De plus, toutes les informations (mesures, connaissances expertes) présentent des incertitudes de natures différentes. Ces sources d'incertitude sont, par exemple, dues :
- au processus de mesure en lui même, le Spectron^{TM},
- au produit mesuré, le raisin,
- à l'échantillonnage de la parcelle,
- aux connaissances expertes du viticulteur.

Cet outil doit être capable de prendre en compte les incertitudes entachant ces informations et les propager sur la *courbe résultat*.

Enfin, le viticulteur doit pouvoir "jouer" différents scénarios pour confronter certaines hypothèses, par exemple temps sec *versus* temps humide ou bien l'influence d'une teneur

cible en anthocyanes sur les autres critères de maturité. Il faut donc pouvoir contraindre les cinétiques à suivre ces scénarios.

L'ensemble de ces problèmes peut se résumer en : "comment construire :

1. **des cinétiques prédictives**, c'est-à-dire la courbe d'évolution d'un critère de maturité dans le temps à partir des premiers points de mesure,
2. en utilisant **les mesures fournies par le Spectron™** qui sont :
   - répétées dans le temps et dans l'espace,
   - multivariées,
   - incertaines,
3. en tenant compte des **connaissances expertes du viticulteur** possédant une incertitude
4. le tout devant être facilement :
   - **exploitable**
   - **interprétable**
5. et capable de **suivre des scénarios**".

La figure 1.10 présente de manière schématisée le problème : modéliser la cinétique d'un critère de maturité, Y_i pour $i \in 1, ..., n$, à partir des premières mesures du Spectron™ et en utilisant les connaissances expertes du viticulteur.

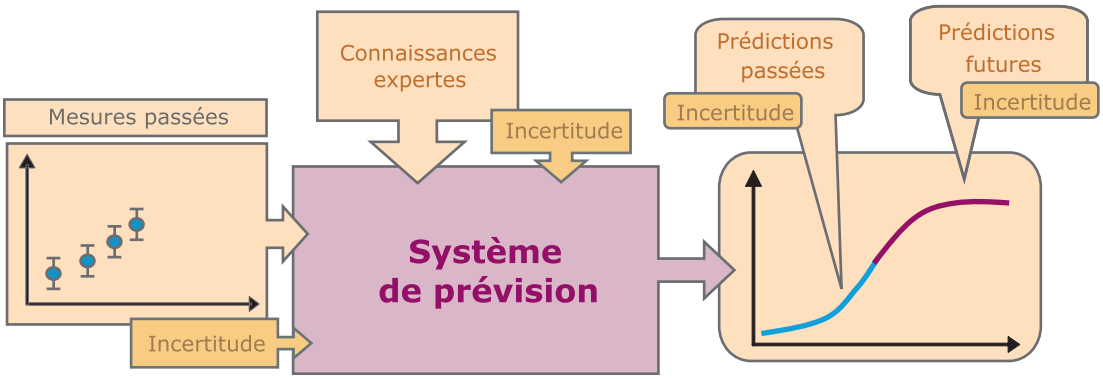


Figure 1.10 – Enjeu : modéliser la cinétique prédictive d'un critère de maturité.

1.8 Organisation du mémoire

Après ce premier chapitre qui a permis de décrire le contexte et le problème opérationnel dans lequel se situe ce travail de thèse, la suite du manuscrit sera structurée en six parties qui constituent autant de chapitres.

Le chapitre II présentera l'analyse de la problématique, à savoir : la modélisation de cinétiques prédictives, à partir de mesures et de connaissances expertes, tout en tenant compte de leur incertitude respective. Une revue bibliographique ciblée sur les différentes questions soulevées par cette analyse sera exposée au chapitre III. Cet état de l'art nous conduira à présenter une proposition scientifique dans le chapitre IV.

Cette approche méthodologique sera ensuite appliquée à la construction de cinétiques prédictives de teneur en sucre et d'acidité totale. Les données utilisées ainsi que l'implémentation de la méthode seront présentées au travers du chapitre V. Ce chapitre détaillera également les modèles utilisés pour la construction des cinétiques ainsi que le système employé pour extraire la valeur des paramètres à partir des connaissances expertes du viticulteur. Le chapitre VI sera dédié à la présentation des résultats obtenus grâce à cette méthode. Ce chapitre intégrera également les discussions générées par les résultats et portera sur :

- la performance du système en général
- la mise en évidence de la nécessité d'utiliser conjointement les mesures et les connaissances expertes du viticulteur,
- l'influence de l'incertitude des connaissances expertes sur les cinétiques prédictives
- l'influence de l'incertitude des connaissances expertes dans la construction d'une bande de confiance autour de la courbe estimée.

Le chapitre VII exposera les conclusions et les perspectives de ce travail de thèse. Il soulignera la validité de la méthodologie proposée ainsi que ses limites. Au travers de ces limites, différentes perspectives seront présentées afin de faire évoluer cette méthodologie.

Chapitre 2

Analyse de la problématique

Sommaire

2.1 Comment construire de telles cinétiques ?

La cinétique d'évolution du critère de maturité correspond, d'un point de vue mathématique, à une courbe $Y_t = \mathcal{M}(X_t)$. Le modèle exact de cette courbe, $\mathcal{M}(.)$, n'a pas forcément de forme analytique et est de toute façon inconnu. Un modèle théorique, $\mathcal{M}^*(.)$, destiné à approcher de manière satisfaisante $\mathcal{M}(.)$ doit être construit. La construction de ce modèle conceptuel peut être guidée par des considérations théoriques relatives au phénomène étudié (p. ex. connaissance des mécanismes relatifs à la biochimie et la physiologie de la vigne). La construction peut également être guidée de manière plus empirique par l'allure générale du phénomène étudié (p. ex. évolution des données de suivi maturité disponibles) [49].

De manière générale, le modèle $\mathcal{M}^*(.)$ peut donc être construit de deux façons :
- à l'aide de modèles conceptuels basés sur les connaissances détaillées des phénomènes,
- à l'aide de modèles empiriques construits à partir des relations structurelles entre les variables d'entrée et de sortie du modèle.

Les mécanismes mis en jeu dans l'évolution des critères de maturité sont très complexes. Les teneurs finales dépendent de nombreux critères : particularité génétique, équilibre hormonal, conditions climatiques, mode de conduite, etc [8, 9, 13, 20, 75]. Il est donc illusoire de vouloir construire un modèle théorique fin de l'évolution de ces différents critères de maturité. De plus, même si la théorie fournissait une bonne représentation de ces phénomènes, leur traduction sous forme de modèles conduirait à des fonctions dépendantes de très nombreux paramètres, difficilement manipulables. La solution analytique de ces problèmes serait alors difficile à obtenir, voire impossible à établir. Cette constatation justifie l'emploi de modèles empiriques pour modéliser la cinétique d'évolution de ces critères de maturité. Une attention particulière doit toutefois être accordée aux modèles qui permettent une certaine interprétation, plutôt qu'un ajustement idéal du modèle aux données [49].

Dans la famille des modèles empiriques, il est possible de distinguer deux groupes : les modèles paramétriques et les modèles non-paramétriques. Le premiers groupe, les modèles paramétriques, sont des fonctions explicites, notées $f_\theta(.)$. Ils sont caractérisés par un certain nombre de constantes ou paramètres, regroupés au sein d'un vecteur θ. Le second groupe, les modèles non paramétriques ou boîtes noires, ne nécessite aucune hypothèse sur la forme de lien entre les variables d'entrée et de sortie du modèle. Il s'agit d'une approche *a*-théorique qui peut aboutir à une représentation satisfaisante du phénomène étudié mais qui ne possède pas de forme analytique. Ces modèles sont donc non-interprétables. Il est possible de citer la régression Kernel, les réseaux de neurones, etc.

Pour l'une des raisons édictée dans le problème opérationnel : l'interprétabilité ; les modèles non paramétriques ne peuvent être envisagés. Seuls des modèles construits à partir de fonctions explicites peuvent être utilisés pour bâtir ces cinétiques prédictives. Les paramètres de ces fonctions n'ont pas forcément de sens précis au regard des théories du phénomène étudié mais permettent de traduire la réalité en terme concret. Par

exemple, le modèle de croissance de la matière sèche de talles de blé est le suivant : $f(x) = \theta_1 \cdot exp(\theta_2 \cdot x)$. Le paramètre θ_2 ne permet pas d'interprétation en terme de physiologie du développement du blé mais peut être assimilé à un taux de croissance [49].

La cinétique des critères de maturité est donc modélisée/construite en utilisant une ou plusieurs fonctions explicites, $f_\theta(.)$, pourvues de paramètres spécifiques, $\{\theta\}$. Les fonctions retenues ont un nombre de paramètres intentionnellement réduit (principe de parcimonie et interprétabilité). Ce problème revient alors à développer une méthode qui permet d'ajuster les paramètres d'une fonction à des jeux d'informations (données) entachées d'incertitude : les mesures de suivi et/ou les connaissances expertes du viticulteur.

En résumé, l'objectif est de développer une méthode capable de fournir une estimation des paramètres, $\{\theta\}$, de la fonction, $f_\theta(.)$, à partir de l'ensemble d'informations imprécises et qui n'ont pas le même cadre de représentation. La qualité de l'estimation des paramètres dépend de manière intrinsèque de la quantité et de la qualité des informations disponibles lors de la procédure d'estimation des paramètres de la fonction.

Comment prendre en compte l'**imprécision des informations**, lors de l'ajustement des paramètres, est **la première question soulevée** par l'analyse de la problématique.

2.2 Ajustement des paramètres

Prises indépendamment, les mesures de suivi et les connaissances expertes du viticulteur permettent l'obtention de valeur probables des paramètres : les mesures en utilisant les techniques d'ajustement de courbe, les connaissances expertes en utilisant celles des systèmes à base de connaissance.

Même si le problème opérationnel impose d'utiliser conjointement ces deux sources d'information, il est nécessaire dans un premier temps d'analyser rapidement et indépendamment ces méthodes d'ajustement. Cette dichotomie permet de mettre en évidence leur limite, étape nécessaire pour construire par la suite une méthode permettant d'utiliser simultanément les deux sources d'information.

2.2.1 Ajustement des paramètres par rapport aux mesures de suivi du critère de maturité

Le problème d'ajustement des paramètres par rapport aux mesures de suivi peut être résolu grâce aux techniques d'ajustement de courbe, plus connues sous le terme anglais de *curve fitting*. Ces techniques permettent d'estimer la valeur des paramètres, $\{\theta\}$, d'une fonction, $f_\theta(.)$, (ici le modèle de la cinétique du critère de maturité) en se basant sur la valeur des observations (ici la mesure du critère de maturité).

La construction d'une "boîte de *fitting*" qui accepterait en entrée les valeurs de suivi et qui fournirait en sortie la valeur la plus probable des paramètres permettrait de répondre à cet objectif. Avec ce type de données, le cadre de représentation des entrées et sorties

de cette boîte sont de type probabiliste. Il est possible de représenter cette proposition par la figure 2.1.

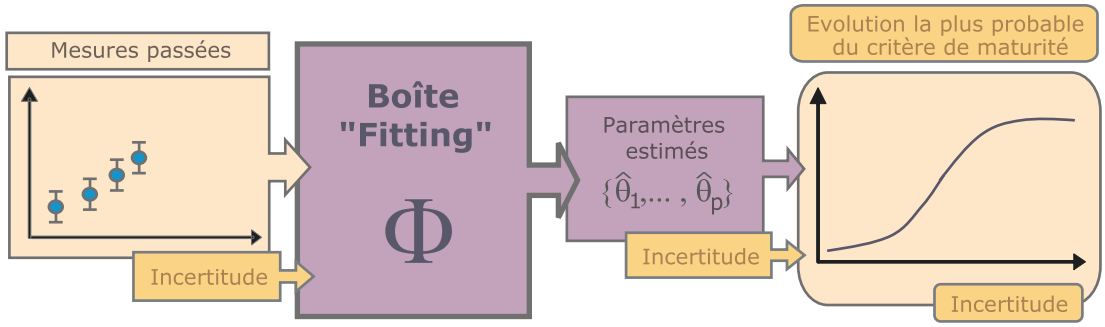


Figure 2.1 – Ajustement des paramètres par rapport aux mesures de suivi.

Concepts de base de l'ajustement de courbe

Soit une série de mesures, y_i, qui a été effectuée pour x_i, où $i = 1$ à n. Afin de relier ces observations $\{y_i\}$, également appelées réponses, à la variable explicative $\{x_i\}$, les mathématiques fournissent une méthode : l'ajustement de courbe. S'il existe un modèle ou une fonction $f_\theta(.)$, qui fournit une image acceptable du phénomène étudié, les outils qui permettront d'ajuster les paramètres, $\{\theta\}$, de cette fonction aux données récoltées, y_i, sont ceux de la régression. L'expression générale peut s'écrire de la manière suivante : $y_i = f(x_i, \theta_p) + \varepsilon_i$ où ε_i symbolise l'erreur et θ_p les p paramètres du modèle[1].

Différentes méthodes d'ajustement, associées à un estimateur, peuvent être envisagées pour résoudre le problème. Le choix de la méthode sera fonction des hypothèses pouvant être émises sur ces erreurs [49, 85]. Dans tous les cas, l'estimateur retenu doit être robuste, c'est-à-dire qu'il doit être résistant aux valeurs rares et/ou extrêmes [85, 93]. Il doit également avoir le biais le plus faible possible et être asymptotiquement non biaisé, c'est-à-dire que son biais tend vers 0 lorsque la taille de l'échantillon tend vers l'infini [85, 93].

L'objectif d'un ajustement se traduit par la volonté de trouver la plus faible erreur possible entre les réponses du modèle, $\widehat{y_i}$, et les observations, y_i. Les méthodes vont consister à chercher parmi l'ensemble des valeurs possibles des paramètres celles qui minimisent cette erreur [85]. L'estimation des paramètres passe donc par l'optimisation (c.-à-d. la maximisation ou la minimisation) d'une fonction objectif. En effet, chaque estimateur est défini par une fonction objectif qui représente une mesure de l'adéquation entre les observations et les réponses du modèle. En d'autre terme, la fonction objectif sert de critère pour déterminer la meilleure solution du problème.

Pour les modèles non-linéaires, le calcul direct de la valeur numérique des paramètres est généralement impossible puisque l'expression explicite des paramètres n'est pas accessible. Différents algorithmes de recherche existent et peuvent être répartis en deux

1. S'il existe plusieurs variables explicatives, x_i devient x_{ij} où $j = 1$ à m le nombre de variables explicatives (régression multivariée)

familles : les méthodes locales et les méthodes globales. La première famille, les algorithmes de recherche locale permettent de déterminer le jeu de paramètres qui minimise (ou maximise) la fonction objectif dans le voisinage du point de départ, en général spécifié par l'utilisateur (p. ex. simplexe de Nelder-Mead, la méthode de quasi-Newton, Levenberg-Marquardt, etc.). Avec ce genre d'algorithme, si le jeu initial de paramètres ne se situe pas dans le voisinage proche de l'optimum global, il peut fréquemment "tomber" dans un optimum local et ne pas s'en extraire (Voir fig. 2.2). En effet, toute tentative de déplacement dans le voisinage immédiat de cet optimum local n'améliore pas le résultat de la fonction objectif. Dans ce contexte, des techniques d'optimisation dites globales ont été développées. Elles permettent d'éviter le piège des minima (ou maxima) locaux. Il en existe trois grandes catégories : les méthodes dites déterministes, énumératives (dans un espace de recherche fini, elles évaluent la fonction objectif en chaque point de l'espace) et stochastiques (elles évaluent la fonction objectif pour un grand nombre de jeux de paramètres initiaux, choisis aléatoirement dans l'espace de recherche).

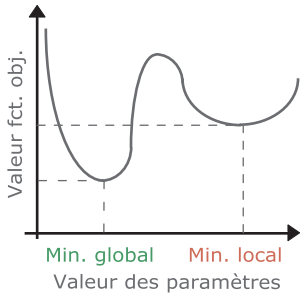


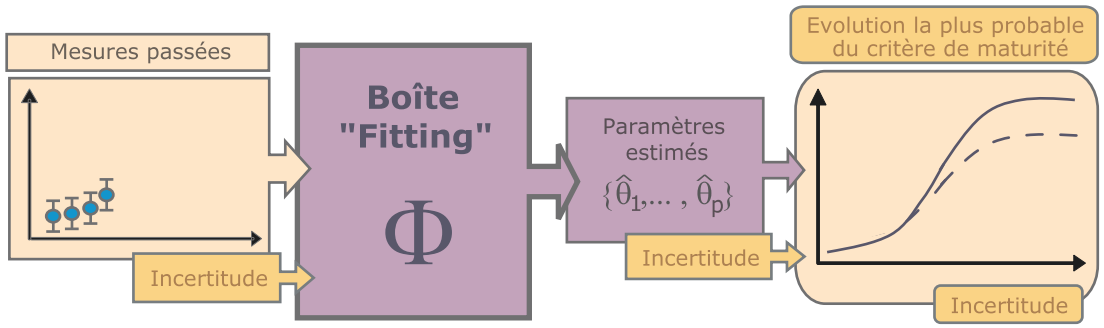
Figure 2.2 – Problème des minima locaux. Evolution de la valeur de la fonction objectif en fonction de la valeur des paramètres.

Après ajustement, l'expression peut s'écrire : $\widehat{y_i} = f(x_i, \widehat{\theta_p})$ où : $\widehat{y_i}$ est la valeur prédite par le modèle pour x_i et $\widehat{\theta_p}$ les p paramètres estimés par la régression.

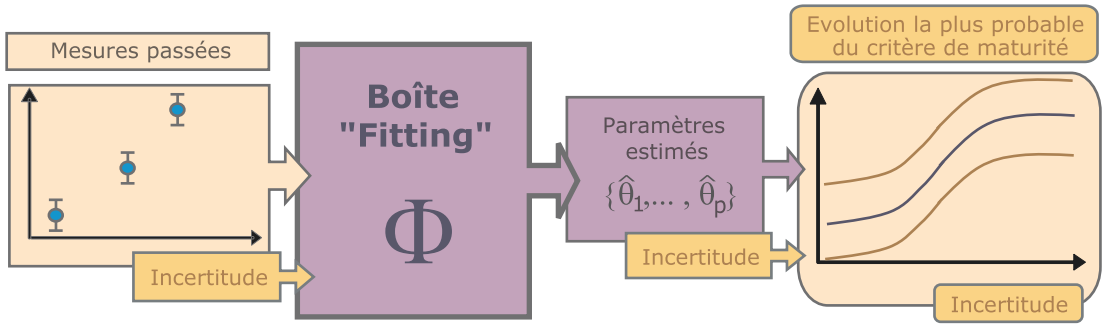
Limites de la "boîte de *fitting*"

La volonté d'obtenir une cinétique prédictive, à partir des premières observations, place l'expérimentateur dans une situation de manque d'information. En effet, les dynamiques de croissance du raisin et de sa maturation sont fortement influencées par des facteurs variétaux et environnementaux [8, 9, 13, 20, 75]. De plus, la précision avec laquelle les paramètres sont estimés par la "boîte de *fitting*" est fonction, certes des mesures faites, mais aussi de leur nombre et surtout de leur répartition sur toute la période de maturation. Les courbes de maturation ne sont que partiellement connues alors que les modèles sont construits pour décrire toute cette période. Ce manque global d'information en début de maturation engendre des difficultés lors de l'estimation des paramètres par les techniques d'ajustement de courbe : tous les paramètres du modèle ne sont plus forcément identifiables, les algorithmes d'optimisation locale convergent vers la solution seulement si le jeu initial des paramètres se situe dans le voisinage proche de l'optimum, l'imprécision des paramètres estimés peut être considérable [70, 85].

Deux exemples de cas limites sont présentés en figure 2.3.



(a)



(b)

Figure 2.3 – Exemple de deux cas limites de la "boîte de *fitting*". a) les points de mesures sont nombreux au début du suivi mais ne permettent pas d'estimer correctement la courbe d'évolution dans son ensemble (trait plein : courbe estimée, trait pointillé : courbe réelle). b) les points de mesures sont répartis sur la quasi totalité de la période de maturation mais leur faible nombre engendre une importante imprécision.

2.2.2 Estimation des valeurs possibles des paramètres par rapport aux connaissances expertes

L'une des principales difficultés avec ce type d'approche est la faible capacité des utilisateurs à retranscrire de manière fine les probabilités. En effet, il est peu envisageable qu'un expert ou un viticulteur puisse fournir les fonctions de distribution de probabilité des paramètres, ou même la valeur la plus probable des paramètres, même si ces derniers permettent de retranscrire la réalité en terme concret. De plus, un expert ou le viticulteur préféra toujours fournir un intervalle plutôt qu'une valeur parce que son savoir est non seulement incertain mais il est imprécis [31]. Cette difficulté peut être contournée par l'utilisation d'un système à base de règles, par exemple, si le viticulteur fait l'hypothèse qu'il fera plus ou moins chaud et humide, alors la vitesse d'accumulation en sucre sera plus ou moins grande.

Un système à base de règles est un *raisonnement* utilisant la *logique*. De manière générale, un *raisonnement* est un processus permettant d'obtenir une nouvelle information à partir d'autres informations existantes et en faisant appel à différentes *lois*.

La traduction, via une "boîte experte", des connaissances expertes du viticulteur en une valeur probable des paramètres permettrait donc de répondre à cet objectif (Voir fig. 2.4). Cette tâche de modélisation serait abordée par la formation d'un ensemble de règles reposant sur la théorie des ensembles flous. Avec ce type d'approche, le cadre de représentation des entrées et sorties de cette boîte est de type possibiliste. Les valeurs des paramètres fournies par ce système ne seront pas les plus probables mais les plus plausibles.

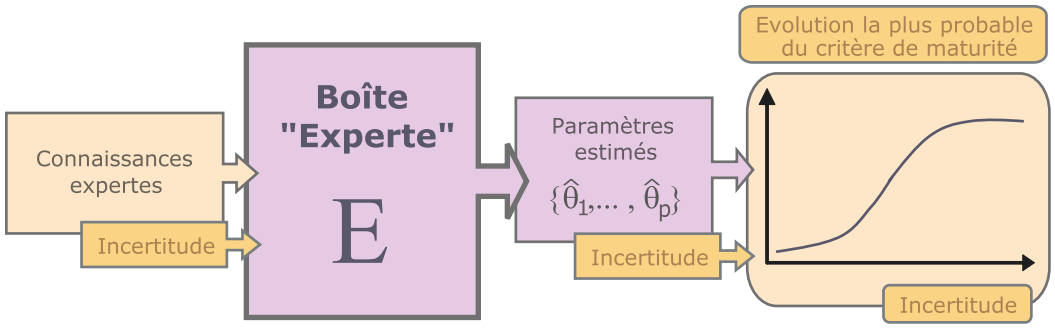


Figure 2.4 – Ajustement des paramètres par rapport aux connaissances expertes.

Concepts de base des systèmes à base de règles floues

Cette "boîte experte" serait basée sur l'utilisation de la logique floue [101]. En effet, la logique floue a été proposée pour modéliser le langage naturel et pour rendre compte du caractère vague des connaissances [102]. Nombre de propositions, comme "la teneur en sucre de cette parcelle est à peu près acceptable", ne peuvent pas être évaluées, comme vraies ou fausses. La logique floue autorise l'utilisation de concepts linguistiques comme "important", "un peu" ou "très bon", lesquels sont modélisés par des ensembles flous.

Les systèmes flous, ou systèmes à base de logique floue, sont donc composés de règles qui font intervenir des ensembles flous. Une règle floue est de la forme : ***Si*** *je rencontre telle situation.* ***Alors*** *j'en tire telle conclusion.* Les ensembles flous sont des ensembles qui permettent de s'affranchir et d'enrichir la notion d'intervalle binaire. Les valeurs sur cet intervalle ne sont plus 1 ou 0 mais sont variables entre 0 et 1. Une classe n'a donc pas forcément des bornes strictes. La logique floue n'est ainsi pas uniquement composée de prédicats vrais ou faux, mais aussi de prédicats de valeurs intermédiaires. Cette représentation se prête tout particulièrement à formaliser le jugement d'un expert. Souvent associée à l'utilisation des sous-ensembles flous, la théorie des possibilités proposée par Zadeh [103] a été essentiellement développée par Dubois et Prade [29].

Supposons, par exemple, qu'un viticulteur doit fournir une estimation de la teneur en sucre acceptable lors des vendanges. Sur la base de son expérience, le viticulteur peut fournir les informations suivantes :

- Il estime que les valeurs se situant entre 22 et 25 ° Brix sont possibles.
- Il n'exclut pas les valeurs comprises entre 20 et 27 ° Brix.

Pour représenter cette information, c'est-à-dire le concept "la teneur en sucre est acceptable", la possibilité des valeurs appartenant à l'intervalle [22 : 25] est normalisée (c.-à-d. égale à 1). Cet intervalle est appelé *le noyau* de la distribution de possibilité. Les valeurs qui sont jugées moins possibles mais non exclues forment *le support* de cette distribution de possibilité. A défaut d'autres informations, il est typiquement supposé que la transition entre les valeurs possibles maximales et minimales soit linéaire. Cet ensemble, A, forme un ensemble flou. La fonction d'appartenance permet de décrire à quel point la teneur en sucre est acceptable pour une teneur x donnée : x appartient à cet ensemble avec un degré d'appartenance $0 \leq \mu_A(x) \leq 1$ (Voir fig. 2.5).

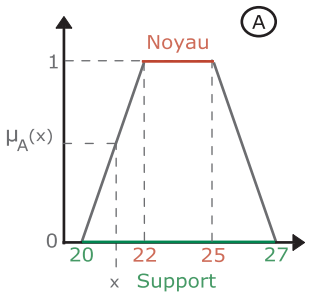


Figure 2.5 – Exemple d'ensemble flou.

Il est ainsi possible de formaliser plusieurs concepts abstraits ou termes linguistiques sur l'univers de la même variable. La figure 2.6 représente les termes "faible" et "élevée" autour de celui "d'acceptable". Les ensembles flous qui représentent les termes linguistiques d'une même variable, définissent le partitionnement de cette variable. Dans le cas de la figure 2.6, x peut être considéré comme une teneur en sucre "acceptable" avec un degré $\mu_A(x)$ et dans le même temps, comme une teneur en sucre "faible" avec un degré $\mu_F(x)$. Ce schéma met en évidence d'une part l'appartenance partielle d'une valeur à un ensemble donné et d'autre part la multi-appartenance d'une valeur à plusieurs ensembles. Cette multi-appartenance permettra d'activer par la suite plusieurs règles (contrairement au cas de la logique classique).

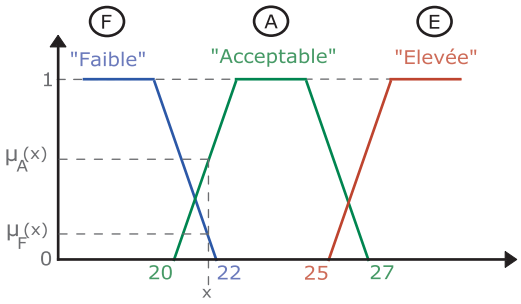


Figure 2.6 – Exemple de partition floue.

La logique floue offre un cadre mathématique permettant des calculs et des raisonnements approchés adaptés au traitement des concepts linguistiques. Le raisonnement approché est basé sur des règles floues ayant la forme suivante : ***Si*** *je rencontre telle situation,* ***Alors*** *j'en tire telle conclusion.* Par exemple, ***Si*** *la température est moyenne,* ***Alors*** *la teneur en sucre est acceptable.* Les concepts linguistiques "moyenne" et "acceptable" sont respectivement représentés par des sous-ensembles flous définis sur l'univers des variables "température" et "teneur en sucre". Le raisonnement approché permet de tirer des conclusions qui tiennent compte du niveau de correspondance entre l'entrée x et la partie condition de la règle appelée prémisse (la situation). Le niveau de correspondance est appelé le *degré de vérité* de la règle (Voir fig 2.7).

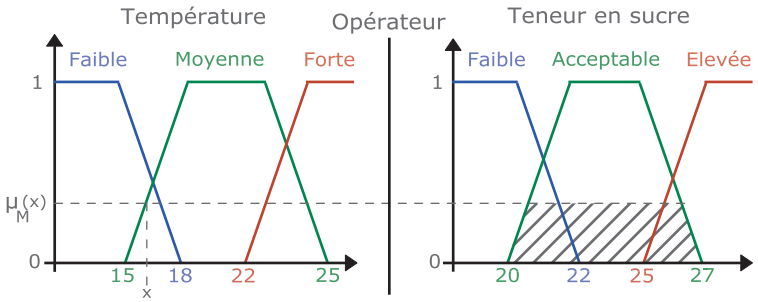


Figure 2.7 – Raisonnement approché. *Si* la température *est* moyenne à un degré $\mu_{MT}(x)$, **alors** la teneur en sucre est acceptable à un degré de vérité $\mu_{FS}(x)$.

Dans le cas multidimensionnel, c'est-à-dire si la prémisse contient plusieurs variables, les degrés d'appartenance sont agrégés (combinés) par un opérateur permettant d'établir une conclusion. Il s'agit des opérateurs de conjonction (le *ET*), les plus utilisés sont le minimum et le produit. (p. ex. Si la teneur en sucre est forte ET si l'acidité totale est faible. Alors le raisin est arrivé à maturation).

De plus, si plusieurs règles existent, l'agrégation des conclusions est faite avec un opérateur de disjonction (le *OU*). Les deux principaux sont le maximum et la somme. C'est-à-dire que les degrés de vérité de la conclusion produits par chacune des règles sont cumulés ou bien seul le maximum est conservé.

Un système d'inférence floue (*SIF*) utilise le raisonnement approché décrit ci-dessus. Il peut présenter plusieurs variables d'entrée et de sortie. Les variables d'entrée apparaissent dans les prémisses des règles et les variables de sortie dans leurs conclusions. Les variables

d'entrée doivent être partitionnées alors que les variables de sortie peuvent être floues (SIF de Mamdani) ou nettes (SIF de Sugeno). Le cœur d'un SIF est le moteur d'inférence floue contenant la base de règles. En pratique, un SIF peut donc être formé de trois blocs (Voir fig. 2.8). Le premier, l'étage de fuzzification transforme les valeurs numériques en degrés d'appartenance aux différents ensembles flous de la partition. Le second bloc est le moteur d'inférence, constitué de l'ensemble des règles. Enfin, un étage de défuzzification permet d'inférer en sortie une valeur nette ou un nombre flou à partir du résultat de l'agrégation des règles. Les opérateurs seront différents en fonction du type de sortie souhaité. De plus amples informations sont disponibles dans [40].

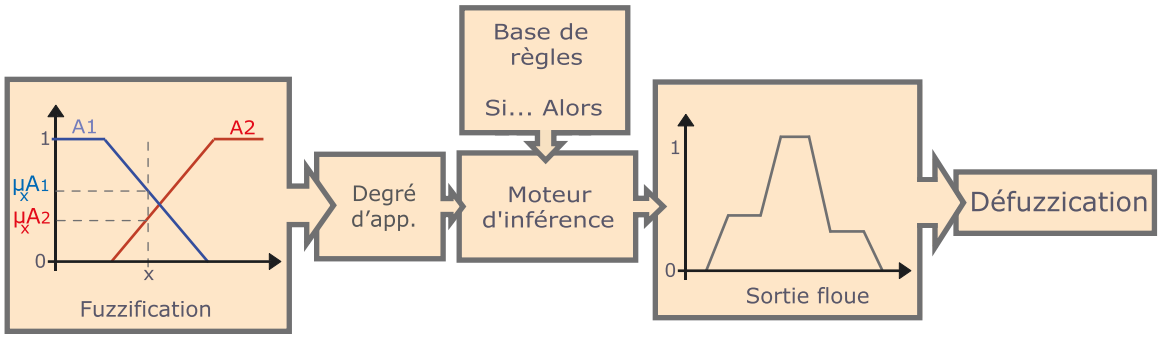


Figure 2.8 – Système d'inférence floue.

Limites de la "boîte experte"

L'estimation des paramètres à partir des connaissances expertes du viticulteur, fondée sur un système à base de règles, représente des conditions standard qu'il faut pouvoir ajuster aux conditions réelles (Voir fig. 2.9).

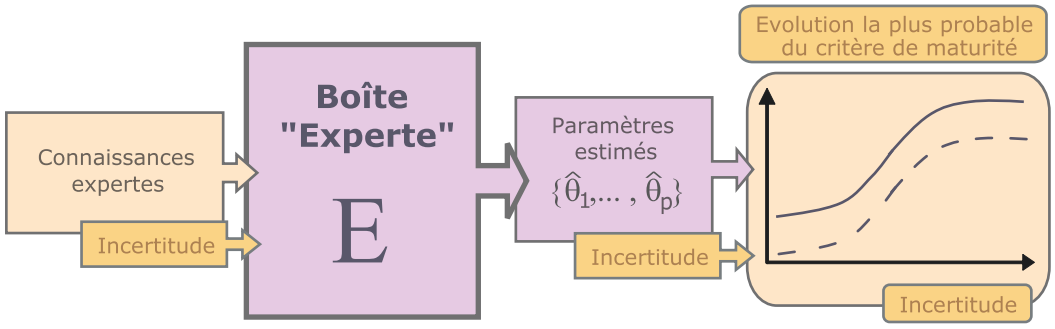


Figure 2.9 – Exemple de cas limite de la "boîte experte". Les connaissances expertes permettent d'estimer les paramètres de la courbe mais avec un biais (trait plein : courbe estimée, trait pointillé : courbe réelle).

2.2.3 Utilisation conjointe des deux sources d'information pour l'ajustement des paramètres

Il est possible de distinguer différentes approches pour utiliser simultanément les mesures de suivi et connaissances expertes du viticulteur :

1. Approches en "parallèle" :
   - L'ensemble des informations est reporté dans un cadre de représentation unique dès le début, c'est-à-dire qu'elles sont fusionnées et utilisées par la même méthode d'ajustement (Voir fig. 2.10.a). Cette fusion est dite de bas niveau.
   - La spécificité de chaque cadre de représentation des informations est préservée le plus longtemps possible avant de fusionner les informations, c'est-à-dire que les deux méthodes d'ajustement travaillent en parallèle et que leurs sorties sont fusionnées (Voir fig. 2.10.b) Cette fusion est dite de haut niveau.

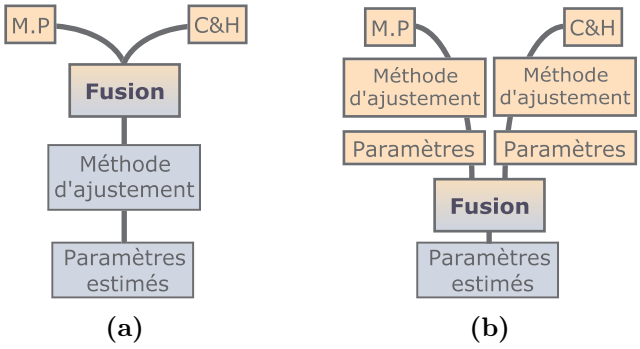


Figure 2.10 – Approches en parallèle. a) fusion de bas niveau b) fusion de haut niveau. M.P : mesures passées, C&H : connaissances expertes.

2. Approches en "série" :
   - Les mesures sont utilisées comme une information *a priori* sur la valeur des paramètres lors de l'ajustement de la fonction par rapport aux connaissances expertes du viticulteur (Voir fig. 2.11.a).
   - Les connaissances expertes du viticulteur sont utilisées comme une information *a priori* sur la valeur des paramètres lors de l'ajustement de la fonction par rapport aux mesures (Voir fig. 2.11.b).

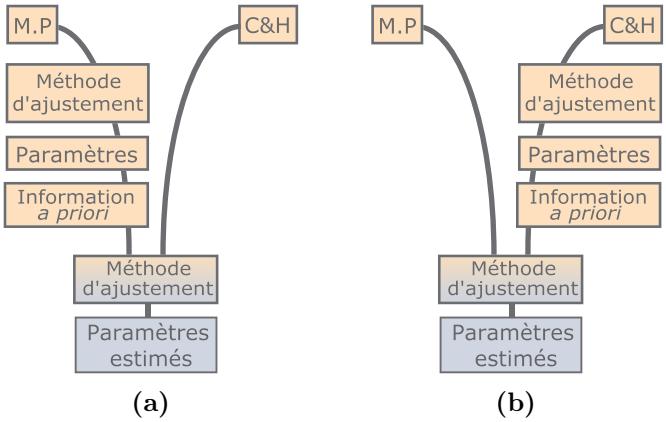


Figure 2.11 – Approches en série. a) mesures utilisées comme information *a priori* sur la valeur des paramètres b) connaissances expertes utilisées comme information *a priori* sur la valeur des paramètres. M.P : mesures passées, C&H : connaissances expertes.

La fusion de bas niveau, qui consiste à concaténer les informations de chaque source afin de les traiter comme une seule et même information, n'est pas envisageable car les

deux sources d'information ne peuvent pas directement être regroupées dans un cadre unique de représentation. La fusion de haut niveau, tout comme l'approche en série utilisant les mesures comme une information *a priori* sur la valeur des paramètres, ne peuvent pas fournir de résultats rapidement après le début du suivi. En effet, la "boîte de *fitting*" ne peut pas fournir d'estimation des paramètres avec un faible nombre de points (Voir chap. 2.2.1). La fusion de haut niveau pose également un problème lorsqu'il s'agit d'agréger des opinions conflictuelles [31]. Par exemple, l'un des ajustements propose une valeur faible pour l'un des paramètres alors que l'autre propose une valeur forte. Dans ce cas, le consensus fournira une solution ni faible ni forte alors que les méthodes d'ajustements s'accordent à la rejeter. L'utilisation des connaissances expertes du viticulteur en une information *a priori* lors de l'ajustement des paramètres par rapport aux mesures semble donc la mieux adaptée à cette problématique.

Cette approche présente l'avantage de pouvoir aisément traduire un scénario comme un jeu d'hypothèses. Ainsi, en appliquant différents jeux d'hypothèses en entrée de ce système, le viticulteur peut facilement "jouer" et confronter plusieurs scénarios (p. ex temps sec *vs.* humide), ces scénarios étant par la suite pris en compte dans l'ajustement des paramètres par rapport aux mesures. Elle présente néanmoins un problème. Elle nécessite de disposer d'une méthode permettant d'employer le cadre de représentation des possibilités comme une information *a priori* lors de l'ajustement des paramètres (par rapport aux mesures), alors que d'autres informations répondent au cadre de représentation probabiliste : par exemple les mesures ou les historiques des suivis.

Comment employer plusieurs sources d'**information *a priori*, ayant des cadres de représentation différents**, lors de l'ajustement des paramètres est **la deuxième question** soulevée par l'analyse de la problématique.

2.3 Gestion et représentation des incertitudes

De nombreuses sources d'erreur surviennent lors de l'estimation des paramètres :
- le choix de la fonction théorique, $f(.)$, destinée à approcher de manière satisfaisante le modèle exact $\mathcal{M}(.)$ est une première source d'erreur.
- la valeur de ces paramètres est inconnue. Leur estimation est une deuxième source d'erreur.
- Le jeu de données, $\{x_i, y_i\}$ pour $i = 1$ à n, qui sert à ajuster le modèle n'est pas parfait. Les mesures ou observations, $\{y_i\}$ dépendent par exemple de la précision de l'appareil de mesure (dans notre cas le Spectron™), de l'opérateur qui prend la mesure et de l'hétérogénéité du produit mesuré. De même, les variables explicatives $\{x_i\}$ peuvent être affectées par une imprécision. Enfin, les connaissances expertes renferment également des incertitudes. Cette imperfection des informations est une troisième source d'erreur.

L'ensemble de ces sources d'imprécision se propage tout au long du processus d'estimation des paramètres. Elles entraînent des effets nuisibles sur les informations extraites

de l'ajustement (p. ex. imprécision des paramètres, biais). Une infinité de courbes, proches de celle estimée, peut donc raisonnablement correspondre à la cinétique recherchée. Ce faisceau de courbes forme une région particulière du plan. La construction d'une bande de confiance autour de la courbe estimée permet de définir cette région, pour un niveau de confiance donné [69, 81].

La figure 2.12 représente le problème.

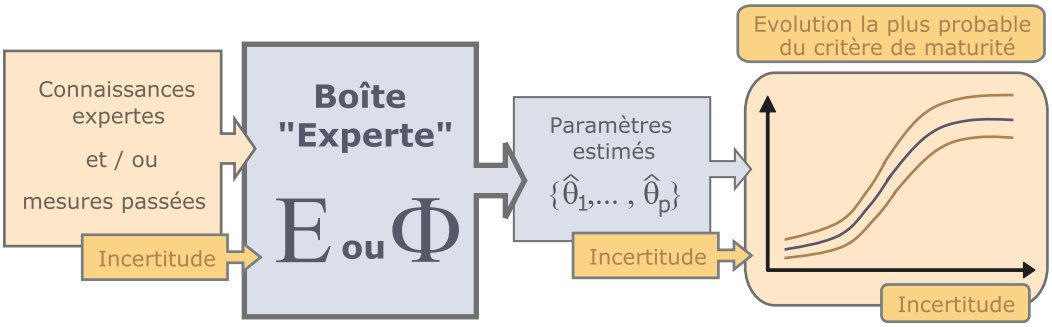


Figure 2.12 – Gestion et représentation des incertitudes.

Comment **construire** une **bande de confiance**, à partir de l'ensemble des sources d'incertitude, est **la troisième et dernière question** soulevée par l'analyse de la problématique.

2.4 Synthèse de la méthode retenue et questions soulevées par l'analyse de la problématique

La méthode retenue, permettant de prédire la cinétique d'évolution la plus probable des critères de maturité, se base sur des fonctions explicites. Les paramètres de ces fonctions sont en partie ajustés par rapport aux mesures de suivi. Comme mises en évidence précédemment, les mesures de suivi représentent une source de connaissance partielle et non suffisante pour l'ajustement des paramètres. Afin de combler ce manque de mesures, une information *a priori* sur la valeur possible des paramètres est incorporée lors de l'ajustement des paramètres. Cette information est issue des connaissances expertes du viticulteur.

Les connaissances expertes du viticulteur sont traduites en une valeur *a priori* des paramètres par un système à base de règles utilisant la logique floue. Le caractère interprétable des règles est des plus importants. En effet, le viticulteur peut aisément comprendre et valider les règles qu'il utilise. Il peut également modifier celles qui ne sont pas en accord avec son expérience. De plus, le système étant ouvert et les règles étant exprimées "dans le langage du viticulteur", il lui est possible d'ajouter de nouvelles règles qui lui sont propres. Enfin, l'utilisation d'un système à base de règles permet naturellement de prendre en compte l'imprécision des connaissances expertes du viticulteur (Voir chap. 2.2.2). La première question soulevée par l'analyse de la problématique (Voir chap. 2.1)

peut donc être reformulée par : "comment prendre en compte l'imprécision des mesures de suivi lors de l'ajustement des paramètres".

Si le viticulteur souhaite jouer un scénario, il peut appliquer différents jeux d'hypothèses en entrée de la "boîte experte" (un scénario = un jeu d'hypothèses défini). Cet avantage peut néanmoins se transformer en un inconvénient. En effet, la méthode favorise, par sa construction, les jeux de paramètres caractérisés par cette information *a priori* sur les paramètres. Ainsi, la recherche est orientée vers le domaine de l'espace des paramètres pressenti *a priori* par le viticulteur.

Un court circuit permettant d'obtenir les cinétiques à partir des seules connaissances expertes, sans forcément renseigner les données de suivi, rendrait le système globalement plus robuste. En effet, le système peut alors continuer à fournir une estimation des paramètres dans une configuration dégradée mais avec une certaine limite (Voir chap. 2.2.2).

L'utilisation de l'incertitude que peut avoir le viticulteur sur ses connaissances doit également être prise en compte. Par exemple, la connaissance de la contrainte hydrique de la parcelle ne sera pas connue avec le même degré de certitude en début ou en fin de maturation. De plus, certaines connaissances peuvent reposer sur des évènement passés. Ces connaissances sont donc connues avec une faible incertitude. A l'inverse, certaines connaissances peuvent être des hypothèses sur la météo portant sur la période de maturation, ces dernières sont donc appréciées avec une grande incertitude.

Enfin, un autre aspect doit également être pris en compte : la fiabilité des sources d'information. Si une source est jugée plus fiable que l'autre, les informations fournies par la première doivent être prioritaires sur celles fournies par la seconde lorsque cette dernière est en désaccord avec les informations de la première source. Il convient alors de faire des hypothèses sur la proportion accordée à chacune des sources. Par exemple, la veille des vendanges un viticulteur peut faire l'hypothèse que le raisin arrivera à 27 $^{\circ}$ Brix alors qu'il est seulement à 20 $^{\circ}$ Brix. Ce cas extrême illustre bien l'incompatibilité des informations et la fiabilité de la source.

La figure 2.13 présente la méthode retenue ainsi que les trois problèmes à résoudre pour mener à bien l'objectif de ces travaux : construire des cinétiques prédictives, à partir de mesures et de connaissances expertes, tout en tenant compte de leurs incertitudes respectives.

- A) prendre en compte l'imprécision des mesures de suivi lors de l'ajustement des paramètres,
- B) employer des informations *a priori*, ayant des cadres de représentation différents, lors de l'ajustement des paramètres,
- C) construire une bande de confiance, pour un risque donné, autour de la courbe estimée en tenant compte de l'ensemble des sources d'incertitude.

Ces différentes questions soulevées par l'analyses de la problématique font l'objet d'une revue bibliographique ciblée exposée dans le chapitre suivant. Cette revue a fait l'objet d'un article en cours de soumission à *Computer and Electronics in Agriculture*.

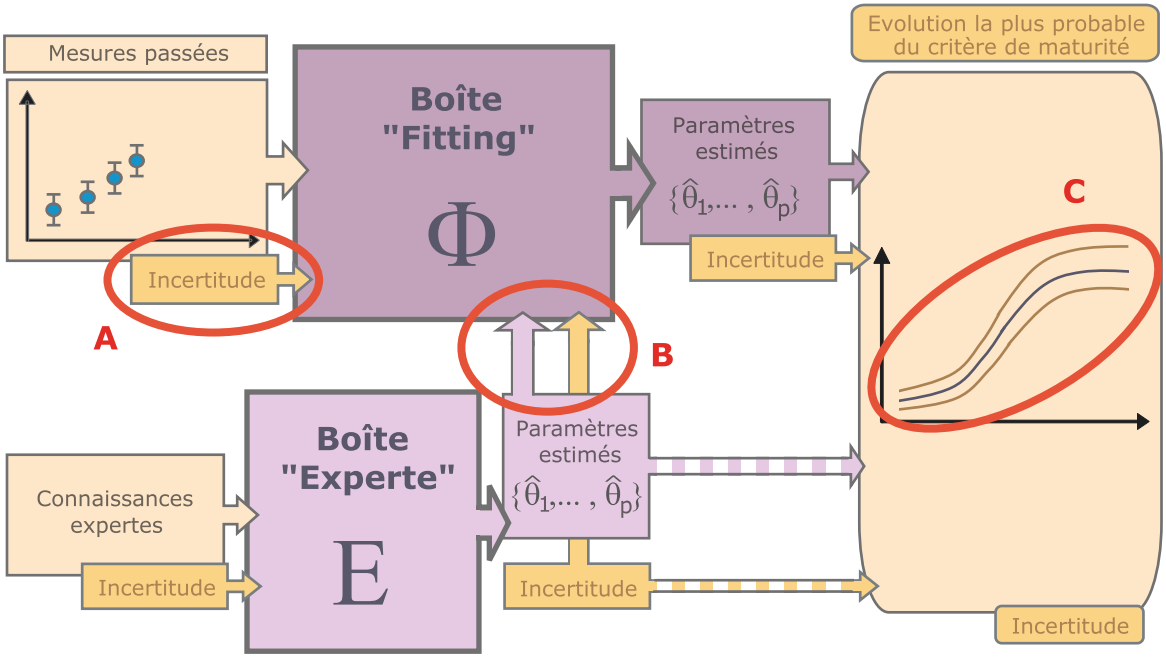


Figure 2.13 – Schéma de méthode retenue et des questions soulevées par l'analyse de la problématique.

Chapitre 3

Etat de l'art

Sommaire

3.1 Introduction

L'analyse de la problématique a soulevé trois questions pour mener à bien l'objectif de ces travaux de thèse :

- Comment prendre en compte l'imprécision des mesures de suivi lors de l'ajustement des paramètres ?
- Comment employer des informations *a priori*, ayant des cadres de représentation différents, lors de ajustement des paramètres ?
- Comment construire une bande de confiance, pour un risque donné, autour de la courbe estimée en tenant compte de l'ensemble des sources d'incertitude ?

Afin de fournir une réponse, une revue bibliographique ciblée sur ces différentes questions est présentée au cours de ce chapitre. Cet état de l'art a fait l'objet d'un article.

3.2 Prise en compte de l'imprécision des mesures

De nombreuses sources d'erreur surviennent lors de l'ajustement d'un modèle (Voir chap. 2.3) :

- erreur due au choix du modèle.
- erreur due à l'estimation des paramètres.
- erreur due aux variables explicatives et aux observations.

Les deux premières sources d'erreur ne sont généralement pas prises en compte dans le processus d'ajustement proprement dit. Elles sont traitées *a posteriori* lors de l'évaluation de la qualité de l'ajustement. De nombreuses informations sont disponibles dans la littérature [52, 62, 69, 85].

La troisième source d'erreur, relative à l'imperfection des données, est prise en compte dans le processus d'ajustement. Différentes méthodes d'ajustement, associées à un estimateur, peuvent être envisagées pour résoudre le problème. Leur choix sera fonction des hypothèses pouvant être émises sur ces erreurs.

De manière générale, l'estimateur retenu doit être robuste, c'est-à-dire qu'il doit être résistant aux valeurs rares et/ou extrêmes [85, 93]. Il doit également avoir le biais le plus faible possible et être asymptotiquement non biaisé (ou convergent), c'est-à-dire que son biais tend vers 0 lorsque la taille de l'échantillon tend vers l'infini [85, 93].

Chaque estimateur est défini par une fonction objectif qui représente une mesure de l'adéquation entre les observations et les réponses du modèle. En d'autres termes, la fonction objectif sert de critère pour déterminer la meilleure solution du problème. Les deux estimateurs les plus couramment utilisés et cités dans la littérature sont l'estimateur du maximum de vraisemblance et celui des moindres carrés.

3.2.1 Prise en compte des erreurs entachant les observations

Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance suppose que les erreurs ε_i, dans le modèle $y_i = f(x_i, \theta_p) + \varepsilon_i$, sont identiquement distribuées. La vraisemblance, L est la probabilité d'observer le jeu de mesures, sachant la valeur des paramètres. Dans la pratique, elle est traduite comme la probabilité d'observer les erreurs, sachant la valeur des paramètres. Plus l'erreur est faible, plus la vraisemblance est importante. La fonction objectif s'écrit :

$$L(\varepsilon|\theta_p) = \prod_{i=1}^{n} P(\varepsilon_i) \tag{3.1}$$

Dans ce cas, l'ajustement consiste à trouver la valeur des paramètres qui maximise cette fonction.

Mais cette fonction ne présente pas d'expression analytique simple, ce qui nécessite de poser certaines hypothèses sur les fonctions de densité de probabilité qui relient les paramètres du modèle aux observations. L'hypothèse de normalité des erreurs permet de simplifier l'expression de la fonction de vraisemblance. Cette hypothèse est largement utilisée dans la pratique. Sous les hypothèses d'indépendance et de normalité des erreurs, les variables aléatoires sont généralement représentées par des lois de distribution gaussienne, centrées sur les prédictions du modèle, et d'écart-type égal à la variabilité des observations (due aux erreurs de mesure du système) par rapport aux prédictions du modèle.

$$P(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_i}^2}} \exp\left[-\frac{1}{2}\frac{(y_i - \widehat{y_i})^2}{\sigma_{\varepsilon_i}^2}\right] \tag{3.2}$$

Dans certains cas, cette variabilité des erreurs de mesure ne peut pas être déterminée expérimentalement : impossibilité de répéter plusieurs fois une mesure sur un échantillon et/ou d'obtenir plusieurs séries de mesures sur un même individu. Un modèle d'erreur est alors proposé. Il constitue une représentation mathématique de la variance des erreurs de mesure.

Estimateur des moindres carrés

Dans le cas particulier où les erreurs sont indépendantes, normalement distribuées, de moyenne nulle et de variance constante ($\forall i, \sigma_{\varepsilon_i}^2 = \sigma_{\varepsilon}^2$, situation d'homoscédasticité), il est possible d'écrire à partir des équations 3.1 et 3.2 :

$$L(\varepsilon|\theta_p) \propto \exp\left[-\frac{1}{2\sigma_{\varepsilon}^2}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2\right] \tag{3.3}$$

Rechercher la valeur des paramètres qui maximise cette fonction (Voir éq. 3.3) revient à rechercher celle qui minimise la somme quadratique des erreurs entre les prédictions du modèle, $\widehat{y_i} = f(x_i, \theta_p)$, et les observations y_i. La fonction objectif peut donc se simplifier et s'écrire sous la forme :

$$S_{OLS}(\theta_p) = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{3.4}$$

$S_{OLS}(\theta_p)$ est l'estimateur des moindres carrés ordinaires (*Ordinary Least Squares*, OLS).

L'utilisation des OLS suppose une variance des erreurs constante. Si cette hypothèse est mise à mal (situation d'hétéroscédasticité), ils fournissent une estimation erronée des paramètres [85]. Il est néanmoins possible de ramener le problème à une approche appartenant à la famille des moindres carrés [2, 72]. A partir des équations 3.1 et 3.2, l'expression de la vraisemblance peut s'écrire :

$$L(\varepsilon|\theta_p) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \widehat{y_i})^T \mathbf{W}(y_i - \widehat{y_i})\right] \tag{3.5}$$

où $\mathbf{W}$ est une matrice de pondération. Comme précédemment, rechercher la valeur des paramètres qui maximise cette fonction (Voir éq. 3.5) revient à rechercher celle qui minimise la somme quadratique des erreurs pondérées. La fonction objectif prend alors la forme :

$$S_{WLS}(\theta_p) = \sum_{i=1}^{n}(y_i - \widehat{y_i})^T \mathbf{W}(y_i - \widehat{y_i}) \tag{3.6}$$

$S_{WLS}(\theta_p)$ est l'estimateur des moindres carrés pondérés (*Weighted Least Square*, WLS).

Les termes de la matrice de pondération, $\mathbf{W}$, sont l'inverse de la variance des erreurs (sous l'hypothèse de non-corrélation des erreurs).

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\sigma^2_{\varepsilon_1}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sigma^2_{\varepsilon_n}} \end{pmatrix} \tag{3.7}$$

Cependant, les valeurs de $\sigma^2_{\varepsilon_i}$ sont généralement inconnues. Les valeurs de la matrice de pondération $\mathbf{W}$ sont donc issues d'une estimation de $\sigma^2_{\varepsilon_i}$. Selon la littérature, plusieurs manières permettent d'estimer ces poids [70]. L'une des plus courantes est l'utilisation de la variance des observations sous différentes formes : $1/\sigma^2_{y_i}$ [24, 54], $\frac{1/\sigma^2_{y_i}}{\sum_{i=1}^{n} 1/\sigma^2_{y_i}}$ [48]. Toutefois, ces estimations peuvent conduire à de mauvais résultats [84], en particulier si le nombre de répétitions réalisées pour une même observation est faible [69]. Une autre approche consiste à envisager les erreurs de mesure comme la seule source de variance. L'étalonnage de l'appareil peut alors fournir cette distribution [88].

D'autres méthodes ont été développées pour prendre en compte l'hétéroscédasticité. Différents auteurs proposent un calcul itératif pour obtenir les pondérations optimales [2, 87, 72]. Il s'agit de la méthode des moindres carrés pondérés itératifs (*Iteratively Reweighted Least Squares*, $IRLS$), dont la trame générale est la suivante :

1. estimation des paramètres par les moindres carrés ordinaires,
2. estimation des poids par une fonction spécifique de variance des erreurs,
3. estimation des paramètres par WLS,
4. ré estimation des poids,

5. répétition des étapes 3 et 4 jusqu'à un critère de convergence défini ou un nombre de cycles fixés.

Les calculs utilisés aux étapes 1 et 2 peuvent différer selon la littérature. De même, il existe différents algorithmes pour estimer les paramètres [14].

Une autre approche est la méthode des moindres carrés étendus (*Extended Least Squares*, ELS) [2, 72, 87, 92]. Le calcul des poids dans l'$IRLS$ est basé sur l'estimation des θ_p de l'itération précédente, les poids dans l'ELS et les θ_p sont calculés simultanément à chaque itération de l'algorithme d'optimisation de la fonction objectif. Un modèle d'erreur représentant la variance des erreurs est introduit dans la fonction d'objectif pour obtenir les pondérations optimales. La dépendance de $\sigma^2_{\varepsilon_i}$ par rapport aux paramètres du modèle, θ_p, et aux paramètres du modèle de bruit, q, est introduit dans un modèle de bruit $M_\sigma(.)$. La fonction d'objectif devient :

$$S_{ELS}(\theta_p, q) = \sum_{i=1}^{n} ln[M_\sigma(\theta_p, q)] + \sum_{i=1}^{n} \left[\frac{(y_i - \widehat{y_i})^2}{M_\sigma(\theta_p, q)}\right] \tag{3.8}$$

Des études basées sur des simulations ont mis en évidence que l'ELS permet d'aboutir à une meilleure estimation des paramètres si la modélisation du bruit affectant les données est correcte [92]. Néanmoins dans certaines situations, les ELS peuvent présenter des problèmes de convergence [48].

Toutes les méthodes présentées ci-dessus tiennent seulement compte des erreurs affectant les observations (y_i). Cela revient à reconnaitre implicitement que les variables explicatives (x_i) ne sont pas entachées d'erreurs. Si cette hypothèse est fausse, l'estimateur des moindres carrés est alors biaisé [85].

3.2.2 Prise en compte des erreurs entachant les observations et les variables explicatives

Les méthodes présentées précédemment minimisent les sommes de carrés d'écarts dans une seule direction (verticale dans le cas d'estimation des y_i à partir des x_i). Si les observations et les variables explicatives sont entachées d'erreurs, les critères d'estimation des paramètres vus précédemment ne peuvent plus être appliqués. Les valeurs des variables y et x sont alors apparentées à leurs vraies valeurs, inconnues. [82].

Deux cas peuvent être envisagés. Les erreurs affectant les observations et les variables explicatives sont soit (supposées) identiques, soit individuellement prises en compte [65]. Différentes approches sont alors utilisées en fonction des hypothèses émises sur les erreurs.

Estimateur du maximum de vraisemblance

Si les erreurs sont supposées additives et indépendantes, la fonction de vraisemblance (Voir éq. 3.1) devient [82] :

$$L(\varepsilon_{xi}\varepsilon_{yi}|\theta_p) = \prod_{i=1}^{n} P(\varepsilon_{xi})P(\varepsilon_{yi}) \tag{3.9}$$

En conservant l'exemple d'erreurs suivant une loi normale de moyenne nulle, l'équation 3.9 s'écrit :

$$L(\varepsilon_{xi}\varepsilon_{yi}|\theta_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{xi}^2\sigma_{yi}^2}} \exp\left[-\frac{1}{2}\left(\left(\frac{(x_i-\widehat{x_i})^2}{\sigma_{xi}^2}\right)+\left(\frac{(y_i-\widehat{y_i})^2}{\sigma_{yi}^2}\right)\right)\right] \tag{3.10}$$

En se basant sur cette approche, plusieurs algorithmes d'optimisation permettant d'estimer les paramètres sont exposés dans la littérature [82, 15].

Estimateur des moindres carrés

Rechercher la valeur des paramètres les plus probables de l'équation 3.10 équivaut à minimiser :

$$S = \sum_{i=1}^{n} \left[(x_i-\widehat{x_i})^T \mathbf{W_x} (x_i-\widehat{x_i}) + (y_i-\widehat{y_i})^T \mathbf{W_y} (y_i-\widehat{y_i})\right] \tag{3.11}$$

Si les variances des erreurs affectant les observations et les variables explicatives sont supposées identiques ($\mathbf{W_x} = \mathbf{W_y}$), l'estimation des paramètres peut alors être réalisée par la méthode des *moindres carrés pondérés perpendiculaires* ou *des moindres distances* [65, 96] (*major axis regression* en anglais). Au lieu de minimiser les carrés des écarts selon un axe vertical, l'objectif est de minimiser les carrés des écarts perpendiculairement à la courbe, impliquant par conséquent les deux variables dans le calcul des résidus (Voir fig. 3.1.a).

Une autre méthode est basée sur celle analogue à la "somme des surfaces triangles" (*Reduced major axis regression* en anglais) définie de la façon suivante : la jonction des points x_i, y_i à la courbe estimée est réalisée par des droites parallèles aux axes [96] (Voir fig. 3.1.b). Néanmoins, les aires sommées et minimisées ne peuvent pas être considérées comme celles de véritables triangles. Ebert et Russel proposent une implémentation de cette méthode [35]. Cette implémentation est limitée aux fonctions monotones. En effet, il existe un seul sens de formation des "triangles" entre les points et la courbe. De plus, la fonction ne doit pas être asymptotique sinon l'un des côtés du triangle tendrait vers l'infini.

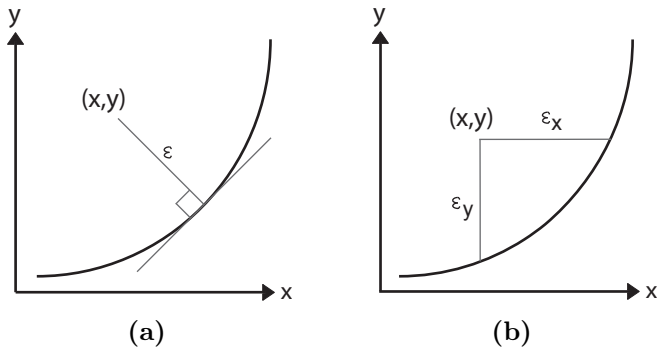


Figure 3.1 – a) distance perpendiculaire b) surfaces "triangles".

Si les variances des erreurs affectant les observations et les variables explicatives sont supposées différentes ($\mathbf{W_x} \neq \mathbf{W_y}$), l'estimation des paramètres peut alors être réalisée par la méthodes des moindres carrés pondérés "des distances obliques" [65].

DeBrauwere propose quant à lui une évolution des moindres carrés pondérés. Il suggère de modifier la matrice de pondération $\mathbf{W}$ (Voir éq. 3.7) [24]. La variance totale de l'erreur est prise en compte en transformant le bruit des variables explicatives en un bruit sur les observations. Ainsi, la nouvelle matrice de pondération, $\mathbf{W_r}$, est toujours composée d'une estimation de l'inverse des variances $\widehat{\sigma}^2_{\varepsilon_i}$ mais qui sont des corrections du premier ordre de la variance de $\widehat{\sigma}^2_{y_i}$:

$$\widehat{\sigma}^2_{\varepsilon_i} = \widehat{\sigma}^2_{y_i} + \sum_{j=1}^{N} \left| \frac{\partial y_i}{\partial x_j} \right|^2 \widehat{\sigma}^2_{x_j} \tag{3.12}$$

En remplaçant $\mathbf{W}$ par $\mathbf{W_r}$ dans l'expression de la fonction objectif S_{WLS} (Voir éq. 3.6), le bruit des variables explicatives est pris en compte lors de l'estimation des paramètres. Cette méthode présente l'avantage d'être applicable même si le bruit des variables explicatives est négligeable. En effet, la matrice $\mathbf{W_r}$ tend vers $\mathbf{W}$ si le bruit des variables explicatives tend vers 0.

3.2.3 Moindres carrés ou maximum de vraisemblance

L'estimateur du maximum de vraisemblance repose sur des bases statistiques et théoriques plus solides. En pratique, c'est l'estimateur des moindres carrés (sous ses différentes formes : ordinaires, pondérés, étendus, etc.) qui est le plus fréquemment mis en œuvre.

Sous les hypothèses d'indépendance et de normalité des erreurs, l'estimateur du maximum de vraisemblance coïncide avec celui des moindres carrés : en cas d'homoscédasticité, maximum de vraisemblance $\Leftrightarrow$ moindres carrés ordinaires et en cas d'hétéroscédasticité, maximum de vraisemblance $\Leftrightarrow$ moindres carrés pondérés [52, 85]. Néanmoins, les moindres carrés sont souvent employés alors que ces hypothèses sont violées. Ce non-respect entraîne une estimation incertaine et biaisée des paramètres. Par exemple, si les erreurs suivent une distribution log-normale, les moindres carrés surestiment la valeur des paramètres [88].

Il n'existe pas de véritable consensus concernant le choix de l'estimateur. Les mérites respectifs des différents estimateurs dépendent de manière complexe et interdépendante de nombreux facteurs : nombre de paramètres à estimer, nombre d'observations, connaissances *a priori* de la variances des erreurs. Le maximum de vraisemblance présente l'avantage, par rapport aux moindres carrés pondérés, de ne pas nécessiter de connaissance *a priori* sur la variance des erreurs afin de procéder à l'optimisation. En effet, un modèle d'erreur est intégré dans la fonction de vraisemblance et attribue un coefficient d'hétéroscédasticité à chaque mesure. De plus, les moindres carrés présentent le défaut de donner beaucoup trop de poids aux observations rares et/ou extrêmes. Ce défaut conduit à des ajustements biaisés [70]. A l'inverse, l'estimateur des moindres carrés nécessite de connaître ou d'estimer la variance des erreurs avant de procéder à l'estimation des paramètres. Si les données sont suffisamment informatives pour émettre des hypothèses

sûres au sujet de la variance des erreurs, les moindres carrés peuvent être préférés à celui du maximum de vraisemblance. Néanmoins, ceci impose des contraintes expérimentales difficilement applicables en agriculture (p. ex. répéter plusieurs fois la mesure sur chaque plant). Enfin, l'estimateur des moindres carrés peut également être préféré si les données disponibles ne sont pas suffisamment informatives pour estimer, en plus des paramètres du modèle, la variance des erreurs ou les paramètres du modèle d'erreur.

En cas d'homoscédasticité ($\forall i, \sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon}$), la méthode des moindres carrés ordinaires joue un rôle central. En cas d'hétéroscédasticité, il existe des estimateurs plus efficaces comme les moindres carrés pondérés. Cependant, dans la pratique, l'expérimentateur peut négliger (consciemment ou inconsciemment) cette hétéroscédasticité et utiliser les moindres carrés ordinaires. En effet, l'expérimentateur est plus souvent familiarisé à cette méthode car elle est plus simple d'utilisation. De plus, les moindres carrés pondérés pose le problème du choix des poids. En partant de ce problème, Shao [86] cherche à savoir si les méthodes statistiques fondées sur les moindres carrés ordinaires sont robustes en cas d'hétéroscédasticité des erreurs. Il arrive à démontrer la convergence de l'estimateur ainsi que celle de la matrice de variance covariance des paramètres (si les erreurs sont de moyenne nulle). La convergence de la matrice de variance covariance des paramètres n'est pas obtenue en utilisant les résidus mais en utilisant la méthode de ré-échantillonnage du *Jackknife*.

Une autre question est de savoir s'il faut prendre en compte les erreurs entachant les observations (y_i) et les variables explicatives (x_i). Legendre & Legendre [60] recommandent la démarche suivante pour choisir une méthode permettant d'estimer les paramètres reliant deux variables aléatoires mesurées avec des erreurs. Si la variance des erreurs sur les y_i est beaucoup plus forte que celle des x_i (c.-à-d. supérieure à trois fois), il est possible d'utiliser les méthodes prenant seulement en compte les erreurs entachant les observations (Voir chap. 3.2.1). Par contre, si les deux variables sont exprimées dans les mêmes unités physiques ou sont sans dimension et que la variance des erreurs est à peu près identique pour les deux variables, il faut utiliser une méthode permettant de prendre en compte les erreurs entachant les observations et les variables explicatives (Voir chap. 3.2.2). De plus, si les variables x et y sont sujettes à des fluctuations aléatoires, cela implique l'utilisation des méthodes dernièrement citées. En résumé, il est possible d'utiliser les méthodes prenant seulement en compte les erreurs entachant les observations si :

- les variables explicatives sont contrôlées (non aléatoires),
- la variance des variables explicatives est très faible par rapport à celle des observations.

Enfin, si les observations et les variables explicatives sont entachées d'erreur et que ces dernières sont corrélées, l'utilisation du maximum de vraisemblance ou de la somme des surfaces triangles n'est pas conseillée [96]. De plus, il est possible de trouver différents algorithmes dans la littérature afin d'estimer les paramètres lorsque les observations et les variables explicatives sont entachées d'erreurs [61, 65, 90].

3.3 Utilisation d'une information *a priori* sur les paramètres

La volonté d'estimer la valeur des paramètres de la cinétique, à partir des premières observations, place l'expérimentateur dans une situation de manque d'informations (Voir chap. 2.2.1). De plus, étant donné la complexité du phénomène étudié, même si les observations sont nombreuses, elles ne constituent pas toujours une source d'information suffisante pour obtenir la cinétique d'évolution la plus probable du critère de maturité (p. ex. influence de la météo, type de cépage, etc. Voir chap. 1.4). L'incorporation d'une information *a priori* sur les paramètres dans la fonction objectif permet de remédier à ce manque d'information.

Différentes sources sont susceptibles de fournir cette information *a priori* sur la valeur des paramètres à estimer (p. ex. historique de la parcelle, avis d'un expert, connaissances du viticulteur, etc.).

Il est peu envisageable qu'un expert puisse fournir les fonctions de distribution de probabilité des paramètres, même si ces derniers permettent de retranscrire la réalité en termes concrets. Cette difficulté peut être contournée par l'utilisation d'un système à base de règles reposant sur la théorie des possibilités [29]. En effet, ce type de système permet de retranscrire de manière formalisée les connaissances recueillies et de fournir une distribution de possibilité (Voir chap. 2.2.2).

De part leurs origines, ces sources informations disposent d'un cadre de représentation différent. Par exemple, les informations issues d'un historique se rattacheront à un cadre probabiliste au travers d'une distribution de probabilité, celles issues d'un système à base de règles se rattacheront à un cadre possibiliste au travers d'une distribution de possibilité.

3.3.1 Utilisation d'une fonction de distribution de probabilité

La méthode bayésienne

La méthode d'estimation bayésienne utilise une information *a priori* concernant le domaine de variation possible des paramètres et leur distribution sur ce domaine [11, 85]. Ces informations sur les paramètres du modèle sont synthétisées au sein d'une fonction de densité de probabilité dite *a priori* (également appelé *prior*). Il s'agit d'évaluer la probabilité de toute proposition en calculant la probabilité qu'elle soit vraie, conditionnée par toute l'information disponible.

Le théorème de Bayes relatif aux distributions conditionnelles permet d'écrire la distribution conjointe $P(\theta_p, \varepsilon)$ de θ et de ε sous deux formes :

$$P(\theta, \varepsilon) = P(\theta|\varepsilon)P(\varepsilon) \; et \; P(\theta, \varepsilon) = P(\varepsilon|\theta)P(\theta) \tag{3.13}$$

Comme P(ε) est indépendante de θ :

$$P(\theta|\varepsilon) \; \propto \; P(\varepsilon|\theta)P(\theta) \tag{3.14}$$

$P(\theta|\varepsilon)$, appelée *posterior*, se lit comme la probabilité de θ, étant donné que ε a été observé. Il s'agit de la distribution de probabilité bayésienne *a posteriori* des paramètres θ. Cette distribution *a posteriori* est la distribution de probabilité conjointe de tous les paramètres. Elle est composée de deux éléments : $P(\varepsilon|\theta)$ et $P(\theta)$. Pour $P(\varepsilon|\theta)$, comme ε est inconnue et que θ est connu, ce terme est donc la vraisemblance $L(\varepsilon|\theta)$. Le *prior*, $P(\theta)$, spécifie les valeurs possibles de θ avant que les mesures soient réalisées et renvoie à la probabilité *a priori* des paramètres. Par exemple, si un paramètre est connu pour être positif, $Pr(\theta \geq 0) = 1$.

La fonction de densité de probabilité dite *a posteriori*, $P(\theta|\varepsilon)$, est donc définie comme le produit de la fonction de vraisemblance $L(\varepsilon|\theta)$ (contenant l'information sur les observations) et de la fonction *a priori* des paramètres $P(\theta)$ (contenant l'information sur les paramètres) [11, 85]. L'estimateur bayésien de θ est celui qui maximise la probabilité *a posteriori* $P(\theta|\varepsilon)$ donc $L(\varepsilon|\theta)P(\theta)$.

Par exemple, si les paramètres suivent une loi normale, $P(\theta)$ s'écrit :

$$P(\theta_p) = \frac{1}{\sqrt{2\pi\sigma_{\theta_p}^2}} \exp\left(-\frac{1}{2}\frac{(\theta_p - \overline{\theta_p})^2}{\sigma_{\theta_p}^2}\right) \tag{3.15}$$

Et si les erreurs aléatoires ε_i sont normalement distribuées autour d'une moyenne nulle et de variance $\sigma_{\varepsilon_i}^2$, alors la distribution *a posteriori* $P(\theta|\varepsilon)$ aura la forme suivante :

$$P(\theta_p|\varepsilon_{\varepsilon_i}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_i}^2}} \frac{1}{\sqrt{2\pi\sigma_{\theta_p}^2}} \exp\left(-\frac{1}{2}\left[\frac{(y_i - \widehat{y_i})^2}{\sigma_{\varepsilon_i}^2} + \frac{(\theta_p - \overline{\theta_p})^2}{\sigma_{\theta_p}^2}\right]\right) \tag{3.16}$$

Les hypothèses émises dans cet exemple reviennent à se placer dans des conditions particulières. Dans cette situation, l'estimation des paramètres est équivalente aux moindres carrés pondérés pénalisés (*Penalized Weighted Least Squares* [72, 3]) qui utilisent une distribution normale des paramètres, de moyenne $\overline{\theta_p}$ et de matrice de covariance $\mathbf{\Sigma}_\theta$. La fonction objectif à minimiser s'écrit alors :

$$S_{WLS_p}(\theta_p) = \sum_{i=1}^{n} (y_i - \widehat{y_i})^T \mathbf{W} (y_i - \widehat{y_i}) + (\theta_p - \mu_p)^T \mathbf{\Sigma}_\theta (\theta_p - \overline{\theta_p}) \tag{3.17}$$

3.3.2 Utilisation d'une distribution de possibilité

Si les informations disponibles sont représentées dans un cadre possibiliste, deux méthodes sont envisageables : transformer la distribution de possibilité en une distribution de probabilité, ce qui est abondamment abordé dans la littérature [28, 77, 78, 89] afin de pouvoir ensuite l'utiliser dans un cadre bayésien dit *classique* ou utiliser directement cette distribution de possibilité dans un cadre bayésien *flou* [38].

Transformation d'une distribution de possibilité en probabilité

Il est possible de transformer une mesure de possibilité, $\pi(.)$, en une mesure de probabilité $p(.)$ en appliquant différentes procédures [78, 89, 77]. Avant de présenter les

méthodes qui permettent cette transformation, il est nécessaire de connaître les conditions autorisant le passage entre possibilité/probabilité ainsi que leur interprétation.

De nombreux principes doivent être respectés lors de la transformation d'une distribution de possibilité en probabilité [77]. La préférence doit être respectée : si un élément x est préféré à un autre élément y selon la distribution de possibilité, cette préférence doit être maintenue dans le cadre probabiliste. Cela se traduit par les relations suivantes :

$$\begin{aligned} p(x) = p(y) &\Leftrightarrow \pi(x) = \pi(y) \\ p(x) > p(y) &\Leftrightarrow \pi(x) > \pi(y) \\ p(x) < p(y) &\Leftrightarrow \pi(x) < \pi(y) \end{aligned} \tag{3.18}$$

La symétrie doit être préservée : la forme générale des données du cadre possibiliste doit être retrouvée dans le cadre probabiliste. Généralement, le respect des préférences (Voir éq. 3.18) entraîne le respect de la symétrie. L'ignorance doit être également conservée : si un état d'ignorance complet apparaît, il doit être traduit. L'ignorance est représentée dans le cadre possibiliste en plaçant tous les éléments de l'univers du discours à 1. Cette ignorance se traduit par une distribution de probabilité uniforme. Enfin, le principe de cohérence possibilité/probabilité (*Possibility/Probability Consistency Principe (PPCP)*) introduit un ensemble d'hypothèses sur les relations entre possibilité et probabilité afin de rendre réalisable la transformation de l'une à l'autre. Ce principe résulte du fait que "ce qui est possible" influence et est influencé par "ce qui est probable" [103]. Formellement, la mesure de cohérence,γ, entre $\pi(.)$ et $p(.)$ est donnée par :

$$\gamma = \sum_{i=1}^{n} \pi(x_i) \cdot p(x_i) \tag{3.19}$$

Ainsi :

- Si $\pi(x) = 0 \ \forall \ x$, alors $\gamma = 0$. Un évènement impossible ne peut pas être probable.
- Si $\pi(x) = 1 \ \forall \ x$, alors $\gamma = 1$. N'importe quelle mesure de probabilité est toujours cohérente avec les mesures de possibilité.

Le principe de cohérence provient du fait que la possibilité est une notion plus faible que la probabilité dans le sens où son affirmation n'est pas aussi forte que celle de la probabilité. En d'autres termes, $\pi(.)$ est cohérente avec $p(.)$ si les informations fournies par $\pi(.)$ sont contenues dans les informations données par $p(.)$ [30]. Le but de cette partie est de présenter de manière non exhaustive les conditions permettant le passage entre possibilité/probabilité ainsi que leur interprétation. De plus amples informations sont disponibles dans la littérature [28, 30, 77, 78, 89, 103].

La méthode de transformation la plus simple est celle du rapport des échelles. Elle consiste seulement à la normalisation des valeurs en divisant par la somme totale des possibilités [77] :

$$p(s_i) = \frac{\pi(s_i)}{\sum_{i=1}^{n} \pi(s_i)} \tag{3.20}$$

La transformée pignistique permet également cette transformation. En pratique si $card(S) = n$ et $\pi(s_i)$ pour $i = 1$ à n avec $\pi(s_1) \geq \pi(s_2) \geq \ldots \geq \pi(s_n)$, la transformée

pignistique est une distribution de probabilité p ordinalement équivalente à $\pi(s_i)$, telle que $p(s_1) \geq p(s_2) \geq \ldots \geq p(s_n)$ pour $i = 1$ à n [28]. Cette transformation s'écrit :

$$p(s_i) = \sum_{j=i}^{n} \frac{\pi(s_k) - \pi(s_{j+1})}{j} \tag{3.21}$$

Cette transformation généralise le principe d'indifférence de Laplace qui dit que tout ce qui est équiplausible doit être équiprobable. En effet, appliquée à une distribution de possibilité uniforme sur un intervalle, elle fournit bien une probabilité uniforme.

D'autres transformations beaucoup plus complexes et donc moins facilement manipulables (Klir's, Moral, approximation par une distribution normale) sont présentées dans [77, 78].

Cadre bayésien avec des données nettes et des *priors* flous

Il est possible d'appliquer le théorème de Bayes avec des données nettes et des *priors* flous [38]. Dans ces circonstances, η_0^* représente l'hyperparamètre flou de la distribution *a priori* noté $\pi(\theta|\eta_0^*)$. La fonction d'appartenance et l'α-coupe correspondantes sont notées $\varphi_{\eta_0^*}(\eta_0)$ et $C(\eta_0^*)_\alpha$, $\alpha \in]0, 1]$. Le *posterior* est noté $g(\theta|\eta_0^*, \varepsilon)$.

Les paramètres de la distribution *a priori* étant *flous*, le *posterior* est une fonction *nette* dépendant d'arguments *nets* et *flous*. Afin de représenter cette valeur, les α-contours $(g_n)_\alpha^L(\theta)$ et $(g_n)_\alpha^U(\theta)$ sont utilisés. Les α-contours sont des courbes de représentation du niveau α. Ils sont obtenus par la minimisation et la maximisation de la fonction "nette".

Les α-contour du *posterior* $g(\theta|\eta_0^*, \varepsilon)$ sont donc reliés aux α-contours de la valeur floue de la distribution *a priori* $\pi(\theta|\eta_0^*)$ par :

$$\begin{aligned} (g_n)_\alpha^L(\theta) &= L(\varepsilon|\theta) \cdot \pi_\alpha^L(\theta) \\ (g_n)_\alpha^U(\theta) &= L(\varepsilon|\theta) \cdot \pi_\alpha^U(\theta) \end{aligned} \tag{3.22}$$

où :

$$\begin{aligned} \pi_\alpha^L(\theta) &= \min_{\eta_0 \in C(\eta_0^*)_\alpha} \pi(\theta|\eta_0) \\ \pi_\alpha^U(\theta) &= \max_{\eta_0 \in C(\eta_0^*)_\alpha} \pi(\theta|\eta_0) \end{aligned} \tag{3.23}$$

3.3.3 Conclusion sur l'utilisation d'une information *a priori* sur les paramètres à estimer

L'estimation bayésienne des paramètres est appliquée avec succès dans de nombreux domaines [3, 18, 36, 64, 71]. S'il existe une information *a priori* fiable sur la distribution des paramètres, elle permet d'améliorer la précision de leur estimation [85]. De plus, ces méthodes sont moins exposées aux problèmes de non-identifiabilité numérique.

Si la fonction de distribution est uniforme, cela revient à borner l'espace de recherche des paramètres dans un intervalle défini. La recherche des valeurs dans cet intervalle est

alors un problème d'optimisation sous contrainte. Il peut être résolu par exemple par l'utilisation des multiplicateurs de Lagrange. Des solutions à ce type de problème spécifique peuvent être trouvées dans [42].

Cependant, la méthode bayésienne favorise, de par sa construction, les jeux de paramètres caractérisés par une forte densité de probabilité *a priori*. Les valeurs des paramètres sont explicitement probabilisées. Ainsi, la recherche est orientée vers le domaine de l'espace des paramètres pressenti *a priori* comme le plus probable. Le choix du *prior* influence donc de manière prédominante les résultats du processus d'estimation. La détermination de cette loi est donc une étape très importante dans la mise en œuvre de cette méthode. Il existe un grand nombre de choix possibles de la distribution *a priori* des paramètres et il n'existe pas une façon unique de la choisir [11, 85]. En effet, il est généralement difficile de proposer une forme exacte des distributions de $\{\theta_p\}$ car elles n'ont pas de réalité propre et correspondent à une paramétrisation de la loi décrivant les paramètres. Cette loi est donc un moyen de résumer l'information disponible sur les paramètres, ainsi que son incertitude.

L'utilisation conjointe des mesures au champ (statistiques) et d'informations subjectives imprécises (possibilistes) permet de construire des cinétiques prédictives. Cette approche pose le problème de disposer de plusieurs cadres de représentation des informations. Avec une approche hybride du type de Fruehwirth-Schnatter [38], peu claire pour un non spécialiste, où une partie des données est possibiliste et l'autre probabiliste, le résultat est un intervalle flou aléatoire. La question est alors d'extraire l'information pertinente et exploitable. De plus, cette démarche demande un certain effort d'implémentation. Il est donc légitime d'utiliser les passages permis entre possibilité et probabilité afin de disposer d'un cadre de représentation unique lors de l'ajustement.

Cas particulier des historiques de suivi

Les historiques de suivi d'une parcelle pourraient être traités d'un point de vue statistique. Dans la pratique, hors cadre d'expérimentation, il est rare en viticulture que le temps et les ressources nécessaires à la construction d'un suivi de qualité, au sens statistique du terme, soient réunis. En effet, la collecte de mesures de suivi est coûteuse (non automatisée, intervention manuelle indispensable). Le nombre de mesures réalisées par suivi au cours d'une même année est donc limité. Cela rend l'utilisation des suivis, d'un point de vue statistique, assez limité pour évaluer la valeur probable des paramètres. Les informations issues d'un historique doivent donc être forcément complétées par d'autres informations sur la parcelle.

Néanmoins, les connaissances du viticulteur peuvent être considérées comme un échantillon représentatif de l'ensemble des situations à modéliser afin de construire une distribution *a priori* informative. L'historique peut alors se traduire directement comme une connaissance du viticulteur (par ex. cette parcelle est plutôt précoce). Cette approche permet d'avoir un cadre unique de représentation de l'ensemble des informations.

3.4 Construction d'une bande de confiance

L'ensemble des sources d'imprécision entraîne des effets nuisibles sur les informations extraites de l'ajustement (imprécision des paramètres, biais). En effet, la variance l'erreur d'ajustement dépend d'une part de la variabilité intrinsèque des observations y_i et d'autre part de l'imprécision de l'estimation des θ_p. L'imprécision de l'estimation des θ_p dépend pour l'essentiel de la taille de l'échantillon. Cette dernière peut être réduite contrairement à la variabilité des observations.

Par conséquent, une infinité de courbes, proches de celle estimée, peut raisonnablement correspondre à la cinétique recherchée. Ce faisceau de courbes forme une région particulière du plan. La construction d'une bande de confiance autour de la courbe estimée permet de définir cette région, pour un niveau de confiance donné [69, 81] (p. ex. la cinétique a 95 % de chance d'être dans cette zone).

Le problème revient donc à trouver pour un niveau de confiance α, la région du plan définie par la limite inférieure $y_L(x)$ et supérieure $y_U(x)$ telle que :

$$\forall x, \quad y_L(x)_\alpha \leq f(x, \widehat{\theta_p}) \leq y_U(x)_\alpha \tag{3.24}$$

Le calcul direct de la bande de confiance est généralement impossible puisque l'expression explicite de la propagation des erreurs sur des paramètres n'est pas accessible [85]. En raison de la complexité de cette évaluation, ces informations pertinentes sur l'interprétation des résultats de l'ajustement sont souvent omises.

Néanmoins, différentes méthodes permettent de construire une bande de confiance globale et continue autour de la courbe estimée : celles basées sur la linéarisation [1, 81, 85] ou celles basées sur le calcul d'un hypervolume de confiance sur les paramètres [17, 19] .

3.4.1 Estimation de la bande de confiance par la méthode de linéarisation

Les méthodes de linéarisation supposent que le modèle est convenablement approximé par une fonction linéaire au niveau des paramètres estimés $\widehat{\theta_p}$ [85, 81, 21]. De manière générale, la bande de confiance basée sur la méthode de linéarisation est donnée par :

$$BC_L = \left\{ (x, \widehat{y}) \ / \ \widehat{y} = f(x, \widehat{\theta_p}) \pm h(x, \widehat{\theta_p}) \right\} \tag{3.25}$$

où :

$$h(x, \widehat{\theta_p}) = \sqrt{pF_{1-\alpha}(p, n-p)} \ \sqrt{\frac{s^2}{m} \ \mathbf{F_x}^T \mathbf{Cov}(\widehat{\theta_p}) \mathbf{F_x}} \tag{3.26}$$

avec :

$$\mathbf{F_x} = \left[\left(\frac{\partial f(x, \widehat{\theta_p})}{\partial \theta_1} \right), \cdots, \left(\frac{\partial f(x, \widehat{\theta_p})}{\partial \theta_p} \right) \right] \tag{3.27}$$

et s^2 la variance résiduelle estimée :

$$s^2 = \frac{\sum_{i=1}^{n} r_i}{n-p} \tag{3.28}$$

La variance résiduelle estimée peut être remplacée par la variance des observations $\sigma^2_{y_i}$. Dans les deux cas, il est de nouveau supposé que la contribution de la variance des variables explicatives est négligeable. Si cette hypothèse n'est pas valide, l'estimation de la variance des paramètres est alors sous-estimée par rapport à la véritable variance [24]. Pour prendre en compte les variations des variables explicatives, les valeurs de la matrice $\mathbf{W_r}$ (Voir éq. 3.12) peuvent être utilisées à la place s^2 ou de $\sigma^2_{y_i}$.

La matrice de variance-covariance des paramètres, $\mathbf{Cov}(\widehat{\theta_\mathbf{p}})$, est couramment estimée par l'une des manières suivantes [1, 27, 14] :

$$\mathbf{Cov}(\widehat{\theta_\mathbf{p}}) = \mathbf{s^2} \left(\mathbf{J}(\widehat{\theta_\mathbf{p}})^\mathbf{T} \mathbf{J}(\widehat{\theta_\mathbf{p}}) \right)^{-\mathbf{1}} \tag{3.29}$$

$$\mathbf{Cov}(\widehat{\theta_\mathbf{p}}) = \mathbf{s^2} \mathbf{H}(\widehat{\theta_\mathbf{p}})^{-\mathbf{1}} \tag{3.30}$$

$$\mathbf{Cov}(\widehat{\theta_\mathbf{p}}) = \mathbf{s^2} \mathbf{H}(\widehat{\theta_\mathbf{p}})^{-\mathbf{1}} \left(\mathbf{J}(\widehat{\theta_\mathbf{p}})^\mathbf{T} \mathbf{J}(\widehat{\theta_\mathbf{p}}) \right) \mathbf{H}(\widehat{\theta_\mathbf{p}})^{-\mathbf{1}} \tag{3.31}$$

où $\mathbf{J}(\widehat{\theta_\mathbf{p}})$ représente la matrice jacobienne de la fonction $f(x, \theta_p)$ pour $\widehat{\theta_p}$, la valeur estimée optimale des paramètres.

$$J = \left. \frac{\partial f(x, \theta_p)}{\partial \theta} \right|_{\theta = \widehat{\theta}} \tag{3.32}$$

et $\mathbf{H}(\widehat{\theta})$ représente la matrice hessienne de la fonction $f(x, \theta_p)$ pour $\widehat{\theta_p}$. Toutefois, il a été démontré que les approximations utilisant la matrice hessienne (Voir éq. 3.30 et 3.31), plus chères et moins stables numériquement, ne permettent pas d'améliorer l'estimation de $\mathbf{Cov}(\widehat{\theta_\mathbf{p}})$ [27].

Dans l'équation 3.26, m peut être égal à 1 ou l'infini [81]. En effet, si y est considérée comme la moyenne d'une infinité d'observation, alors $m \to \infty$. Dans ce cas, la bande de confiance calculée correspond à celle de la courbe ajustée, c'est-à-dire les limites dans lesquelles se situe une valeur individuellement lue sur la courbe. A l'inverse, si y est vue comme une valeur individuelle, $m = 1$ [1]. La bande de confiance ainsi calculée correspond à la prédiction d'une estimation, c'est-à-dire les limites dans lesquelles se situent une nouvelle observation issue de la même population statistique que l'échantillon. La bande de prédiction correspond donc à la bande de confiance à laquelle a été ajoutée la variance d'une nouvelle observation [25, 81] (Voir fig. 3.2).

Les bandes ainsi obtenues sont dites *simultanées*. Il est possible d'obtenir à partir de la courbe ajustée, pour un niveau de confiance donné, une valeur de y à partir de toute la plage de valeurs de x. Si la prédiction est limitée à un seul x, il faut faire appel à un intervalle dit *non simultané*. Pour obtenir cet intervalle, le terme $\sqrt{pF_{1-\alpha}(p, n-p)}$ de l'équation 3.26 doit être remplacé par $t(n-p, \alpha)$ [25, 81].

A l'inverse, afin d'obtenir à partir de la courbe estimée, pour un niveau de confiance donné, une valeur de x basée sur une valeur y_0, deux approches *inverses* sont possibles. La bande de prédiction, construite selon l'équation 3.25 pour $m = 1$, peut directement être utilisée [1]. Un exemple est présenté figure 3.3.a. Il est également possible d'utiliser la bande de confiance, construite selon l'équation 3.25 pour $m \to \infty$, en ajoutant une

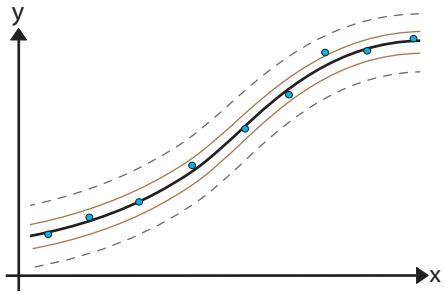


FIGURE 3.2 – Bandes de confiance (trait fin) et de prédiction (trait pointillé) d'une courbe ajustée (trait gras).

erreur à la valeur y_0. Un intervalle de confiance à $(1-\alpha)100\%$ autour de la valeur y_0 est construit sur l'axe, puis projeté horizontalement sur la bande de confiance [12]. Un exemple est présenté figure 3.3.b.

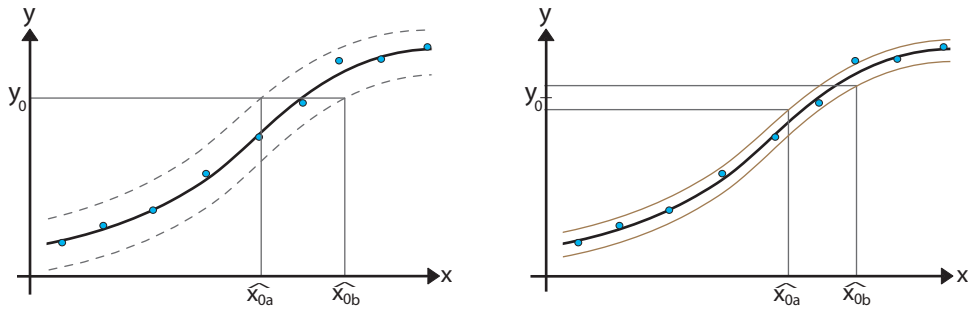


FIGURE 3.3 – Construction de l'intervalle *inverse* (sur x) à partir a) de la bande de prédiction et b) de la bande de confiance avec un intervalle sur y_0.

3.4.2 Estimation de la bande de confiance basée sur le calcul d'un hypervolume de confiance des paramètres

L'estimation d'un hypervolume de confiance sur les paramètres consiste à simuler de très nombreux jeux de paramètres. Plusieurs voies sont possibles.

Des jeux de paramètres ajustés $\widehat{\theta^*}$ sont obtenus en réalisant des ajustements à partir de B sous-échantillons $\{y_i^*\}$ de l'ensemble des observations $\{y_i\}$. Ces sous-échantillons peuvent être obtenus par *bootstrap* [17, 45]. A partir des jeux de paramètres simulés, il est possible d'écrire :

$$(\widehat{\theta^*}-\widehat{\theta})^T\mathbf{Cov}(\widehat{\theta})^{-1}(\widehat{\theta^*}-\widehat{\theta}) \widetilde{\rightarrow} \chi_p^2 \tag{3.33}$$

où $\widetilde{\rightarrow}$ signifie approximativement distribué selon un χ_p^2 à p degré de liberté. Ainsi :

$$Pr\left\{(\widehat{\theta^*}-\widehat{\theta})^T\mathbf{Cov}(\widehat{\theta})^{-1}(\widehat{\theta^*}-\widehat{\theta}) \leq \chi_p^2(\alpha)\right\} = 1-\alpha \tag{3.34}$$

En définissant :

$$R(\alpha) = \left\{\theta : (\widehat{\theta^*}-\widehat{\theta})^T\mathbf{Cov}(\widehat{\theta})^{-1}(\widehat{\theta^*}-\widehat{\theta}) \leq \chi_p^2(\alpha)\right\} \tag{3.35}$$

$R(\alpha)$ est la région de confiance à $(1-\alpha)$ 100% pour θ. Pour chaque valeur de x, il est possible de trouver les valeurs $\theta_L(x)$ et $\theta_U(x)$ qui minimisent et maximisent $f(x,\theta)$. C'est-à-dire :

$$\begin{aligned} \theta_L(x) &= arg[\min_{\theta \in R(\alpha)} f(x,\theta)] \\ \theta_U(x) &= arg[\max_{\theta \in R(\alpha)} f(x,\theta)] \end{aligned} \tag{3.36}$$

En utilisant les valeurs de paramètres ainsi trouvées, la bande de confiance est obtenue par :

$$\begin{aligned} y_L(x) &= f(x,\theta_L(x)) \\ y_U(x) &= f(x,\theta_U(x)) \end{aligned} \tag{3.37}$$

En utilisant les mêmes conditions citées précédemment et la méthode du *boodstrap*, il est possible de générer des $\widehat{\theta^*}$ en utilisant l'expression :

$$\widehat{\theta^*} \sim \mathcal{N}\left(\widehat{\theta}, (\widehat{\theta^*} - \widehat{\theta})^T \mathbf{Cov}(\widehat{\theta})^{-1} (\widehat{\theta^*} - \widehat{\theta})\right) \tag{3.38}$$

puis, pour chaque itération, de calculer $h(x, \widehat{\theta_p^*})$ à partir des $\widehat{\theta_p^*}$ obtenus (Voir éq. 3.26). Il est ainsi possible d'obtenir l'estimation de la fréquence cumulée de $h(x, \widehat{\theta_p^*})$. La partie de la distribution obtenue au niveau $(1-\alpha)$ correspond au percentile limite de confiance [19].

3.4.3 Conclusion sur l'estimation de bande de confiance

Les méthodes de linéarisation permettent de construire une bande de confiance ainsi qu'une bande de prédiction autour de la courbe ajustée. Dans les deux cas, elles peuvent être simultanées ou non-simultanées. Leur utilisation permet d'exploiter de manière rigoureuse la courbe ajustée. En effet, elles permettent d'estimer, pour un risque donné, une valeur de y pour un x donné ou inversement, une valeur de x pour un y donné [1, 12]. Ces méthodes présentent de nombreux avantages. Elles fournissent généralement des résultats corrects dans de nombreuses situations [81]. Une grande partie des calculs est déjà réalisée au cours de l'optimisation de la fonction d'objectif. Cela permet d'avoir des calculs très rapides. De plus, il existe dans la littérature des solutions déjà implémentées pour accéder à la précision des paramètres estimés, nécessaires à l'utilisation de ces méthodes [14, 16].

Les méthodes basées sur le calcul d'un hypervolume de confiance des paramètres servent également à évaluer l'incertitude sur les paramètres [19, 67, 97]. Les mêmes remarques peuvent donc être faites pour leur utilisation dans la construction de la bande de confiance. Ces méthodes de simulation sont des outils efficaces sous certaines conditions. L'une de ces conditions est l'utilisation d'un estimateur non biaisé [17, 24]. En effet, ces méthodes simulent des perturbations autour de l'estimation des paramètres considérée comme optimale (pour le jeu de mesure original). L'un des avantages est d'éviter d'approximer $f(.)$. Néanmoins, ces méthodes sont difficiles à intégrer dans un calcul automatique de routine. Des problèmes d'optimisation peuvent survenir pour une réalisation particulière d'un sous-échantillon [24]. Enfin, la lenteur de ces méthodes (fortement liée

à la puissance de calcul des ordinateurs), peut également être un inconvénient en raison du grand nombre de simulations à traiter (M compris entre 500 et 1000 pour l'intervalle de confiance [54, 67]).

Toutes les solutions proposées reposent sur l'utilisation de la matrice de variance covariance des paramètres. Le calcul de cette matrice fait appel aux résidus $(y_i - \widehat{y_i})$ et dépend de la variable explicative (x). Aucune méthode ne prend en compte de manière explicite l'incertitude que peut avoir l'expérimentateur sur ses connaissances expertes.

3.5 Conclusion

Cet état de l'art ciblé sur les questions soulevées par l'analyse de la problématique met en évidence les solutions possibles à mettre en œuvre afin de mener à bien l'objectif opérationnel de la thèse : la modélisation de cinétiques prédictives, à partir de mesures et de connaissances expertes, tout en tenant compte de leur incertitude respective.

Différents estimateurs fournissent le moyen de prendre en compte l'imprécision des données lors de l'estimation des paramètres : maximum de vraisemblance et les moindres carrés. Plusieurs hypothèses peuvent être émises sur les erreurs d'estimation du modèle, $\sigma^2_{\varepsilon_i}$, et sur de l'imprécision des données. Ces estimateurs sont déclinés sous différentes formes afin de prendre en considération ces hypothèses. De manière générale, l'estimateur du maximum de vraisemblance repose sur des bases statistiques et théoriques plus solides.

Le cadre bayésien présente deux avantages. Son emploi permet :

- l'utilisation conjointe des mesures de suivi et d'information *a priori* sur les paramètres lors de leur estimation,
- la prise en compte de l'hétéroscédasticité des erreurs d'estimation du modèle (utilisation de la vraisemblance).

Dans le cadre bayésien, les informations *a priori* sur les paramètres sont exploitées lors de l'ajustement grâce aux *priors*, assimilables à des lois décrivant les paramètres.

Les connaissances expertes sont une source d'information nécessaire au bon ajustement des paramètres. Malheureusement, leur traduction en une information *a priori*, au travers d'un système à base de règles floues, fournit une distribution de possibilité des paramètres. Ce cadre de représentation est incompatible avec celui des mesures de suivi. La transformation de ces distributions de possibilité en distributions de probabilité offre une réponse à ce problème. Cette transformation permet de disposer de *priors* :

- issus des connaissances expertes,
- ayant un cadre de représentation identique aux mesures de suivi.

Cet état de l'art met également en évidence les propriétés de la matrice de variance covariance des paramètres pour manipuler les incertitudes.

Enfin, il met en lumière de nombreux points méthodologiques qui ne sont pas maîtrisés pour la construction de la bande de confiance. Toutes les méthodes reposent sur les résidus et fournissent une bande de confiance dépendante de x. Mais aucune ne permet

de prendre en compte clairement le degré d'incertitude des informations *a priori*.

Chapitre 4

Proposition scientifique et implémentation

Sommaire

4.1 Proposition scientifique

Les réponses apportées par la revue bibliographique, associées à l'analyse de la problématique permettent d'élaborer une proposition scientifique.

Cette proposition repose sur :
- la construction d'une information *a priori* sur les paramètres basée sur les connaissances expertes et utilisables dans un cadre probabiliste,
- l'estimation des paramètres en employant conjointement les mesures de suivi et les informations *a priori*,
- une méthode de calcul de la bande de confiance qui tient compte de l'incertitude des mesures de suivi et des connaissances expertes.

4.1.1 Construction des *priors* à partir des connaissances expertes

Les connaissances expertes, qui sont des informations subjectives imprécises, sont traduites en une valeur possible du paramètre au travers d'un système à base de règles floues (ou système d'inférence floue, SIF). Les propriétés de ces systèmes fournissent un moyen de gérer l'incertitude liée à ces informations. Dans la pratique, les entrées du système sont renseignées sous la forme d'un intervalle. La largeur de cet intervalle traduit l'incertitude que peut avoir l'utilisateur au sujet des connaissances expertes à renseigner. Cet intervalle est transformé en un nombre flou triangulaire. Le support de ce nombre flou correspond à la largeur de l'intervalle et le noyau à sa valeur centrale (Voir fig. 4.1).

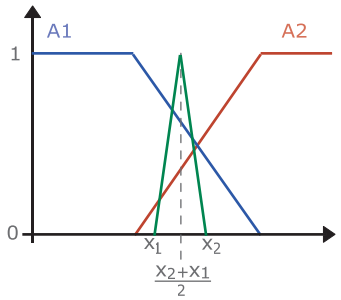


Figure 4.1 – Traduction de l'incertitude des connaissances expertes.

Des règles de type *implicatives* doivent êtré utilisées pour prendre en compte l'incertitude des connaissances expertes. L'agrégation de ces règles se fait de manière *conjonctive* dont l'opérateur est le *ET*. Les règles *implicatives* imposent donc des contraintes sur la conclusion au moment de leur activation. L'activation d'une règle supprime des valeurs possibles de sortie. A l'inverse, si aucune règle n'est activée, toutes les valeurs de sortie sont alors possibles.

La distribution de possibilité (fournie par le SIF) est transformée en une distribution de probabilité. Cette étape permet de disposer d'une information *a priori* sur les paramètres dont le cadre de représentation est identique à celui des mesures. Cette information est donc directement utilisable dans le cadre bayésien "classique" (Voir chap. 3.3 et 4.1).

4.1.2 Estimation des paramètres en utilisant conjointement les mesures de suivi et les informations *a priori*

La méthode bayésienne est employée pour estimer la valeur des paramètres la plus probable, en utilisant conjointement les mesures de suivi et les informations *a priori* sur les paramètres (obtenues à partir des connaissances expertes).

La méthode bayésienne prend en compte l'imprécision des données grâce à l'usage de l'estimateur du maximum de la vraisemblance. Les variances des erreurs d'estimation du modèle, $\sigma^2_{\varepsilon_{yi}}$ et $\sigma^2_{\varepsilon_{xi}}$, sont estimées à partir des variances des données de suivi, $\sigma^2_{y_i}$ et $\sigma^2_{x_i}$. En effet, si une observation est issue de k sous-observations, où k est grand, il est possible d'approcher convenablement les $\sigma^2_{\varepsilon_i}$ à partir des données.

4.1.3 Construction de la bande de confiance

Les méthodes de calcul de la bande de confiance trouvées dans la littérature ne peuvent pas être mises en œuvre pour cette problématique. En effet, ces méthodes fournissent une bande de confiance très fortement liée aux résidus et dépendante de la variable explicative (x). L'amplitude de la bande sur la partie *prédictive* de la cinétique augmente donc de manière démesurée (Voir fig. 4.2). En effet, l'expérience montre que la courbe ajustée à partir de l'ensemble des informations (mesures et connaissances expertes) sera moins proche des observations que celle issue d'un ajustement *classique* s'appuyant seulement sur les mesures. Cette amplitude rend les résultats inexploitables et pourrait faire douter de la validité de la courbe ajustée.

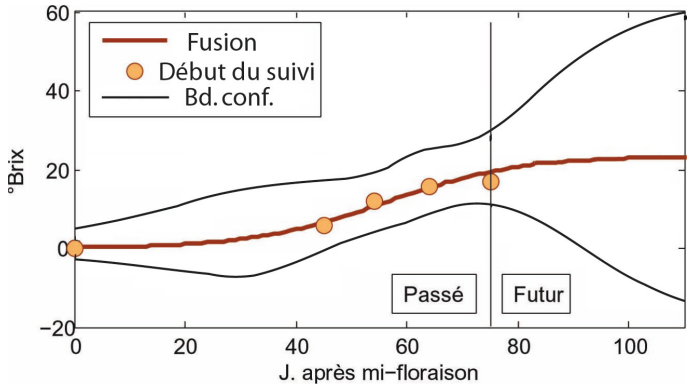


Figure 4.2 – Bande de confiance calculée à partir des mesures ayant servies à l'ajustement.

De plus, l'incertitude que peut avoir l'utilisateur sur les connaissances expertes doit être prise en compte. Certaines connaissances expertes reposent sur des évènement passés et sont donc connues avec une faible incertitude. A l'inverse, d'autres connaissances expertes sont des hypothèses, ces dernières sont donc très incertaines. Mais les méthodes référencées n'en tiennent pas compte de manière explicite.

Il est donc proposé de construire une bande de confiance en utilisant une matrice de variance-covariance des paramètres estimés, $\mathbf{Cov}(\widehat{\theta_\mathbf{p}})$, basée :

– sur la variance des mesures de suivi (les mesures sont supposées indépendantes) afin de prendre en compte leur incertitude,

– et sur la variance-covariance des paramètres *a priori* émanant des SIF, afin de tenir compte de l'incertitude des connaissances expertes.

A partir de cette matrice de variance-covariance globale des paramètres estimés, $\mathbf{Cov}(\widehat{\theta_\mathbf{p}})$, et des paramètres estimés, $\widehat{\theta_p}$, un hypervolume de confiance est formé. En d'autres termes, de nombreux jeux de paramètres $\widehat{\theta_p^*}$ sont simulés en s'appuyant sur cette matrice de variance-covariance. La bande de confiance, au niveau α, est ensuite obtenue par (Voir chap. 3.4.2) :

$$\begin{aligned} y_L(x) &= \{\underset{\theta \in R(\alpha)}{min}\ f(x, \widehat{\theta_p^*})\} \\ y_U(x) &= \{\underset{\theta \in R(\alpha)}{max}\ f(x, \widehat{\theta_p^*})\} \end{aligned} \tag{4.1}$$

4.2 Implémentation de la méthode

La mise en œuvre de la proposition scientifique est soumise à des contraintes liées au logiciel utilisé pour son implémentation. Ces contraintes engendrent donc certaines modifications de la proposition. De plus, certaines simplifications peuvent être proposées.

De manière générale, un système est construit de manière indépendante pour chacun des critères de maturité. Ce système est défini par :

1. une fonction permettant de modéliser la cinétique du critère de maturité étudié,
2. une "boîte experte" pour obtenir une information *a priori* sur les paramètres de cette fonction.
3. une "boîte de *fitting*" afin d'estimer les valeurs les plus probables de la cinétique en fonction des mesures et des informations *a priori*
4. et d'un module pour calculer la bande de confiance.

4.2.1 Implémentation de la "boîte experte" pour obtenir les informations *a priori*

La "boîte experte" est formée de différents systèmes d'inférence floue (SIF) ; un pour chaque paramètre de la fonction modélisant la cinétique du critère de maturité étudié.

Les règles contenues dans ces SIF sont induites par un apprentissage supervisé utilisant une base de données. Ce type d'apprentissage consiste à construire, par des procédures automatiques, des règles qui permettent de reproduire la relation entre les données d'entrée (connaissances expertes du viticulteur) et celles de sortie (valeurs possibles du paramètre). Malheureusement, à l'heure actuelle, les systèmes d'apprentissage automatiques générant des règles *implicatives* n'en sont qu'à leurs prémisses. Les méthodes permettant de générer ces règles ne sont pas encore validées.

Les règles produites par apprentissage sont de type *conjonctive*. La philosophie associée à ces règles est à l'inverse des règles *implicatives*. C'est-à-dire que l'activation d'une règle garantit une valeur de conclusion comme possible à un certain niveau [32, 33]. Par exemple, pour une variable x "Plus la valeur prise par x appartient au sous-ensemble

A, *alors* plus il est possible que la valeur prise par la conclusion y appartient au sous-ensemble B de la conclusion". L'agrégation de ces règles se fait de manière *disjonctive* dont l'opérateur est le *OU*. En d'autres termes, l'activation d'une règle offre une valeur garantie comme possible et la conclusion est l'union de ces possibilités. Plus les règles activées sont nombreuses, plus la distribution de possibilité en sortie sera élargie. Par conséquent, l'utilisation d'un nombre flou pour traduire l'incertitude des connaissances expertes n'entraîne pas de modification notable sur la distribution de possibilité de la sortie (Voir fig. 4.3).

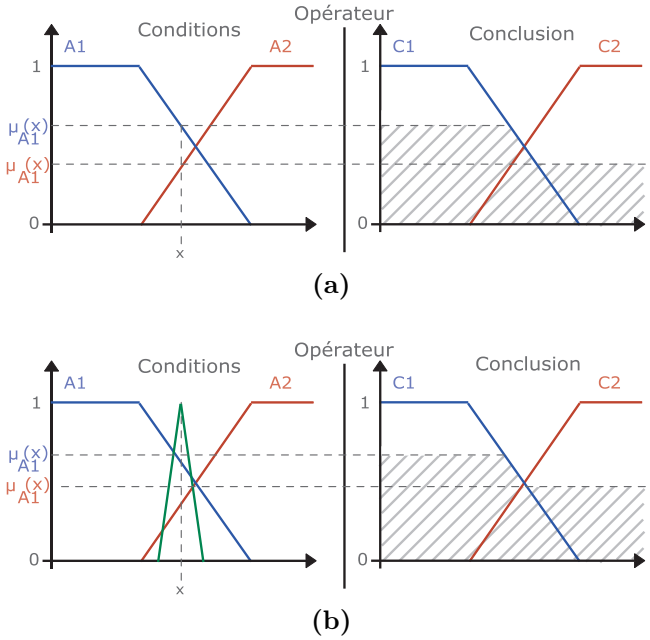


Figure 4.3 – Conclusion issue de règles *conjonctives* avec (a) une entrée nette (b) une entrée floue.

De plus, une fois les règles générées, les SIF sont implémentés grâce au logiciel Matlab® et sa *Fuzzy Logic Toolbox*™. Ce dernier ne fournit pas la fonction permettant de construire la distribution de possibilité des sorties inférées.

Ces deux contraintes dues à l'implémentation : utilisation de règles *conjonctives* et impossibilité d'avoir la distribution de possibilité en sortie, obligent à modifier la méthode exposée dans la proposition scientifique.

La certitude des connaissances expertes est toujours traduite par un intervalle mais son utilisation est différente. De nombreux jeux de valeurs d'entrée du SIF sont obtenus par un tirage aléatoire avec remise sur cet intervalle. La méthode de ré-échantillonnage du *bootstrap* est employée pour générer ces valeurs. Les différentes sorties du SIF ainsi obtenues sont alors défuzzifiées en un nombre *net*. La défuzzification appliquée est celle de *Sugeno*, c'est-à-dire que la sortie est calculée comme la somme pondérée des conclusions de chacune des règles, le poids étant le degré de vérité de la règle. Ensuite, la variance de la sortie est calculée.

Les informations *a priori* sur les paramètres sont ensuite modélisées par une loi de probabilité appropriée aux données.

4.2.2 Implémentation de la "boîte de *fitting*"

La méthode bayésienne est toujours employée pour estimer la valeur des paramètres la plus probable, afin d'utiliser conjointement les mesures de suivi et les informations *a priori* sur les paramètres.

La variance de la variable explicative (x) est négligée. En effet, dans la pratique les variations de x, au cours de la réalisation d'une observation, sont négligeables par rapport à la durée de la maturation et surtout par rapport à celle des observations. Par conséquent, les erreurs de mesure sont considérées comme la seule source de variance au moment de l'ajustement bayésien.

La variance des erreurs d'estimation du modèle est toujours estimée à partir de la variance des observations, $\sigma^2_{y_i}$.

Les valeurs d'initialisation de l'algorithme d'optimisation sont celles des paramètres obtenus à partir de la valeur centrale de l'intervalle défini comme entrée du SIF.

4.2.3 Construction de la bande de confiance

La matrice de variance-covariance des paramètres *a priori* est obtenue en additionnant la variance des erreurs d'étalonnage des SIF et la variance-covariance des valeurs de sortie inférées à partir des différents SIF. Cette matrice est ensuite utilisée pour simuler, à partir des paramètres estimés, $\widehat{\theta_p}$, de nombreux jeux de paramètres, $\widehat{\theta^*_p}$.

De la même manière, à partir de la variance des observations ($\sigma^2_{y_i}$) et des observations ($\widehat{y_i}$), de nombreux jeux d'observations, $\widehat{y^*_i}$, sont simulés.

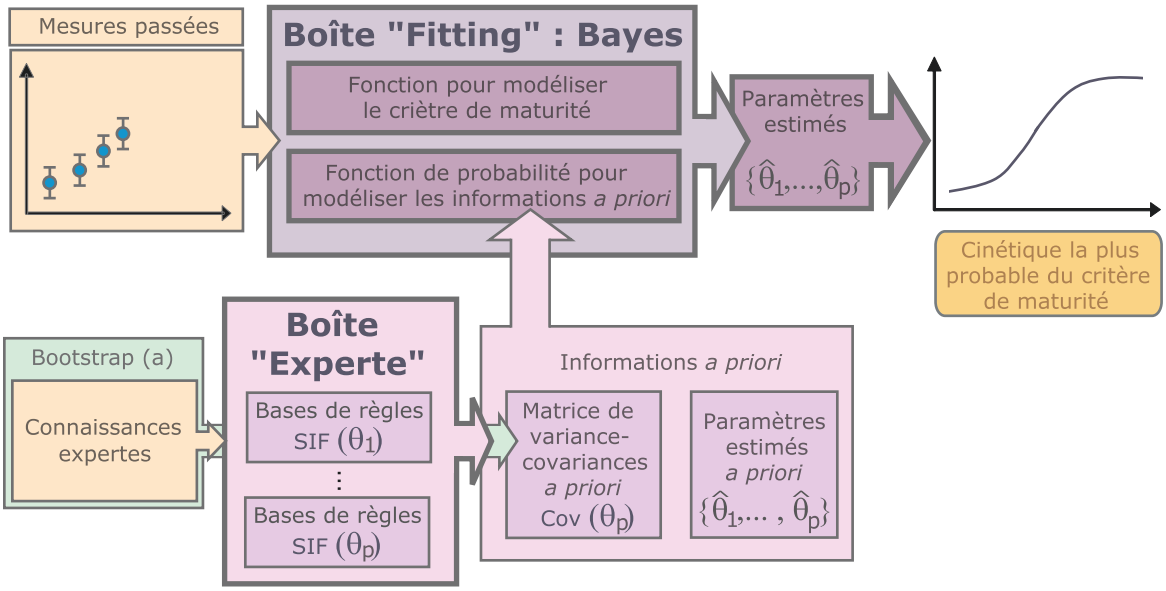
Pour chacun de ces jeux de données simulés ($\widehat{\theta^*_p}$ et $\widehat{y^*_i}$), les paramètres de la cinétique sont estimés. La bande de confiance est obtenue comme exposé précédemment (Voir chap. 4.1.3).

4.2.4 Schéma de synthèse de la proposition scientifique implémentée

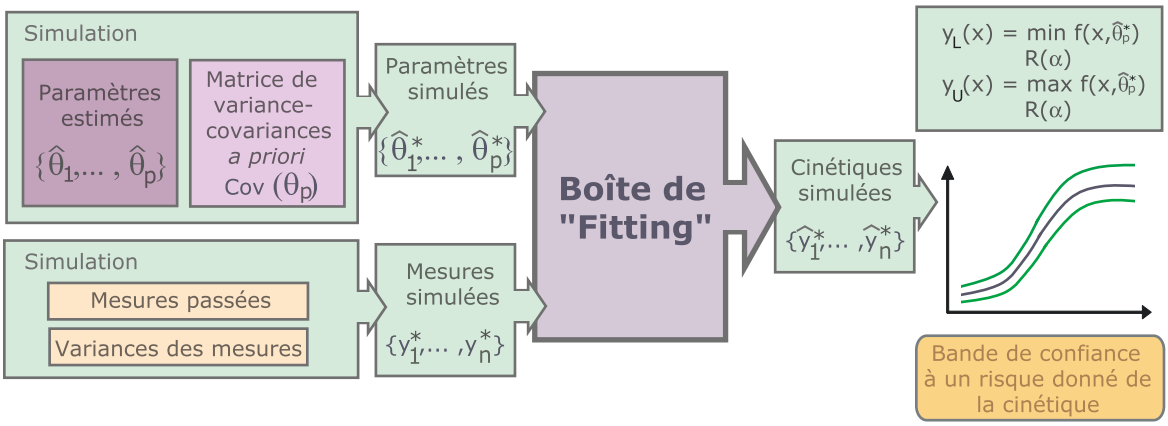
Les figures 4.4 présentent de manière schématisée la méthodologie adaptée aux contraintes liées à son implémentation.

4.3 Conclusion

La proposition scientifique, ainsi implémentée, permet d'estimer les paramètres d'une cinétique prédictive à partir des premières mesures de suivi et de connaissances expertes. La méthode de construction de la bande de confiance retenue est directement le reflet de l'imprécision de l'ensemble des données.



(a)



(b)

Figure 4.4 – Implémentation de la proposition scientifique (a) estimation des paramètres (b) calcul de la bande de confiance.

Cette méthode est appliquée à la modélisation de cinétiques prédictives de teneur en sucre et d'acidité totale afin de mettre en relief ses capacités et ses limites.

Le chapitre suivant détaille :
- les fonctions utilisées pour chacun de ces systèmes : système "teneur en sucre" et système "acidité totale",
- la base de données utilisée pour inférer les règles contenues dans les différents SIF,
- la méthode de construction des "boîtes expertes",
- la méthode de construction des "boîtes de *fitting*",
- et la méthode de calcul de la bande de confiance.

Chapitre 5

Matériels et méthodes

Sommaire

5.1 Fonctions utilisées pour modéliser les cinétiques d'évolution des critères de maturité

Les fonctions, aussi simples que possible, ont été choisies par rapport à l'allure des dynamiques d'évolution des critères de maturité observés. En effet, elles n'ont pas de justification biologique ou mathématique mais ont été uniquement retenues pour leur capacité à reproduire de manière satisfaisante les phénomènes étudiés. Néanmoins, les paramètres de ces équations sont interprétables en termes concrets. Ces propriétés permettent de facilement paramétrer ces fonctions à partir des connaissances expertes du viticulteur. Compte-tenu des objectifs de ces travaux, cette démarche est pleinement justifiable [49].

Il a été décidé de modéliser dans un premier temps les cinétiques de teneur en sucre et d'acidité totale. En effet, de par leur facilité de mesure, un grand nombre de suivi est disponible.

5.1.1 Teneur en sucre

La dynamique de croissance de la teneur en sucre est modélisée par une sigmoïde :

$$T_S = \frac{T_{S_{max}}}{1 + exp^{\frac{(t_{1/2} - j_{amf})}{\tau}}} \tag{5.1}$$

où :
- T_s la teneur en sucre,
- $T_{S_{max}}$ la teneur finale en sucre au moment de la vendange,
- $t_{1/2}$ le temps nécessaire pour atteindre la moitié de la valeur finale,
- j_{amf} jour après mi-floraison (variable explicative appelée x dans les chapitres précédents),
- τ paramètre qui permet d'accentuer ou ralentir l'évolution. Par la suite, ce paramètre sera improprement appelé *pente*.

Cette sigmoïde possède pour asymptotes les droites d'équation $T_S = 0$ et $T_S = T_{S_{max}}$. Elle possède un point d'inflexion à $t = t_{1/2}$, qui représente le temps où la teneur a atteint la moitié de sa valeur maximale. En ce point, la dérivée est nulle.

5.1.2 Acidité totale

La dynamique de décroissance de l'acidité totale est modélisée par l'équation suivante :

$$T_{AT} = T_{AT_0} - \pi . \frac{j_m}{(t^{\pi}_{1/2} + j_m)} \tag{5.2}$$

où :
- T_{AT} l'acidité totale,
- T_{AT_0} l'acidité totale au moment de la première mesure,
- π le taux de décroissance maximale de l'acidité totale,
- j_m jour après la première mesure (variable explicative appelée x dans les chapitres précédents),

– $t^{\pi}_{1/2}$ égale au nombre de jours lorsque le taux de décroissance a atteint la moitié de sa valeur maximale.

Cette équation possède pour asymptote la droite d'équation T_{AT} = acidité totale finale. Elle est monotone décroissante et n'a pas de point d'inflexion.

5.2 Base de données utilisée pour inférer les règles

Comme expliqué précédemment dans le chapitre 4.2.1, les règles contenues dans les SIF sont induites par un apprentissage supervisé utilisant une base de données.

Cette base de données comprend :
– des mesures de suivi,
– des connaissances expertes au sujet de chacun des suivis.

Il a été décidé de prendre comme connaissances expertes :
– les données météorologiques de l'année,
– la contrainte hydrique globale de la parcelle.

En effet, ces connaissances expertes sont couramment utilisées par la profession, sont facilement accessibles et ont une grande influence sur la maturation (Voir chap. 1.4). De plus, il est aisé d'émettre ou de trouver des hypothèses à leur sujet.

Cette base de données est issue d'expérimentations anciennes ou actuelles de l'Unité Expérimentale de Pech Rouge, INRA de Gruissan, France (latitude 43 ° 08' 35"N ; longitude 3 ° 7' 59'E). Le choix de cette base de données a été motivé par le fait qu'elle comprend des suivis de maturité très complets, réalisés sur plusieurs parcelles et sur plusieurs années. Néanmoins, il a été décidé de restreindre la base de données à un seul cépage, la Syrah.

Le principal critère de sélection des suivis présents dans la base de données est le nombre de mesures réalisées par suivi. Il doit être supérieur au nombre de paramètres à estimer pour éviter tout problème d'identifiabilité lors de l'estimation. Ces mesures de suivi doivent également être réparties sur l'ensemble de la période de maturation (de la mi-véraison à la vendange).

Une fois les suivis sélectionnés, ils sont divisés en deux groupes de manière aléatoire :
– 80% vont dans l'ensemble d'apprentissage,
– 20% vont dans l'ensemble de test.

L'ensemble d'apprentissage sert à inférer les règles. L'ensemble de test sert, quant à lui, à vérifier la capacité des systèmes à générer des cinétiques prédictives, à partir de données inconnues.

5.2.1 Parcelles

Les parcelles retenues sont donc des parcelles de Syrah suivies durant les saisons 2003 à 2008. Ces parcelles présentent l'avantage d'avoir différentes modalités de conduite, de sol et d'orientation :

- Parcelle 63 dite *Syrah Vila* :
  - clone 99, porte-greffe R-140,
  - plantée en 1990,
  - espacée de 2,50 m x 1 m,
  - conduite espalier haute, taillée en guyot
  - sol calcaire dur fissuré
  - orientation des rangs nord-ouest sud-est, la parcelle est orientée vers le sud
  - non irriguée.

- Parcelle 76 dite *Syrah Israël*
  - clone 174, porte-greffe sur R-140,
  - plantée en 1993,
  - espacée de 2,50 m x 1 m,
  - conduite espalier haute, taillée en cordon
  - sol calcaire dur fissuré
  - orientation des rangs ouest-est, la parcelle est orientée vers le sud
  - non irriguée.

- Parcelle 22 dite *Syrah Bedarde*
  - clone "collection", porte-greffe R-140,
  - plantée en 1995,
  - espacée de 2,50 m x 1 m,
  - conduite espalier haute, taillée en cordon
  - sol sablo-limoneuse
  - orientation des rangs nord-sud
  - irriguée ou non par un goutte à goutte à partir de 2008.

5.2.2 Données météorologiques

De manière générale, le climat de ce vignoble est de type méditerranéen (hivers doux et étés secs et chauds). Les jours de pluie sont peu nombreux (moins de 80 jours par an), irréguliers et mal repartis, avec des épisodes torrentiels à l'automne (pluviométrie annuelle entre 400 et 800mm). Les vents sont secs et violents (le mistral, le cers et la tramontane). L'insolation annuelle est d'environ 2700 heures.

Les données météorologiques sont obtenues à partir d'une station météorologique automatique, située à l'U.E. de Pech Rouge. Les données sont exprimées par pas de temps horaire et journalier. Parmi toutes les variables météorologiques disponibles, celles retenues le sont pour leur influence sur la maturation :

- l'évapotranspiration Penman (mm),
- le rayonnement global (J/cm^2),
- les précipitations (mm),

– les humidités journalières (%) minimales, maximales et moyennes,
– les températures journalières (° c) minimales, maximales et moyennes.

Les sommes de ces différents paramètres climatiques ont été réalisées sur différentes périodes :
– du premier avril à la mi-floraison,
– de la mi-floraison à la mi-véraison,
– du premier avril à la mi-véraison,
– de la mi-véraison aux vendanges.

5.2.3 Caractérisation de l'état hydrique

L'état hydrique de la vigne est obtenu à partir de la mesure du potentiel hydrique foliaire de base (Ψ_b). Cette mesure s'obtient au moyen d'une chambre à pression [83]. Le but est d'estimer la capacité des cellules d'une feuille à retenir l'eau lorsque cette dernière est soumise à la pression d'un gaz neutre. Moins il y aura d'eau libre dans la plante, plus la pression à exercer sur la feuille pour la faire sortir par le pédoncule sectionné sera grande. Les résultats sont exprimés en *bar* ou en MPa.

Les mesures sont effectuées en fin de nuit. A ce moment là, l'état hydrique de la plante est en équilibre avec celui du sol dans sa zone racinaire car la transpiration nocturne est considérée comme nulle [56].

Les relations entre le niveau de contrainte hydrique et le potentiel foliaire de base utilisées dans ces travaux sont présentées dans le tableau 5.1.

Niveau de contrainte hydrique	Seuils de potentiel foliaire de base
Absence de contrainte hydrique	De 0 à -2 bars ou 0 à -0,2 MPa
Contrainte hydrique faible	De -2 à -4 bars ou -0,2 à -0,4 MPa
Contrainte hydrique moyenne	De -4 à -6 bars ou -0,4 à -0,6 MPa
Contrainte hydrique forte	$<$ -6 bars ou 0 à -0,6 MPa

Table 5.1 – Relations entre le niveau de contrainte hydrique et le potentiel foliaire de base utilisées pour ces travaux.

5.3 Construction de la "boîte experte"

5.3.1 Variables de sortie

Afin d'obtenir les valeurs *a priori* des paramètres, les suivis de l'ensemble d'apprentissage sont modélisés en effectuant une régression. En première approche, les variances des erreurs d'estimation du modèle, $\sigma^2_{\varepsilon_{yi}}$ sont supposées normales. L'estimateur des moindres carrés pondérés (équivalant à celui du maximum de vraisemblance dans ces conditions, voir chap. 3.2) est donc employé pour réaliser ces régressions.

Afin d'être le plus proche possible des conditions d'utilisation du Spectron™, chacune des mesures de suivi est affectée d'une variance issue de mesures de suivi effectuées avec ce capteur [1].

Les paramètres sont estimés avec la fonction *nlinfit* du logiciel Matlab® v7.0 (The Math Works, Inc., Natick, MA) :

$$\widehat{\theta_p} = \ nlinfit\ (\{x_i\}, \{y_i^v\}, \mathcal{M}^v(.), \theta_{p0}) \tag{5.3}$$

où
- $\widehat{\theta_p}$ vecteur contenant les valeurs estimées des p paramètres du modèle,
- $\{x_i\}$ vecteur contenant les valeurs de la variable explicative,
- $\{y_i^v\}$ vecteur contenant les valeurs des observations affectées de leur variance,
- $\mathcal{M}^v(.)$ la fonction permettant de prendre en compte la variance des observations,
- θ_{p0} vecteur contenant les valeurs d'initialisation pour l'algorithme d'optimisation.

L'algorithme d'optimisation utilisé par la fonction *nlinfit* est celui de Levenberg-Marquardt.

Les suivis ne permettant pas une bonne estimation des paramètres sont supprimés de la base d'apprentissage.

Au total, les paramètres de 29 cinétiques de teneur en sucre et 33 cinétiques d'acidité totale, réparties sur l'ensemble des saisons et des parcelles, ont pu être estimés.

5.3.2 Induction des règles floues

L'induction des règles floues contenues dans les différents SIF s'appuie sur un apprentissage supervisé (Voir chap. 4.2.1).

L'ensemble de ces étapes est fait au moyen du logiciel *FisPro* [2] [43]. Il s'agit d'un logiciel dédié à la conception et l'optimisation de systèmes d'inférence floue interprétables.

La construction d'un SIF se découpe en trois étapes : la sélection des variables d'entrée, la construction du partitionnement flou des entrées et l'induction des règles.

Construction du partitionnement flou des variables d'entrée

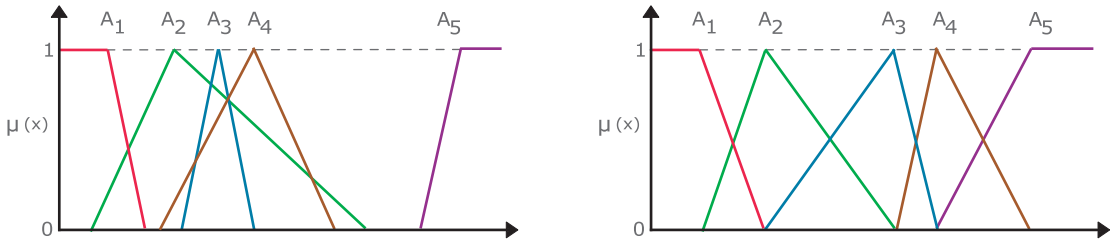
Le partitionnement des variables d'entrée, c'est-à-dire la définition des sous-ensembles flous correspondant aux termes linguistiques, revêt une grande importance. En effet l'intégrité sémantique de la partition floue doit être respectée. Si cette intégrité est enfreinte, les résultats du SIF ne sont plus interprétables en termes linguistiques, ce qui enlève son principal intérêt à ce type d'approche. La construction des partitions associées aux SIF est donc soumise à des contraintes qui assurent cette intégrité. Les conditions d'interprétabilité d'une partition floue sont [95] :
- des sous-ensembles identifiables,
- en nombre raisonnable,

1. Ces suivis ne sont pas utilisés car ils ne sont pas rattachés à des connaissances expertes
2. http ://www.inra.fr/internet/Departements/MIA/M/fispro/

– présentant un recouvrement significatif,
– la couverture du domaine de variation de la variable doit être complète
– et les sous-ensembles doivent être normalisés ($\forall i,\ \exists x,\ \mu_{A_i}(x) = 1$).

La figure 5.1.a présente un exemple de partition floue impossible à interpréter.

Pour garantir l'ensemble de ces conditions, des partitions floues fortes sont utilisées (Voir fig. 5.1.b). En plus de respecter les conditions édictées précédemment, avec ce type de partition, $\forall x$, x appartient à deux sous-ensembles au plus et $\forall x$, $\sum \mu_{A_i}(x) = 1$. Enfin, dans le logiciel *FisPro*, la conception des partitions est assurée à partir des seules valeurs de la variable correspondante, sans tenir compte des relations entre les variables, notamment celles de sortie. Ce choix permet d'exploiter les valeurs de la distribution pour elles-mêmes, sans présumer de leur éventuel pouvoir explicatif [44].



(a) **(b)**

Figure 5.1 – Exemple de deux partitions avec cinq sous-ensembles flous. a) Partition impossible à interpréter : : le sous-ensemble A_2 recouvre quasiment les sous-ensembles A_3 et A_4, une partie de la variable entre les sous-ensembles A_4 et A_5 n'est pas recouverte, etc. b) Partition floue forte interprétable.

Les variables d'entrée, hormis celle de la contrainte hydrique, sont divisées en sous-ensembles flous par la méthode des *k-means*. Cette méthode est introduite par J. McQueen [66]. Elle partitionne le domaine de variation de la variable en différentes classes isolées les unes des autres, dans lesquelles sont répartis les exemples de l'ensemble d'apprentissage. L'algorithme des *k-means* vise à minimiser la variance intra-classe :

– *a)* Initialisation des centres de chaque classe,
– *b)* Affectation des exemples au centre le plus proche,
– *c)* Ré-évaluation de la position des centres,
– *d)* Réitération des étapes *a* et *b* jusqu'à stabilisation.

Sélection des variables d'entrée

Même si toutes les variables connaissances expertes sont potentiellement intéressantes [3], il n'est pas envisageable de toutes les utiliser. En effet, plus il y a de variables d'entrée, plus il y a de règles possibles. Par exemple, un système comprenant 5 variables d'entrée et comptant chacune 4 labels linguistiques (sous-ensemble de la partition) peut conduire à 4^5 règles possibles, soit 1024. Ce nombre de règles rend leur analyse impossible et fournit donc un SIF non-interprétable. Une étape de sélection des variables d'entrée

3. neufs paramètres météorologiques sur quatre périodes plus la contrainte hydrique

est donc nécessaire pour réduire leur nombre et ainsi conserver les plus pertinentes.

Dans le cas présent, la sélection des variables est réalisée avec des arbres de décision flous [50]. Les arbres de décision flous sont une extension des arbres de décision classiques [79]. Leur construction est une procédure itérative. L'algorithme cherche à produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire, à partir des variables d'entrée. Le critère d'homogénéité des groupes est le maximum d'entropie.

Le *nœud racine* est le point de départ du processus. Le *nœud racine* correspond à l'une des variables d'entrée (p. ex. la contrainte hydrique). Les individus de la base d'apprentissage sont alors partagés entre différents *nœuds fils*. Le nombre de nœud fils est fonction du partitionnement de la variable d'entrée (p. ex. 4 pour la contrainte hydrique). La variable d'entrée utilisée au niveau du nœud est sélectionnée afin de produire des groupes d'individus les plus homogènes possibles, du point de vue de la variable à prédire, au niveau des *nœuds fils*. Pour choisir cette variable, l'algorithme teste toutes les variables d'entrée et choisit celle qui optimise le critère d'homogénéité des groupes. A partir de chacun des *nœuds fils*, une nouvelle segmentation est réalisée de la même manière : rechercher une variable d'entrée qui permet d'obtenir de nouveaux groupes les plus homogènes possibles, toujours du point de vue de la variable à prédire. Le processus est répété tant que les nœuds n'ont pas atteint un critère de pureté suffisant. Les *nœuds terminaux* sont appelés des *feuilles* (Voir fig. 5.2). Cette procédure permet donc d'ordonner les variables, les plus pertinentes sont celles utilisées en premier.

Les arbres de décision flous proposés dans *FisPro* sont basés sur une implémentation floue de l'algorithme *ID3* [100].

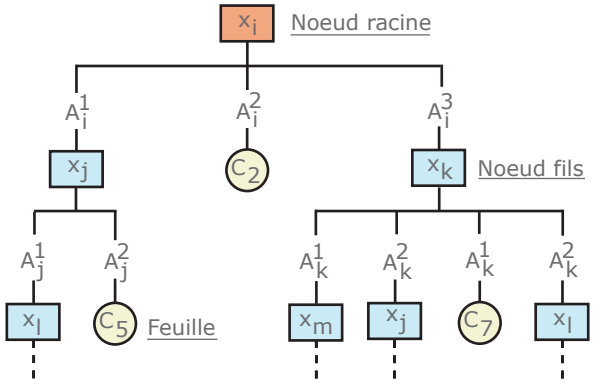


Figure 5.2 – Exemple d'arbre de décision flou [43].

En pratique, un arbre de décision est généré pour chacun des paramètres des fonctions. En recoupant les différents arbres ainsi obtenus pour une même fonction, les variables d'entrée qui semblent les plus pertinentes pour la fonction dans son ensemble sont sélectionnées. Chacun de ces différents ensembles de variables d'entrée est ensuite testé pour induire les règles permettant d'estimer la valeur des paramètres.

Induction des règles du système d'inférence floue

Les règles du système d'inférence floue sont également induites à partir d'arbres de décision flous (Voir chap. précédent). En effet, pour des valeurs de sous-ensembles flous de chacune des variables, il est possible de relier le nœud racine à chacune des feuilles. Ces chemins peuvent donc être interprétés comme des règles. Par exemple, par rapport à la figure 5.2, cela donne (Voir fig. 5.3) :

- Pour la feuille 2 : si x_i est A_i^2, alors y est C_2
- Pour la feuille 5 : si x_i est A_i^1 et si x_j est A_j^2 alors y est C_5.

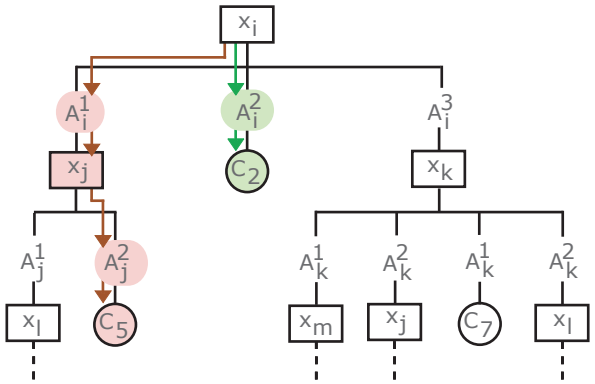


Figure 5.3 – Exemple de cheminement des règles dans un arbre de décision flou.

Les arbres sont ensuite *élagués*. C'est-à-dire que des nœuds sont transformés en feuilles, si la perte de performance qui en découle est faible. Cette procédure automatique dans *FisPro* permet de simplifier le système. Elle est basée sur la formule 5.4 qui représente le rapport de la performance de l'arbre sur sa complexité.

$$E_{(s,h)} = \frac{MCAE(s) - MC(h)}{N(s)NbF(s)} \tag{5.4}$$

où :

- MCAE le nombre d'exemples mal classés dans l'arbre élagué (sans le nœud s),
- MC le nombre d'exemples mal classés dans l'arbre entier,
- N(s) le nombre d'exemples qui passent par le nœud s
- NdF(s) le nombre de feuilles dans l'arbre sans le nœud s.

Le SIF ainsi généré (arbre élagué) est conservé si :

- ses règles permettent de prédire la valeur du paramètre avec un R^2 proche de 0,7.
- si le nombre de règles est inférieur à 15.

Si le SIF ne répond pas à ces critères, les étapes de construction du partitionnement flou des variables d'entrée ainsi que leur sélection sont de nouveau réalisées.

5.4 Transformation des valeurs possibles de paramètres en une information *a priori*

Une fois les différents SIF définis, il est possible de traduire les connaissances de l'utilisateur en une valeur possible des paramètres. Toutefois, ces valeurs ne sont pas

une source d'information *a priori* suffisante pour estimer les paramètres des cinétiques prédictives. Il faut également connaître leur variance-covariance.

5.4.1 Obtention de la matrice de variance-covariance des paramètres *a priori*

L'obtention de la matrice de variance-covariance des paramètres *a priori* se décompose en plusieurs étapes.

Dans un premier temps, 500 jeux de variables d'entrée des SIF sont simulés en réalisant un *boostrap* sur l'intervalle défini par l'utilisateur (Voir chap. 4.2.1). La variance-covariance des valeurs de sortie, inférées à partir des 500 jeux de variables d'entrée, est ensuite calculée.

Enfin, en additionnant la variance-covariance des valeurs possibles des paramètres à la variance des erreurs d'étalonnage des SIF, la matrice de variance-covariance des paramètres *a priori* est obtenue.

5.4.2 Modélisation des informations *a priori* par une loi de probabilité

Les informations *a priori* sur les paramètres (valeur possible et matrice de variances-covariance), sont employées selon la méthode présentée dans l'implémentation de la proposition scientifique (Voir chap. 4.2.1).

Les informations *a priori*, de chacun des paramètres, sont modélisées à partir d'une loi normale :
- de moyenne égale au paramètre obtenu à partir de la valeur centrale de l'intervalle défini comme entrée du SIF,
- et d'écart-type calculé à partir de la variance calculée précédemment.

La densité de probabilité est générée avec la fonction *normpdf* de Matlab® :

$$D_{\theta_p i} = normpdf(\theta_{pi}, \overline{\theta_p}, \sigma_{\theta_p}) \tag{5.5}$$

où
- $D_{\theta_p i}$ la valeur de la densité de probabilité associée à θ_p pour θ_{pi}
- θ_{pi} la valeur utilisée lors de l'estimation des paramètres,
- $\overline{\theta_p}$ la valeur du paramètre obtenue en sortie du SIF,
- σ_{θ_p} l'écart-type du paramètre obtenu en sortie du SIF.

5.5 Construction de la "boîte de *fitting*"

Les variances des erreurs d'estimation du modèle, $\sigma^2_{\varepsilon_{yi}}$ sont supposées normales. Elles sont toujours estimées à partir de la variance des observations, $\sigma^2_{y_i}$.

L'estimateur est celui du maximum de vraisemblance. La fonction objectif est définie comme le produit des valeurs de la fonction de densité de probabilité des résidus et des

priors.

Les paramètres sont estimés grâce à la fonction *fminsearch* de Matlab® :

$$\widehat{\theta_p} = fminsearch(-1 * FO(\mathcal{M}(.)), theta_p 0); \tag{5.6}$$

où

- $\widehat{\theta_p}$ valeur des paramètres qui optimisent la fonction objectif,
- $FO(\mathcal{M}(.))$ la fonction objectif correspondant au modèle,
- $theta_p 0$ la valeur d'initialisation de l'algorithme d'optimisation pour chacun des paramètres.

Cet algorithme s'appuie sur la méthode de *Nelder-Mead* qui utilise le concept de simplexe [57].

5.6 Calcul de la bande de confiance

500 jeux de paramètres, $\widehat{\theta_p^*}$, sont simulés selon une loi normale de moyenne $\widehat{\theta_p}$ et d'écart-type calculé à partir de la variance des paramètres *a priori*. Cette simulation tient également compte la matrice de variance-covariance des paramètres *a priori*.

500 jeux de mesures de suivi sont également simulés selon une loi normale de moyenne $\overline{y_i}$ et d'écart-type calculé à partir de la matrice de variance de la mesure σ^2_{yi}.

La bande de confiance est ensuite calculée comme expliquée dans le chapitre 4.2.3.

5.7 Réunion des deux boîtes de chacun des systèmes dans un même programme

L'utilisateur futur d'un tel outil n'est pas un théoricien ni un informaticien. Il est donc indispensable de construire une interface conviviale. Dans cette optique, les boîtes "experte" et "*fitting*" des systèmes "teneur en sucre" et "acidité totale" sont implémentées dans un *guide* Matlab®. Ainsi une interface graphique commune permet de charger les données de suivi et renseigner les connaissances expertes nécessaires au fonctionnement des différents SIF.

La figure 5.4 présente une impression d'écran de ce *guide*.

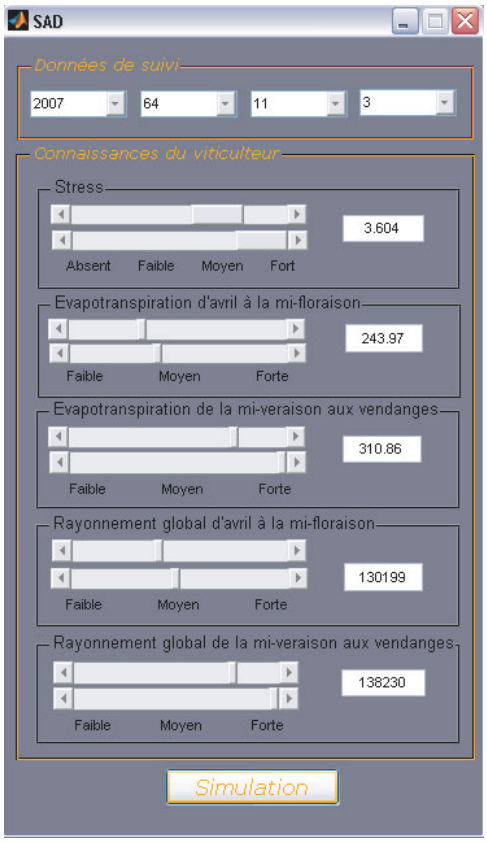


Figure 5.4 – *Guide* implémenté pour fusionner les différentes "boîtes" permettant de tracer les courbes d'évolution les plus probables de la teneur en sucre et de l'acidité totale.

Chapitre 6

Résultats et discussion

Sommaire

6.1 Induction de règles floues

Les méthodes exposées dans le chapitre précédent ont permis de construire les "boîtes expertes" des systèmes teneur en sucre et acidité totale. Dans cette partie, les variables d'entrée retenues ainsi que les règles inférées pour chacun des systèmes sont présentées.

6.1.1 Système "teneur en sucre"

Variables d'entrée retenues pour le système "teneur en sucre"

Trois variables d'entrée ont été sélectionnées pour le système "teneur en sucre". L'évapo-transpiration d'avril à la mi-floraison et celle de la mi-véraison aux vendanges et le niveau de contrainte hydrique. Les deux premières variables sont découpées en trois sous-ensembles flous, correspondant respectivement au label linguistique "faible", "moyen" et "fort". Le niveau de contrainte hydrique est, quant à lui, découpé en quatre sous-ensembles flous, correspondant respectivement au label linguistique "absente", "faible", "moyenne" et "forte". La valeur de teneur en sucre finale inférée sert d'entrée pour les SIF "Pente" et "$t_{1/2}$".

Le tableau 6.1 présente les variables sélectionnées comme entrée des SIF du système "teneur en sucre".

Variables d'entrée du SIF	Variable de sortie du SIF
Contrainte hydrique Evapo-transpiration d'avril à la mi-floraison Evapo-transpiration de la mi-véraison aux vendanges	Teneur en sucre finale
Contrainte hydrique Teneur en sucre finale Evapo-transpiration d'avril à la mi-floraison	Pente
Contrainte hydrique Teneur en sucre finale Evapo-transpiration d'avril à la mi-floraison	$t_{1/2}$

Table 6.1 – Variables d'entrée pour le système "teneur en sucre".

Règles générées pour le système "teneur en sucre"

Le tableau 6.2 présente le nombre de règles générées pour chacun des SIF du système "teneur en sucre" ainsi que leur coefficient de détermination et l'erreur standard d'étalonnage (Voir chap. 1.5.2).

Quelques règles générées pour le SIF permettant d'approcher la teneur en sucre finale sont reportées, sous forme linguistique, dans le tableau 6.3. Au premier abord, la

SIF	Nb. de règles générées	R^2	SE
Teneur en sucre finale	13	0,66	1,63 ($^\circ$Brix)
Pente	9	0,73	2,48 (jour)
$t_{1/2}$	8	0,78	5,3 (jour)

Table 6.2 – Nombre de règles générées pour les SIF du système "teneur en sucre", le coefficient de détermination (R^2) et l'erreur standard (SE).

conclusion des règles peut paraître choquante. Mais les SIF utilisés sont du type Sugeno. Ils fournissent donc un nombre net en sortie.

Règle	Contrainte hydrique	Evapo.1	Evapo.2	Teneur en sucre finale
1	Absente	Faible	Forte	16,83
2	Faible	Faible	Forte	23
3	Moyenne	Faible	Forte	21,23
4	Forte	Faible	Forte	18,92
5	Moyenne	Moyen	Forte	24,9

Table 6.3 – Exemples de règles générées pour le SIF teneur en sucre finale ($^\circ$Brix). Evapo.1 : évapo-transpiration d'avril à la mi-floraison, Evapo.2 : évapo-transpiration de la mi-véraison aux vendanges.

Ces règles s'écrivent, par exemple pour la n°1, ***Si*** *la contrainte hydrique est absente* ***ET si*** *l'évapo-transpiration d'avril à la mi-floraison est faible* ***ET si*** *évapo-transpiration de la mi-véraison aux vendanges est forte* ***Alors*** *la teneur en sucre finale est de 16,83* $^\circ$*Brix.*

Les règles 1 à 4 conduisent toutes trois à des teneurs en sucre finales différentes. Elles diffèrent par une variable : le niveau de contrainte hydrique. Quand la contrainte hydrique est absente ou forte, la teneur en sucre finale sera moins élevée que pour une contrainte hydrique faible ou moyenne. De la même manière, les règles 3 et 5 diffèrent par l'évapo-transpiration d'avril à la mi-floraison. Quand l'évapo-transpiration d'avril à la mi-floraison est faible, la teneur en sucre finale sera moins élevée que pour une l'évapo-transpiration moyenne.

Ces quelques règles présentées dans le tableau 6.3 sont bien en accord avec l'effet de ces variables sur la teneur en sucre finale.

De manière plus générale, après discussion avec un expert, l'ensemble des règles générées pour le système "teneur en sucre" sont cohérentes (elles ne se contredissent pas) et sont également en accord avec la littérature.

6.1.2 Système "acidité totale"

Variables d'entrée retenues pour le système "acidité totale"

Trois variables d'entrée ont été sélectionnées pour le système "acidité totale". Le rayonnement global d'avril à la mi-floraison et celui de la mi-véraison aux vendanges et le niveau de contrainte hydrique. Les deux premières variables sont découpées en trois sous-ensembles flous, correspondant respectivement au label linguistique "faible", "moyen" et "forte". Comme pour le système "teneur en sucre", le niveau de contrainte hydrique est, quant à lui découpé, en quatre sous-ensembles flous.

Le tableau 6.4 présente les variables sélectionnées comme entrée des SIF du sytème "acidité totale".

Variables d'entrée du SIF	Variable de sortie du SIF
Contrainte hydrique Rayonnement global d'avril à la mi-floraison Rayonnement global de la mi-véraison aux vendanges	Acidité totale finale
Contrainte hydrique Acidité totale finale Rayonnement global d'avril à la mi-floraison Rayonnement global de la mi-véraison aux vendanges	Taux de décroissance max.
Contrainte hydrique Acidité totale finale Rayonnement global de la mi-véraison aux vendanges	$t^{\pi}_{1/2}$

Table 6.4 – Variables d'entrée pour le système "acidité totale".

Règles générées pour le système "acidité totale"

Le tableau 6.5 présente le nombre de règles générées pour chacun des SIF du sytème "acidité totale" ainsi que leur coefficient de détermination et l'erreur standard d'étalonnage (Voir chap. 1.5.2).

SIF	Nb. de règles générées	R^2	SE
Acidité totale finale	6	0,53	0,44 ($g.l^{-1} H_2SO_4$)
Taux de décroissance	12	0,74	4,51 ($jour^{-1}$)
$t^{\pi}_{1/2}$	9	0,64	8,8 ($jour^{-1}$)

Table 6.5 – Nombre de règles générées pour les SIF du système "acidité totale", le coefficient de détermination (R^2) et l'erreur standard (SE).

Un couple de règle n'a pas pu être interprété :
- **Si** la contrainte hydrique est faible, moyenne ou forte **ET si** le rayonnement global d'avril à la mi-floraison est moyen **ET si** le rayonnement global de la mi-véraison aux vendanges est faible **Alors** l'acidité totale est de 3,57 $g.l^{-1}H_2SO_4$.
- **Si** la contrainte hydrique est faible, moyenne ou forte **ET si** le rayonnement global d'avril à la mi-floraison est fort **ET si** le rayonnement global de la mi-véraison aux vendanges est faible **Alors** l'acidité totale est de 4,41 $g.l^{-1}H_2SO_4$.

Hormis le couple de règles pour lequel l'expert n'avait pas d'interprétation, de la même manière que pour le sytème "teneur en sucre", les règles générées pour le système "acidité totale" sont en accord avec la littérature.

6.1.3 Conclusion sur l'induction des règles

Les différents SIF générés permettent bien de retrouver des lois en accord avec les experts de la viticulture. Ces systèmes sont également capables de reproduire les relations entre les entrées et les sorties de façon satisfaisante.

Le caractère interprétable des règles facilite la compréhension de ces SIF. Ils ont ainsi pu être le support d'un dialogue avec un expert du domaine dans un but de validation.

De plus, contrairement à des systèmes de type boîte noire, les méthodes qui visent à produire des systèmes interprétables ne peuvent gérer qu'un nombre limité de variables. Ce fait souligne l'importance de la sélection des variables d'entrée. Dans le cas présent, cette sélection met en relief l'importance de la contrainte hydrique sur les deux critères de maturité étudiés. En effet, il s'agit de la seule variable d'entrée présente dans tous les SIF générés.

Toutefois, il faut avoir conscience que la démarche d'apprentissage est limitée par le contenu de la base d'exemples : plus celle-ci reflétera l'ensemble des possibilités de comportement de la vigne, plus le modèle sera général, c'est-à-dire utilisable dans une large gamme de situations. Dans notre situation, le faible effectif des données d'apprentissage et le caractère mono-cépage de la base de données interdit de généraliser les règles induites à l'ensemble des cépages et des régions viticoles.

De plus, aucune explication n'a pu être trouvée dans la littérature afin de comprendre le rejet de la période comprise entre la mi-floraison et la mi-véraison lors de la sélection des variables d'entrée. Différents essais ont été menés afin de la prendre en compte, mais sans succès (remplacement de la période d'avril à mi-floraison par avril à mi-véraison et remplacement de mi-véraison à la vendange par mi-floraison à la vendange).

6.2 Performance générale des systèmes et nécessité de fusion des sources d'information

La performance des systèmes "teneur en sucre" et "acidité totale" est évaluée dans cette partie.

Le critère de jugement retenu pour caractériser la qualité globale des cinétiques est l'erreur standard moyenne : ESM (Voir éq. 6.1). Ce critère a l'avantage de s'exprimer dans la même unité de mesure que celle du phénomène étudié. Il est calculé pour l'ensemble d'apprentissage et pour l'ensemble de test.

$$ESM = \sqrt{\frac{\sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_i - \widehat{y}_i)^2}{\sum_{j=1}^{p} n_j}} \tag{6.1}$$

où :

- p le nombre de suivis de l'ensemble d'apprentissage ou de test,
- n_j le nombre d'observations du suivi j,
- y_i la valeur des observations,
- $\widehat{y}_i$ la valeur prédite par le système.

Dans la pratique, les erreurs standards moyennes sont calculées sur la globalité des cinétiques modélisées. C'est-à-dire que toutes les mesures de suivi, celles appartenant au passé et celles appartenant au futur, sont prises en compte dans ce calcul. En effet, les cinétiques prédictives sont normalement formées de deux parties (Voir fig. 6.1) :

- Une première partie correspond au passé. Il s'agit de la partie connue de la cinétique. Elle est prise en compte lors de la modélisation.
- Une seconde partie correspond au futur. Il s'agit de la partie prédictive à proprement parlé. Elle est normalement inconnue et est entièrement modélisée.

La partie prédictive des cinétiques débute en moyenne un mois avant les vendanges. Cela dépend : de la date de début du suivi et de la fréquence des mesures.

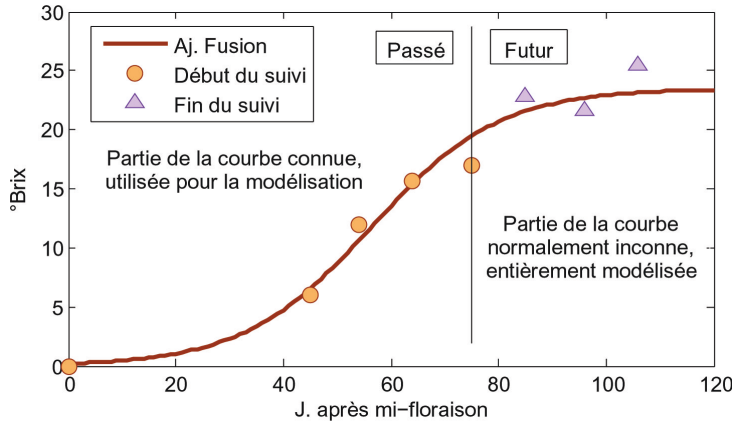


Figure 6.1 – Cinétique prédictive.

6.2.1 Performance générale des systèmes

Les cinétiques prédictives sont comparées aux cinétiques réelles afin d'évaluer les performances générales des systèmes "teneur en sucre" et "acidité totale".

Les cinétiques réelles sont modélisées en effectuant une régression sur les jeux complets de suivis (Voir chap. 5.3.1). Les cinétiques prédictives sont modélisées, quant à elles, en utilisant les trois premières mesures de suivi et les connaissances expertes. Les erreurs standards moyennes sont ensuite calculées (Voir chap. précédent) pour ces deux types de modélisation. Le tableau 6.6 présente les résultats obtenus pour l'ensemble d'apprentissage et l'ensemble de test.

Suivi	Teneur en sucre		Acidité totale	
	ESM. Reg.	ESM. CP.	ESM. Reg.	ESM. CP.
Ens. app.	0,69	1,89	0,83	2,45
Ens. test	1,94	1,87	1,13	1,80

Table 6.6 – Erreurs standards moyennes (ESM.) des systèmes "teneur en sucre" ($^\circ$Brix) et "acidité totale" ($g.l^{-1}H_2SO_4$) calculées sur les cinétiques réelles (Reg.) et sur les cinétiques prédictives (CP.) pour l'ensemble d'apprentissage (Ens. app.) et l'ensemble de test (Ens. test).

Pour l'ensemble de test, les erreurs standards moyennes calculées sur les cinétiques prédictives sont semblables à celles calculées sur les cinétiques réelles. Mais elles sont différentes pour l'ensemble d'apprentissage. Ce fait est à mettre en rapport avec la sélection réalisée sur les suivis de l'ensemble d'apprentissage. Les suivis ne permettant pas une bonne régression ont été supprimés (voir chap. 5.3.1). A l'inverse, aucune sélection n'a été faite sur l'ensemble de test. Par conséquent, il est normal que la régression fournisse de meilleurs résultats pour l'ensemble d'apprentissage que la modélisation des cinétiques prédictives.

Ces résultats démontrent les capacités de la méthode développée pour modéliser les cinétiques prédictives en utilisant conjointement les mesures de suivi et les connaissances expertes.

Afin d'évaluer spécifiquement la capacité prédictive de la méthode, les erreurs standards moyennes sont calculées seulement sur la partie future des cinétiques prédictives (Voir fig. 6.1). Ces erreurs standards moyennes sont comparées à celles calculées précédemment sur les cinétiques prédictives dans leur globalité. Le tableau 6.7 présente ces résultats pour l'ensemble d'apprentissage et l'ensemble de test.

Ces résultats montrent que sur la partie future des cinétiques prédictives, les erreurs standards moyennes ne sont pas plus importantes que celles calculées sur les cinétiques prédictives dans leur globalité.

Suivi	Teneur en sucre		Acidité totale	
	ESM. CP.	ESM. PF.	ESM. CP.	ESM. PF.
Ens. app.	1,89	1,99	2,45	2,80
Ens. test	1,87	1,61	1,80	2.04

Table 6.7 – Erreurs standards moyennes (ESM.) des systèmes "teneur en sucre" ($^\circ$ Brix) et "acidité totale" ($g.l^{-1}H_2SO_4$) calculées sur la partie future des cinétiques prédictives (PF.) et sur les cinétiques prédictives dans leur globalité (CP.), pour l'ensemble d'apprentissage (Ens. app.) et l'ensemble de test (Ens. test).

6.2.2 Nécessité de fusion des sources d'information

La nécessité de fusionner les sources d'information (les mesures et les connaissances expertes) est évaluée dans cette partie. Pour cela, les erreurs standards moyennes des deux ensembles de suivi (apprentissage et test) sont calculées pour les cinétiques prédictives obtenues :

- par régression à partir des seules premières mesures de suivi,
- à partir des seules connaissances expertes,
- à partir de la fusion des premières mesures de suivi et des connaissances expertes.

Système "teneur en sucre"

Le tableau 6.8 présente les valeurs de l'erreur standard moyenne des cinétiques prédictives de teneur en sucre obtenues dans les conditions explicitées précédemment. Ces résultats démontrent l'impossibilité de construire des cinétiques prédictives de teneur en sucre à partir des seules mesures de suivi ou des seules connaissances expertes.

Suivi	ESM. M.	ESM. C.	ESM. F.
Ens. app.	3,74	3,61	1,89
Ens. test	3,21	4,65	1,87

Table 6.8 – Erreurs standards moyennes de l'ensemble d'apprentissage (Ens. app.) et de l'ensemble de test (Ens. test) pour la teneur en sucre ($^\circ$ Brix), obtenues par : régression à partir des seules premières mesures (M.), à partir des seules connaissances expertes (C.) et à partir de la fusion de ces deux sources d'information (F.).

Les figures suivantes présentent deux exemples de cinétiques prédictives de teneur en sucre pour illustrer la nécessité de fusionner les mesures et les connaissances expertes.

Les cinétiques de teneur en sucre, obtenues à partir des seules mesures de suivi, débutent au bon moment (courbes en trait pointillé). Mais ces dernières se poursuivent de manière aléatoire. En effet, si le point d'inflexion de la courbe n'est pas défini à partir des premières mesures, la teneur en sucre tend vers des valeurs très importantes (Voir fig. 6.2.a). A l'inverse, si un point d'inflexion se forme de manière prématurée, suite par exemple à une importante modification temporaire des conditions météorologiques (pluie,

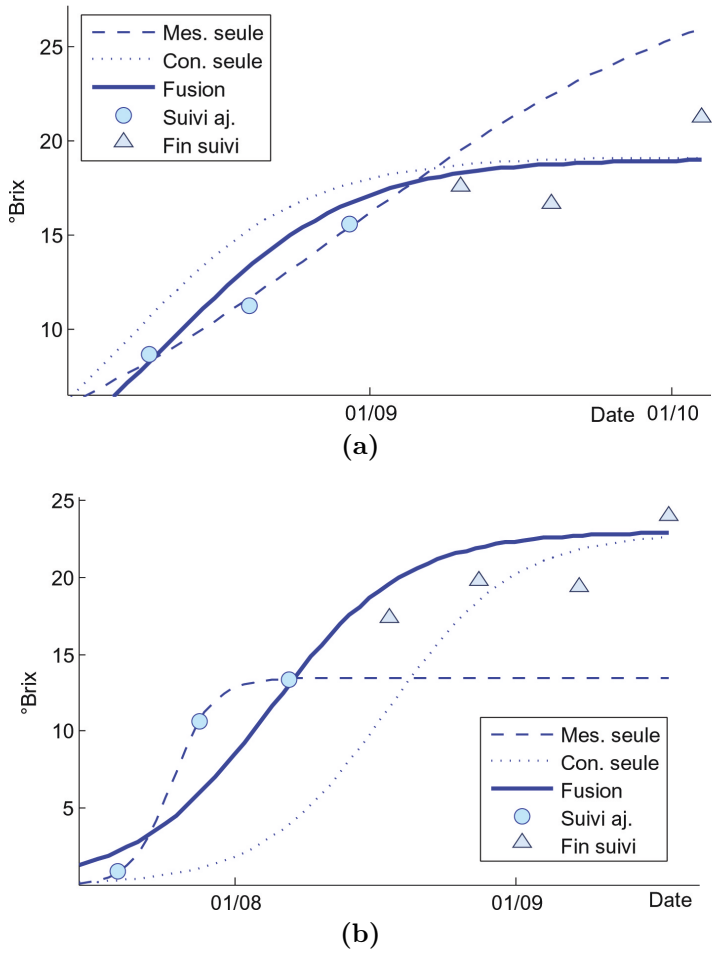


Figure 6.2 – Exemples de cinétiques prédictives de teneur en sucre issues de l'ensemble de test, obtenues par régression à partir des seules premières mesures de suivi (Mes. seule), des seules connaissances expertes (Con. seule) et à partir de la fusion des deux sources d'information (Fusion).

période plus fraîche, etc.), la teneur en sucre tend vers des valeurs très faibles (Voir fig. 6.2.a).

Les cinétique obtenues à partir des seules connaissances expertes tendent vers des teneurs en sucre correctes (courbes en trait discontinu). Mais ces dernières ne débutent pas forcément au bon moment. Les courbes peuvent débuter de manière prématurée (Voir fig. 6.2.a), ou avec du retard (Voir fig. 6.2.b). En effet, les connaissances expertes fournissent des valeurs de paramètres "standards" qu'il faut recaler à la réalité de la parcelle.

Dans les deux cas, l'utilisation d'une seule des deux sources d'information (mesures ou connaissances expertes) conduit à des cinétiques prédictives erronées.

A l'inverse, les cinétiques prédictives obtenues à partir de la fusion des deux sources d'information (mesures et connaissances expertes) débutent au bon moment et tendent vers des teneurs en sucre correctes (courbes en trait gras). La fusion permet de recaler les connaissances expertes grâce aux mesures. De la même manière, les informations concernant le point d'inflexion et la teneur en sucre finale, fournies par les mesures, sont corrigées par les connaissances expertes. En effet, les connaissances expertes portent sur

de grandes périodes et ne sont donc pas assujetties aux fluctuations temporaires de la météo ou de l'état de contrainte hydrique de la parcelle.

Système "acidité totale"

Le tableau 6.9 présente les valeurs de l'erreur standard moyenne des cinétiques prédictives d'acidité totale obtenues dans les conditions explicitées précédemment. De la même manière que pour le système "teneur en sucre", ces résultats montrent l'impossibilité de construire des cinétiques prédictives d'acidité totale à partir des seules mesures de suivi ou des seules connaissances expertes.

Suivi	ESM. M.	ESM. C.	ESM. F.
Ens. app.	4,02	6,12	2,45
Ens. test	4.02	5,05	1,80

Table 6.9 – Erreurs standards moyennes de l'ensemble d'apprentissage (Ens. app.) et de l'ensemble de test (Ens. test) pour la teneur en sucre ($g.l^{-1}H_2SO_4$), obtenues par : régression à partir des seules premières mesures (M.), à partir des seules connaissances expertes (C.) et à partir de la fusion de ces deux sources d'information (F.).

Les figures suivantes présentent deux exemples de cinétiques prédictives d'acidité totale pour illustrer la nécessité de fusionner les mesures et les connaissances expertes.

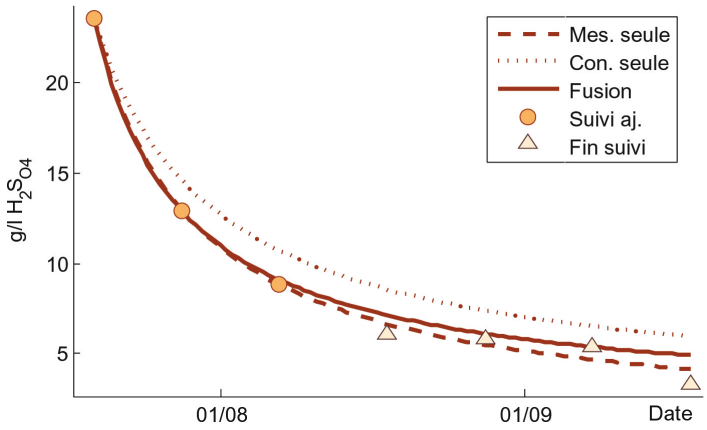
Dans les deux cas, comme pour la teneur en sucre, l'utilisation d'une seule des sources d'information (mesures ou connaissances expertes) conduit à des cinétiques prédictives d'acidité totale erronées.

A l'inverse, les cinétiques prédictives obtenues à partir de la fusion des deux sources d'information (mesures et connaissances expertes) décroissent correctement et tendent vers des valeurs d'acidité totale finale correctes (courbes en trait gras). La fusion permet de recaler les connaissances expertes grâce aux mesures (Voir fig. 6.3.a). De la même manière, les informations concernant le taux de décroissance, fournies par les mesures, sont corrigées par les connaissances expertes (Voir fig. 6.3.b).

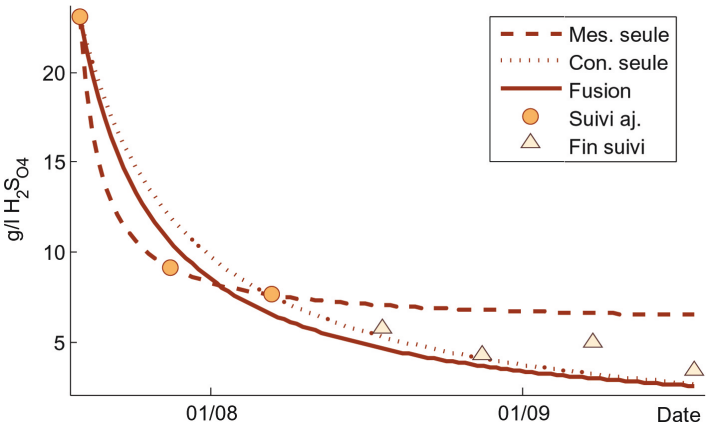
6.2.3 Conclusion sur les performances du système

Ces résultats mettent en évidence l'impossibilité d'obtenir des cinétiques prédictives de bonne qualité à partir des seules mesures de suivi ou des seules connaissances expertes. A l'inverse, les cinétiques prédictives obtenues en fusionnant ces deux sources d'information présentent de biens meilleurs résultats selon le critère de l'erreur standard moyenne.

Toujours selon le critère de l'erreur standard moyenne, la qualité globale des cinétiques issues de la fusion des mesures et des connaissances expertes peut être considérée comme tout à fait satisfaisante au regard des erreurs standards moyennes obtenues par régression.



(a)



(b)

Figure 6.3 – Exemples de cinétiques prédictives d'acidité totale issues de l'ensemble de test, obtenues par régression à partir des seules premières mesures de suivi (Mes. seule), des seules connaissances expertes (Con. seule) et à partir de la fusion des deux sources d'information (Fusion).

Toutefois, ces résultats sont à mettre en relation avec les tailles des ensembles d'apprentissage de test qui sont limitées. Pour confirmer ces performances, il serait intéressant, dans un premier temps, d'appliquer cette méthode à de nombreux de suivis issus du même cépage, la Syrah et de la même région viticole.

6.3 Influence de l'incertitude des connaissances expertes

Des modélisations sont réalisées avec deux niveaux d'incertitude afin de mettre en évidence l'influence de l'incertitude des connaissances expertes sur la cinétique prédictive et sur la bande de confiance. Ces deux niveaux correspondent à une faible incertitude (les intervalles de l'entrée des SIF correspondent à environ ±1% de la valeur centrale) et forte incertitude (les intervalles de l'entrée des SIF correspondent à environ ±20% de la valeur centrale).

6.3.1 Influence de l'incertitude des connaissances expertes sur la cinétique prédictive

Les figures suivantes présentent des cinétiques de teneur en sucre obtenues à partir de la fusion des deux même sources d'information mais avec les niveaux d'incertitude définis précédemment sur les connaissances expertes.

La figure 6.4.a est la cinétique obtenue avec une faible incertitude. Dans ce cas, les connaissances expertes ont un poids plus important lors de la recherche des paramètres les plus probables. La figure 6.4.b est la cinétique obtenue avec une forte incertitude. Dans ce cas, à l'inverse du précédent, les mesures de suivi ont un poids plus important lors de la recherche des paramètres.

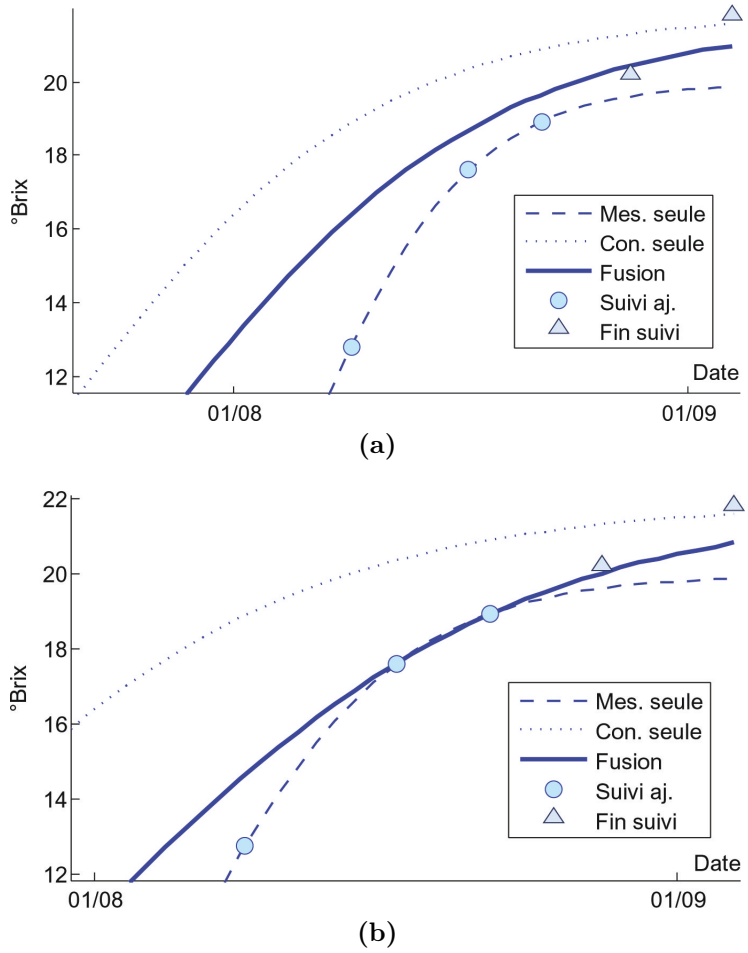


Figure 6.4 – Cinétiques prédictives de teneur en sucre (° Brix) obtenues avec (a) une faible incertitude et (b) une forte incertitude sur les connaissances expertes.

L'influence de l'incertitude des connaissances expertes est nettement visible en comparant ces cinétiques prédictives (courbe en trait gras). Il est important de rappeler qu'elles sont obtenues à partir de la fusion des deux même sources d'information mais avec des niveaux d'incertitude différents sur les connaissances expertes. Pour un fort niveau d'incertitude sur les connaissances expertes (Voir fig. 6.4.b), la cinétique prédictive est beaucoup plus proche de la cinétique obtenue à partir des seules mesures de

suivi (courbes en trait pointillé). A l'inverse, pour un faible niveau d'incertitude sur les connaissances expertes (Voir fig. 6.4.a), la cinétique prédictive se décolle de la cinétique obtenue à partir des seules mesures de suivi et se rapproche de celle obtenue à partir des seules connaissances expertes (courbes en trait discontinu).

Ces cinétiques démontrent la capacité de la méthode à prendre en compte l'incertitude des connaissances expertes dans l'estimation des paramètres des cinétiques prédictives.

L'une des conséquences pratiques de cette capacité, associée aux avantages de la fusion des sources d'information, est la faculté d'absorber, en partie, les erreurs pouvant être faites sur les connaissances expertes. Les figures suivantes illustrent cette capacité avec un exemple pour l'acidité totale.

Les deux cinétiques prédictives ont été modélisées avec des connaissances expertes erronées, associées aux deux niveaux d'incertitude utilisés précédemment. La figure 6.5.a est la cinétique obtenue avec une faible incertitude. La figure 6.5.b est la cinétique obtenue avec une forte incertitude. Dans les deux cas, la cinétique prédictive est plus proche des mesures (avantage de la fusion des sources d'information), mais si une incertitude est affectée à ces connaissances erronées, la courbe se rapproche plus des mesures de suivi.

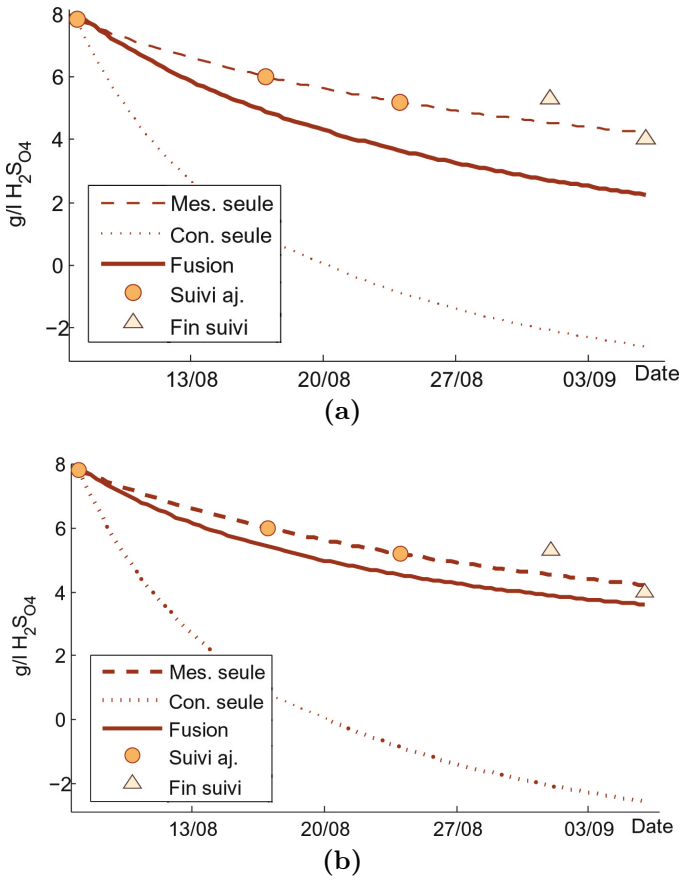


Figure 6.5 – Cinétiques prédictives d'acidité totale ($g.l^{-1}H_2SO_4$) obtenues en tenant compte des premières mesures et de connaissances expertes erronées affectées (a) d'une faible incertitude et (b) d'une forte incertitude.

6.3.2 Influence de l'incertitude des connaissances expertes sur la bande de confiance

La prise en compte des incertitudes, dans la construction des bandes de confiance des cinétiques prédictives de teneur en sucre et d'acidité totale, est évaluée dans cette partie.

Le critère de jugement retenu pour évaluer la taille de la bande de confiance est l'aire définie par cette bande. Le tableau 6.10 présente les aires moyennes calculées sur l'ensemble d'apprentissage et l'ensemble de test, pour la teneur en sucre et l'acidité totale, dans les conditions édictées précédemment.

Suivi	Teneur en sucre		Acidité totale	
	Aire FI.	Aire II	Aire FI.	Aire II.
Ens. app.	294	334	818	1113
Ens. test	299	331	318	328

Table 6.10 – Aires moyennes formées par la bande de confiance, sur l'ensemble d'apprentissage (Ens. app.) et l'ensemble de test (Ens. test) pour la teneur en sucre (S., °Brix) et l'acidité totale (AT., $g.l^{-1}H_2SO_4$), calculées sur les cinétiques prédictives obtenues avec des connaissances affectées d'une faible incertitude (FI.) d'une forte incertitude (II.).

La faible différence au niveau de ces aires peut paraître très surprenante. Mais cette différence peut s'expliquer par deux raisons.

La méthode développée a pour objectif de prendre en compte l'incertitude des différentes sources d'information (utilisation de l'estimateur de maximum de vraisemblance associé à des lois de probabilité des paramètres). Par conséquent, la méthode permet d'absorber une grande partie des incertitudes. Cette capacité d'absorption est mise en évidence en comparant par exemple la variance des paramètres estimés à celle des paramètres *a priori*. Le tableau 6.11 présente celles du système teneur en sucre.

Variances	Ens. App.			Ens. test		
	$T_{S_{max}}$	$t_{1/2}$	*Pente*	$T_{S_{max}}$	$t_{1/2}$	*Pente*
Paramètres *a priori*	3,28	29,48	6,30	3,99	54,08	8,42
Paramètres estimés	1,06	4,26	1,89	1,28	5,88	2,30

Table 6.11 – Variances des paramètres estimés et celles des paramètres *a priori* des cinétiques de teneur en sucre, pour l'ensemble d'apprentissage et de test.

L'incertitude des connaissances expertes dans son ensemble est traduite par la matrice de variance-covariance des paramètres *a priori* (Voir chap. 4.2.3). La majeure partie de la variance des paramètres *a priori* est due aux erreurs de prédiction des SIF. Par exemple pour le système teneur en sucre, les incertitudes directement liées à l'utilisateur représentent en moyenne :

– 1% de la variance des paramètres *a priori* pour une faible incertitude,
– 29% de la variance des paramètres *a priori* pour une forte incertitude.

Les figures présentent deux exemples de bandes de confiance calculées pour une cinétique prédictive de teneur en sucre (Voir fig. 6.6.a) et pour une cinétique prédictive d'acidité totale (Voir fig. 6.6.b).

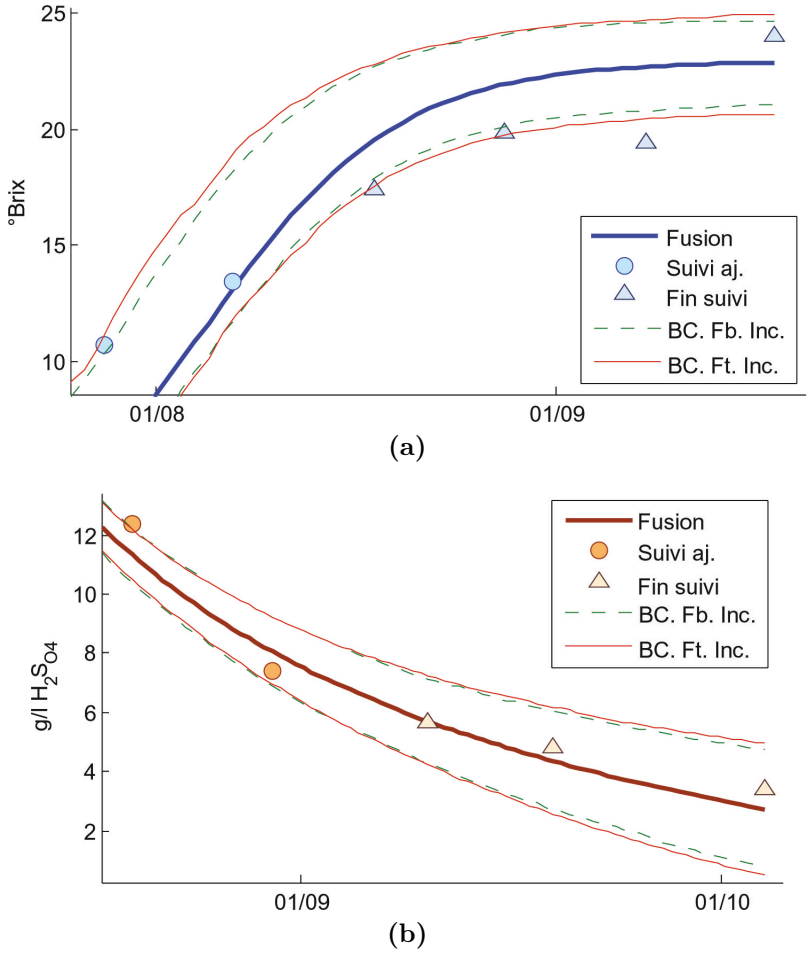


Figure 6.6 – Cinétiques prédictives (a) de teneur en sucre (S., °Brix) et (b) d'acidité totale ($g.l^{-1}H_2SO_4$) obtenues avec des connaissances expertes affectées d'une faible incertitude (Fb.) et d'une forte incertitude (Ft.).

6.3.3 Conclusion sur l'influence de l'incertitude des connaissances expertes

Ces résultats confirment que l'incertitude est bien prise en compte par la méthode proposée au niveau de la cinétique prédictive.

Toutefois, l'influence de l'incertitude des connaissances expertes sur la bande de confiance est très limitée car :
– l'incertitude des connaissances expertes liée à l'utilisateur est partiellement "masquée" par les erreurs de prédiction des SIF,

– l'incertitude des connaissances expertes dans son ensemble est en grande partie absorbée par la méthode d'estimation des paramètres.

Chapitre 7

Conclusion et perspectives

Cette thèse s'inscrit dans le projet Spectron™ initié par l'UMR-ITAP du Cemagref de Montpellier et la Société Pellenc SA. Le cahier des charges portait sur :

- le développement d'un procédé de mesure, rapide et non destructif, de suivi de maturité,
- le développement d'un outil d'aide à la décision fondé sur les informations recueillies grâce à ce capteur.

Les deux axes de ce projet ont été développés en parallèle.

Ce capteur est un spectrophotomètre portable capable de fournir des évaluations parcellaires de taux de sucre, d'acidité totale, de teneur en eau et de potentiel d'extraction d'anthocyanes. Il peut fournir un suivi temporel très fin de la parcelle pour ces quatre critères de maturité, ou de tout autre indice basé sur ces critères. La seule limite est le temps que souhaite consacrer le viticulteur au suivi des parcelles. Le développement du capteur s'est déroulé depuis les essais de différents prototypes jusqu'aux tests d'un produit de présérie.

Le développement de l'outil d'aide à la décision, qui accompagne le viticulteur dans l'exploitation des données fournies par le Spectron™ était l'objectif opérationnel de cette thèse.

L'analyse de cet objectif dans son contexte a mis en évidence que l'aide la plus utile consistait à construire au plus tôt des cinétiques d'indices de maturité, basées sur les mesures du Spectron™ et sur les connaissances et les hypothèses fournies par l'utilisateur.

Une analyse détaillée du problème posé a soulevé trois questions :

- la prise en compte de l'imperfection des mesures dans l'estimation des paramètres d'un modèle,
- l'utilisation d'une information *a priori* sur la valeur des paramètres d'un modèle lors de leur estimation, quel que soit son cadre de représentation (probabiliste ou possibiliste).
- la construction d'une bande de confiance autour de la courbe résultat.

Un état de l'art ciblé sur ces trois questions a montré :

- les capacités de la méthode d'ajustement bayésienne afin d'estimer la valeur des paramètres d'un modèle, en utilisant conjointement les mesures de suivi et des informations *a priori* sur les paramètres, en tenant compte de leur incertitude respective.
- la possibilité d'obtenir une information *a priori*, utilisable dans le cadre bayésien (probabiliste), à partir d'une distribution de possibilité (fournie à partir des connaissances expertes).
- les limites des méthodes de calcul de la bande de confiance. Ces calculs reposent en grande partie sur les résidus et fournissent une bande de confiance dépendante de la variable explicative. Aucune ne tient compte explicitement de l'incertitude que peut avoir l'utilisateur au sujet de ses connaissances.
- les capacités de la matrice de variance-covariance des paramètres pour transporter l'incertitude des informations à la cinétique prédictive et à la bande de confiance.

Cet état de l'art s'est traduit par une publication en cours de soumission à *Computer and Electronics in Agriculture*.

Sur la base de ce constat, une proposition scientifique a été émise. Elle propose une méthode basée sur :
- l'utilisation d'un système à base de règles pour traduire les connaissances expertes en une valeur possible des paramètres du modèle et leur incertitude,
- l'estimation des paramètres en associant les mesures aux valeurs des paramètres précédemment estimés dans un système de fusion bayésien,
- et la construction autour de cette cinétique d'une bande de confiance basée sur l'incertitude totale du modèle.

Cette approche méthodologique a ensuite été implémentée et appliquée à la construction de cinétiques prédictives de deux paramètres mesurés par le Spectron™ : la teneur en sucre et l'acidité totale.

La mise en œuvre de la méthode proposée a ainsi permis de confirmer sa pertinence, ainsi que ses performances.

Sa capacité à générer des résultats interprétables et techniquement exploitables a été soulignée. Les règles générées ont pu être validées par un expert de la profession. Cette validation a permis d'affirmer que les règles sont bien en accord avec les grands principes de la viticulture.

La modélisation de cinétiques prédictives sur des jeux de données test a conduit à des résultats satisfaisants, selon le critère de l'erreur standard moyenne. Ces résultats ont également montré la nécessité de fusionner les mesures de suivi et les connaissances expertes.

Les performances et les limites de l'approche proposée pour prendre en compte l'incertitude ont également été mises en évidence. Les incertitudes ont bien été traduites au niveau des cinétiques prédictives. Toutefois, l'influence de ces incertitudes au niveau de la bande de confiance est moins visible. En effet, l'incertitude liée à l'utilisateur est masquée par celle des SIF et les incertitudes dans leur ensemble sont partiellement absorbées par la méthode d'estimation des paramètres.

Cette méthode peut être présentée comme une alternative aux approches habituellement utilisées pour modéliser le comportement de la parcelle, qui nécessitent souvent l'emploi de modélisation complexe. En effet, la modélisation des cinétiques sous la forme de fonctions explicites, associées aux possibilités laissées au viticulteur de retranscrire ses connaissances en une valeur possible de paramètres, offre un cadre de développement interprétable, ouvert et évolutif. Dans la pratique, ce cadre de développement permet par exemple à l'utilisateur de modifier les règles qui ne sont pas en accord avec son expérience. De plus, le système étant ouvert et les règles étant exprimées de manière linguistique, il lui est possible d'ajouter de nouvelles règles qui lui sont propres.

Les avantages de cette méthode sont :
- l'utilisation conjointe d'information de suivi et de connaissances expertes pour la construction de cinétiques prédictives,
- le faible nombre de mesures de suivi nécessaires pour initialiser des simulations,
- sa capacité à gérer l'incertitude sur les sources d'information,
- de fournir une information qui permet le diagnostic en temps réel de la maturation,
- son implémentation facile et pratique avec un temps de calcul raisonnable (quelques secondes).

Cette méthode, associée au Spectron™, constitue donc à l'évidence un apport potentiel à l'exploitation des données de suivi. L'implémentation de la méthode proposée peut rapidement évoluer vers le développement d'un logiciel complet. Ce couple " capteur - service " répondrait aux attentes des professionnels de la viticulture puisqu'il leur permettrait :
- de connaître *in situ* et très rapidement l'état de maturation d'une parcelle,
- de simuler très tôt l'évolution de la maturation et d'en tirer les conclusions concernant sa gestion,
- d'estimer le degré de confiance associé à ces simulations.

Ainsi, à partir de l'ensemble des cinétiques, le viticulteur peut approcher la date des vendanges en fonction de ses objectifs et du type de vin qu'il cherche à produire.

Ces travaux ont donc permis de mettre en lumière le potentiel de la méthode développée. Cependant, cette dernière comporte certaines simplifications rendues nécessaires par son implémentation. De plus, les données expérimentales utilisées ont été limitées à un cépage et une région viticole. En effet, l'objectif premier de ces travaux était de valider la méthode et non de construire un outil achevé. Cet objectif étant atteint, il est alors possible de proposer différentes perspectives afin de faire évoluer cette méthode.

Le développement des systèmes (c.-à-d. les "boîtes de *fitting*" et les "boîtes expertes") pour le potentiel d'extraction d'anthocyanes et la teneur en eau doit être mis en œuvre. Ainsi, les capacités du Spectron™ pourront être pleinement exploitées.

Il serait intéressant de réaliser les simulations sur différentes périodes afin de déduire à partir de quel moment les estimations relatives à un suivi deviennent stables. De la même manière, des simulations avec différents niveaux d'incertitude sur les connaissances expertes permettraient de savoir à partir de quel degré d'incertitude les estimations deviennent stables pour un même nombre de points de suivi.

Afin de valider la généricité de la méthode, elle doit être appliquée sur d'autres cépages et d'autres suivis issus de régions viticoles différentes. Il faudrait pour cela trouver des données de suivi, liées à des connaissances expertes, suffisamment nombreuses pour réaliser les phases d'apprentissage et de test. Une autre approche consisterait à appliquer la méthode en redéfinissant les partitions des concepts linguistiques avec un expert de la région viticole. Les valeurs sous-jacentes par exemple aux concepts " faible évapotranspiration " ou " fort rayonnement global " sont différentes selon les régions mais leurs effets relatifs sur la maturation sont similaires.

Deux perspectives peuvent être envisagées pour le type de règles mis en œuvre dans les SIF :

- Si le choix d'un apprentissage supervisé devait perdurer (gain de temps dans la génération des règles, possibilité de travailler avec un grand nombre de données), la base d'apprentissage devra d'une part refléter l'ensemble des conditions de maturation possibles et d'autre part comprendre différents cépages.
- Des travaux pour implémenter des règles implicatives seraient intéressants. En effet, elles semblent permettre une meilleure gestion des incertitudes concernant les connaissances expertes.

Dans tous les cas, il n'est pas garanti qu'un système générique à l'ensemble des cépages et des régions puisse être défini.

Les applications de cette thèse ont porté sur le taux de sucre et l'acidité totale. Mais ce ne sont pas les seuls critères utiles pour approcher la maturation. Par la suite, il conviendrait d'implémenter des modèles de cinétiques d'indices de maturité plus performants, comme par exemple la quantité de sucre par baie.

L'utilisation d'autres connaissances ou hypothèses serait à explorer comme par exemple l'analyse sensorielle des baies, qui constitue un outil d'évaluation de la maturité. Le principale problème de l'analyse sensorielle est sa très forte dépendance aux "goûteurs", à son humeur, son environnement, etc. De plus, il existe plusieurs référentiels couramment utilisés, comme la méthodologie ASDQ [1] de l'ICV. De la même manière, le risque de maladie serait une donnée intéressante à utiliser. Mais là aussi, il existe différentes manières d'apprécier ce risque en fonction des cépages, des porte-greffes, des clones, etc.

Les cinétiques sont construites de manière indépendante les unes des autres. Un travail sur l'aspect multi-varié des cinétiques lors de leur construction permettrait de prendre en compte leur interdépendance.

Enfin, la généricité de la méthode développée, associée au potentiel de la spectrométrie simplifiée (capteur portable), offre la possibilité de traiter d'autres problèmes que ceux liés à la viticulture. En effet, dans un monde qui évolue sans cesse vers une exigence incontournable de qualité, les outils d'aide à la décision se développent pour assister les professionnels dans la conduite de leur production. Cette méthode constitue une réponse aux problèmes de fusion d'informations numériques statistiques à des informations subjectives imprécises, en tenant compte des incertitudes de chacune. Elle ouvre ainsi des pistes permettant d'affiner les connaissances sur les paramètres d'un modèle de prédiction, lorsque leurs valeurs sont mal connues.

1. Analyse Sensorielle Descriptive Quantifiée

Table des figures

Bibliographie

[1] W.G. Alvord and J.L. Rossio. Determining confidence limits for drug potency in immunoassay. *Journal of Immunological Methods*, 157(1-2) :155–163, 1993.

[2] S.L. Beal and L.B. Sheiner. Heteroscedastic nonlinear regression. *Technometrics*, 30(3) :327–338, 1988.

[3] C.S. Berkey. Bayesian approach for a nonlinear growth model. *Biometrics*, 38(4) :953–961, 1982.

[4] B. L. Bishop, D. C. Ferree, J. F. Gallander, T. E. Steiner, D. M. Scurlock, and J. C. Schmid. Sources of variation in maturation soluble solids of three white grape cultivars. *Journal of the American Pomological Society*, 59(3) :153–160, 2005.

[5] J. Blouin and G. Guimberteau. *Maturation et Maturité des raisins*. Féret, 2000.

[6] J. Blouin and E. Peynaud. *Connaissance et travail du vin*. Edition La VIGNE, Dunod, 3 edition, 2001.

[7] A. Carbonneau. Theory of grape berry maturation and typicity. *Progrès Agricole et Viticole*, 124(13-14) :275–284, 2007.

[8] A. Carbonneau, P. Casteranp, and P. Leclair. Essai de détermination, en biologie de la plante entière, de relations essentielles entre la bioclimat naturel, la physiologie de la vigne et la composition du raisin. méthodologie et premiers résultats sur les systèmes de conduite. *Ann. Amélior. Plantes*, 28 :195–221, 1978.

[9] A. Carbonneau, A. Deloire, and B. Jaillard. *La vigne. Physiologie, terroir, culture.* Dunod, 2007.

[10] A. Carbonneau, A. Moueix, N. Leclair, and J. Renoux. Proposition d'une méthode de prélèvement de raisins à partir de l'analyse de l'hétérogénéité de maturation sur un cep. *Bull. de l'OIV*, 727/728 :679–690, 1991.

[11] B.P Carlin and T.A. Louis. *Bayes and Empirical Bayes methods for data analysis*. Chapman and Hall / CRC, 2 edition, 2000.

[12] R.J. Carroll and C.H. Spiegelman. Quick and easy multiple-use calibration-curve procedure. *Technometrics*, 30(2) :137–141, 1988.

[13] F. Champagnol. *Elements de physiologie de la vigne et de la viticulture générale*. Champagnol Saint-Gely-du-Fesc, 1984.

[14] I.S. Chan, A.A. Goldstein, and J.B. Bassingthwaighte. Sensop : A derivative-free solver for nonlinear least squares with sensitivity scaling. *Annals of Biomedical Engineering*, 21(6) :621–631, 1993.

[15] J. P. Chandler. On an iterative procedure for estimating functions when both variables are subject to error. *Technometrics*, 14(1) :71–76, 1972.

[16] N. Checchi, E. Giusti, and S. Marsili-Libelli. Peas : A toolbox to assess the accuracy of estimated parameters in environmental models. *Environmental Modelling and Software*, 22(6) :899–913, 2007.

[17] R.C.H. Cheng. Bootstrapping simultaneous confidence bands. In *Proceedings - Winter Simulation Conference*, volume 2005, pages 240–247, 2005.

[18] M.C. Coleman and D.E. Block. Bayesian parameter estimation with informative priors for nonlinear systems. *AIChE Journal*, 52(2) :651–667, 2006.

[19] R.L. Cooley. Confidence intervals for ground-water models using linearization, likelihood, and bootstrap methods. *Ground Water*, 35(5) :869–880, 1997.

[20] B. G. Coombe and M. G. McCarthy. Dynamics of grape berry growth and physiology of ripening. *Australian Journal of Grape and Wine Research*, 6(2) :131–135, 2000.

[21] C. Cox and G. Ma. Asymptotic confidence bands for generalized nonlinear regression models. *Biometrics*, 51(1) :142–150, 1995.

[22] D. Cozzolino, R. G. Dambergs, L. Janik, W. U. Cynkar, and M. Gishen. Analysis of grapes and wine by near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 14(5) :279–289, 2006.

[23] G. César. Rapport d'information sur l'avenir de la viticulture française. Technical report, Sénat, 2002.

[24] A. De Brauwere, F. De Ridder, M. Elskens, J. Schoukens, R. Pintelon, and W. Baeyens. Refined parameter and uncertainty estimation when both variables are subject to error. case study : Estimation of si consumption and regeneration rates in a marine environment. *Journal of Marine Systems*, 55(3-4) :205–221, 2005.

[25] S. De Gryze, I. Langhans, and M. Vandebroek. Using the correct intervals for prediction : A tutorial on tolerance intervals for ordinary least-squares regression. *Chemom. Intell. Lab. Syst*, 87 :147–154, 2007.

[26] A. Deloire, A. Carbonneau, Z. Wang, and H. Ojeda. Vine and water, a short review. *J. Int. Sci. Vigne Vin*, 37(4) :199–211, 2003.

[27] Janet R. Donaldson and Robert B. Schnabel. Computatinal experience with confidence regions and confidence intervals for nonlinear lest squares. *Technometrics*, 29(1) :67–82, 1987.

[28] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. *Fuzzy Inf and Decis Processes*, pages 167–181, 1982.

[29] D. Dubois and H Prade. *Théorie des possibilités*. Masson, 1988.

[30] D. Dubois and H. Prade. On possibility/probability transformation. In *Proc. Fourth IFSA Congress, Mathematics, Brussels*, pages 50–53, 1991.

[31] D. Dubois and H. Prade. Possibility theroy and data fusion in poorly informed environments. *Control Engineering Practice*, 2(5) :881–823, 1994.

[32] D. Dubois and H. Prade. What are rules and how to use them. *Fussy sets and systems*, 84(2) :169–185, 1996.

[33] D. Dubois, H. Prade, and L. Ughetto. A new perspective on reasoning with fuzzy rules. *International Journal of Intelligent System*, 18(5) :541–567, 2003.

[34] D. Dubourdieu, A. Lonvaud, P. Ribéreau-Gayon, and B. Donèche. *Traité d'oenologie*, volume 1, Microbiologie du vin, Vinification. Edition La VIGNE, Dunod, 5 edition, 2004.

[35] T.A. Ebert and M.P. Russell. Allometry and model ii non-linear regression. *Journal of Theoretical Biology*, 168(4) :367–372, 1994.

[36] A.M. Ellison. Bayesian inference in ecology. *Ecology Letters*, 7(6) :509–520, 2004.

[37] J. Fernandez-Novales, M.-I. Lopez, M.-T. Sanchez, J. Morales, V. Gonzalez-Caballero, J. Fernandez-Novales, M.-I. Lopez, M.-T. Sanchez, J. Morales, and V. Gonzalez-Caballero. Shortwave-near infrared spectroscopy for determination of reducing sugar content during grape ripening, winemaking, and aging of white and red wines. *Food Research International*, 42(2) :285–291, 2009.

[38] S. Fruhwirth-Schnatter. On fuzzy bayesian inference. *Fuzzy Sets and Systems*, 60(1) :41–58, 1993.

[39] P. Galet. *Précis de viticulture*. JF Impression, 7 edition, 2000.

[40] L. Gascogne. *Elements de logique floue*. Hermes, 1 edition, 1997.

[41] V. Geraudie, JM. Roger, JL Ferrandis, and JM. Gialis. A revolutionnary device for predicting grape maturity based on nir-spectroscopy. In *Fruitic Chile 2009 - 8th International Symposium of Information for the Sustainable Production of Fruit and Vegetables, Nuts, Wines and Olives.*, 2009.

[42] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic press, Inc., 1981.

[43] S. Guillaume. Designing fuzzy inference systems from data : an interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, 9(3) :426–443, 2001.

[44] S. Guillaume and B. Charnomordic. Generating an interpretable family of fuzzy partitions. *IEEE Transactions on Fuzzy Systems*, 12(2) :324–335, 2004.

[45] W.W. Hauck. A note on confidence bands for the logistic response curve. *Am. Stat.*, 37 :158–160, 1983.

[46] E. Hellman. How to judge grape ripeness before harvest. In *Southwest Regional Vines and Wine Conference*, 2004.

[47] J. Herrera, A. Guesalaga, and E. Agosin. Shortwave-near infrared spectroscopy for non-destructive determination of maturity of wine grapes. *Measurement Science and Technology*, 14(5) :689–697, 2003.

[48] J. C. van Houwelingen. Use and abuse of variance models in regression. *Biometrics*, 44(4) :1073–1081, 1988.

[49] S. Huet, E. Jolivet, and A. Messéan. *La régression non-linéaire : méthodes et applications en biologie*. INRA edition, 1992.

[50] H. Ichihashi, T. Shirai, K. Nagasaka, and T. Miyoshi. Neuro-fuzzy id3 : a method of inducing fuzzy decision trees with linear programming for maximizing entropy and analgebraic method for incremental learning. *Fuzzy Sets and Systems*, 81 :157–167, 1996.

[51] D. I. Jackson and P. B. Lombard. Environmental and management practices affecting grape composition and wine quality - a review. *Am. J. Enol. Vitic.*, 44(4) :409–430, 1993.

[52] K. Jaqaman and G. Danuser. Linking data to models : Data regression. *Nature Reviews Molecular Cell Biology*, 7(11) :813–819, 2006.

[53] C. Jaren, J. C. Ortuno, S. Arazuri, J. I. Arana, and M. C. Salvadores. Sugar determination in grapes using nir technology. *International Journal of Infrared and Millimeter Waves*, 22(10) :1521–1530, 2001.

[54] M.L. Johnson. Parameter correlations while curve fitting. *Methods in Enzymology*, 321 :424–446, 2000.

[55] G. V. Jones and R. E. Davis. Climate influences on grapevine phenology, grape composition, and wine production and quality for bordeaux, france. *American Journal of Enology and Viticulture*, 51(3) :249–261, 2000.

[56] N. Katerji, F. Daudet, and N. Carbonneau, A. Ollat. Etude à l'échelle de la plante entière du fonctionnement hydrique et photosynthétique de la vigne : comparaison des systèmes de conduite traditionnel et en lyre. *Vitis*, 33 :197–203, 1994.

[57] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the nelder-mead simplex : Method in low dimensions. *SIAM Journal of Optimization*, 9(1) :112–147, 1998.

[58] M. Larrain, A. R. Guesalaga, and E. Agosin. A multipurpose portable instrument for determining ripeness in wine grapes using nir spectroscopy. *IEEE Transactions on Instrumentation and Measurement*, 57(2) :294–302, 2008.

[59] M. Le Moigne, C. Maury, D. Bertrand, and F. Jourjon. Sensory and instrumental characterisation of cabernet franc grapesaccording to ripening stages and growing location. *Food Quality and Preference*, 19(2) :220–231, 2008.

[60] P. Legendre and L Legendre. *Numerical ecology.* Elsevier Science BV, 2 edition, 1998.

[61] J.M. Lisy and P. Simon. Evaluation of parameters in nonlinear models by the least squares method. *Computers and Chemistry*, 22(6) :509–513, 1998.

[62] M. Markatou and G. Manos. Robust tests in nonlinear regression models. *Journal of Statistical Planning and Inference*, 55(2) :205–217, 1996.

[63] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics*, volume Part A. Elseiver, 1997.

[64] M.K. McAllister and G.P. Kirkwood. Bayesian stock assessment : A review and example application using the logistic model. *ICES Journal of Marine Science*, 55(6) :1031–1060, 1998.

[65] R.B. McCammon. Nonlinear regression for dependent variables. *Journal of the International Association for Mathematical Geology*, 5(4) :365–375, 1973.

[66] J. McQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[67] G. Meinrath, C. Ekberg, A. Landgren, and J. O. Liljenzin. Assessment of uncertainty in parameter evaluation and prediction. *Talanta*, 51(2) :231–246, 2000.

[68] P. Menesatti. In-field spectrophotometric measurement to esimate maturity stage of wine grapes. In *Optics for Natural Resources, Agriculture, and Foods II*, volume 6761, 2007.

[69] H.J. Motulsky and A. Christopoulos. *Fitting Models to Biological Data Using Linear and Nonlinear Regression a Practical Guide to Curve Fitting*. Oxford University Press, 2008.

[70] H.J. Motulsky and L.A. Ransnas. Fitting curves to data using nonlinear regression : a practical and nonmathematical review. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 1(5) :365–374, 1987.

[71] R.A. Moyeed and R.T. Clarke. The use of bayesian methods for fitting rating curves, with case studies. *Advances in Water Resources*, 28(8) :807–818, 2005.

[72] R.F. Muzic Jr and B.T. Christian. Evaluation of objective functions for estimation of kinetic parameters. *Medical Physics*, 33(2) :342–353, 2006.

[73] B. M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron, and J. Lammertyn. Nondestructive measurement of fruit and vegetable quality by means of nir spectroscopy : A review. *Postharvest Biology and Technology*, 46(2) :99–118, 2007.

[74] H. Ojeda. Irrigation qualitative de précision de la vigne. *Le progrès Agricole et Viticole*, 7 :1–13, 2007.

[75] H. Ojeda, C. Andray, E. Kraeva, A. Carbonneau, and A. Deloire. Influence of pre and postveraison water deficit on synthesis and concentration of skin phenolic compounds during berry growth of vitis vinifera l., cv shiraz. *Am. J. Enol. Vitic.,*, 53(4) :261–267, 2002.

[76] H. Ojeda, A. Deloire, and A. Carbonneau. Influence of water deficits on grape berry growth. *Vitis*, 40 :141–145, 2001.

[77] M. Oussalah. On the probability/possibility transformations : a comparative analysis. *International Journal of General Systems*, 29(5) :671–718, 2000.

[78] C. Prakash Gupta. A note on the transformation of possibilistic information into probabilistic information for investment decisions. *Fuzzy Sets and Systems*, 56(2), 1993.

[79] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1 :81–106, 2001.

[80] A Reynier. *Manuel de viticulture*. Edition TEC et DOC, Lavoisier, 10 edition, 2007.

[81] J.F. Robison-Cox. Multiple estimation of concentrations in immunoassay using logistic models. *Journal of Immunological Methods*, 186(1) :79–88, 1995.

[82] V. Rod and V. Hancl. Iterative estimation of model parameters when measurements of all variables are subject to error. *Computers and Chemical Engineering*, 4(2) :33–38, 1980.

[83] P.F. Scholander, H.T. Hammel, E.T. Brandstreet, and E.A. Hemmingsen. Sap pressure in vascular plants. *Science*, 148 :339–346, 1965.

[84] M. Scholze, W. Boedeker, M. Faust, T. Backhaus, R. Altenburger, and L.H. Grimme. A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environmental Toxicology and Chemistry*, 20(2) :448–457, 2001.

[85] G. A. F. Seber and C. J. Wild. *Non linear regression*. John Wiley and Sons, 1989.

[86] J. Shao. Asymptotic theory in heteroscedastic nonlinear models. *Statistics and Probability Letters*, 10(1) :77–85, 1990.

[87] L.B. Sheiner and S.L. Beal. Pharmacokinetic parameter estimates from several least squares procedures : Superiority of extended least squares. *Journal of Pharmacokinetics and Biopharmaceutics*, 13(2) :185–201, 1985.

[88] W. Silvert. Practical curve fitting. *Limnology and Oceanography*, 24(4) :767–773, 1979.

[89] P. Smets. Constructing the pignistic probability function in a context of uncertainty. *Uncertainty in Artificial Intelligence*, 5, 1990.

[90] W.H. Southwell. Fitting experimental data. *Journal of Computational Physics*, 4(4) :465–474, 1969.

[91] S. E. Spayd, J. M. Tarara, D. L. Mee, and J. C. Ferguson. Separation of sunlight and temperature effects on the composition of vitis vinifera cv. merlot berries. *American Journal of Enology and Viticulture*, 53(3) :171–182, 2002.

[92] M.E. Spilker and P. Vicini. An evaluation of extended vs weighted least squares for parameter estimation in physiological modeling. *Journal of Biomedical Informatics*, 34(5) :348–364, 2001.

[93] Joel Tellinghuisen. Stupid statistics ! *Methods in Cell Biology*, 84 :739–780, 2008.

[94] J. Tonietto and A. Carbonneau. A multicriteria climatic classification system for grape growing regions wordwide. *Agriccultural Forest Meteology*, 124 :81–97, 2004.

[95] J. Valente de Oliveira. Semantic constraints for membership functions optimization. *IEEE Transactions on Systems, Man and Cybernetics*, 29(1) :128–138, 1999.

[96] G. Valsami, A. Iliadis, and P. Macheras. Non-linear regression analysis with errors in both variables : Estimation of co-operative binding parameters. *Biopharmaceutics and Drug Disposition*, 21(1) :7–14, 2000.

[97] M.A.J.S. Van Boekel. Statistical aspects of kinetic modeling for food science problems. *Journal of Food Science*, 61(3) :477–485, 1996.

[98] K.B. Walsh, J.A. Guthrie, and J.W. Burney. Application of commercially available, low-cost, miniaturised nir spectrometers to the assessment of the sugar content of intact fruit. *Australian Journal of Plant Physiology*, 27(12) :1175–1186, 2000.

[99] Z.-P. Wang, A. Deloire, A. Carbonneau, B. Federspiel, and F. Lopez. An in vivo experimental system to study sugar phloem unloading in ripening grape berries during water deficiency stress. *Annals of Botany*, 92 :1–6, 2003.

[100] R. Weber. Fuzzy-id3 : A class of methods for automatic knowledge acquisition. In *2nd International Conference On Fuzzy Logic and Neural Netwoks*, pages 265–268, 1992.

[101] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3) :338–353, 1965.

[102] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-i. *Information Sciences*, 8(3) :199–249, 1975.

[103] L. A. Zadeh. Fussy sets as a basis for a therory of possibility therory. *Fuzzy Sets and Systems*, 1(1) :3–28, 1978.