An exploratory penalized regression to identify combined effects of temporal variables -Application to agri-environmental issues

Bénédicte Fontez, Patrice Loisel, Thierry Simonneau, Nadine Hilgert

Biometrics 000, 000–000 DOI: 000
000 0000

An exploratory penalized regression to identify combined effects of temporal variables - Application to agri-environmental issues

Bénedicte Fontez[1,*], Patrice Loisel[1], Thierry Simonneau[2]

and Nadine Hilgert[1]

[1]MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France.

[2]LEPSE, Université Montpellier, INRAE, Institut Agro, Montpellier, France.

email: benedicte.fontez@supagro.fr

Summary: The development of sensors in several fields of activity opens new avenues. With regard to agricultural crops, complex combination of agri-environmental dynamics, such as soil and climate variables, are now commonly recorded. These new kinds of measurements are an opportunity to improve knowledge on the drivers of yield and quality at harvest. This involves renewing statistical approaches to take into account the combined variations of these dynamic variables, which are considered here as temporal variables. The objective of the paper is to infer an interpretable model to study the influence of the two combined inputs on a scalar output. A Sparse and Structured Procedure is proposed to Identify Combined Effects of Formatted temporal Predictors, denoted SpiceFP. It is based on a transformation of both temporal variables into categorical variables by defining joint modalities, from which a collection of multiple regression models is derived. The regressors are the frequencies associated to the joint class intervals. Selection of class intervals and related regression coefficients are performed through a Generalized Fused Lasso. SpiceFP is a generic and exploratory approach. Simulations performed show that it is flexible enough to select the non null or impacting modalities of values. A motivating example for grape quality is also presented.

Key words: Generalized Fused Lasso, information criteria, interpretable coefficient, joint distribution, penalized linear regression

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Nowadays, several fields of activity are being revolutionized by the emergence of sensor data. With regard to agricultural crops, the setting up of harvest management can now be based on monitoring with the aim of including/modelling the influence of multiple environmental conditions. Specifically, water scarcity and temperature increase are two major factors which have long been analyzed and considered to be determining factors causing huge variations in crop yield. However, relationships between fluctuating climatic conditions and quality of the harvest are still poorly understood and modelling approaches are still lacking.

A recent European project Innovine has funded research for combining innovation (like an increased use of sensor and non destructive measurements) in vineyard management for a sustainable European viticulture. Our motivating example on the quality of the grape berry comes from this context. Results and Expert knowledge indicate a multi-factor impact of the climate. As for the composition of anthocyanin, a determinant of grape berry colour, it often results from a complex system of ordinary differential equations (Dai et al., 2017) and complex interaction with abiotic factors (temperature and irradiance mainly). This complexity has prevented the emergence of an anthocyanin composition model equivalent to the growth model. There is a need for methods able to explore which combination of climatic variables influences the quality of harvest and at which stage of plant development. Climate variables constitute multivariate temporal data which can be input of supervised learning black box tools. All these black box tools are based on complex combinations of the regressors whose individual or combined effects are difficult to interpret.

The objective of the present paper is to infer an interpretable model to study the combined influence of two temporal variables on a scalar output.

A classical approach could be to consider these temporal data as a classical multivariate sample. In this case, LASSO-type regularisation methods (Tibshirani, 1996) allow the se-

lection of the most relevant explanatory variables, even for data sets where the number of explanatory variables can be greater than the number of individuals (Zhang and Huang, 2008; Meinshausen and Yu, 2009; Fan and Tang, 2012). In the context of temporal data, such methods allow to select, for each variable, the time instants during which this variable had an influence on the scalar output (Zhou, Wang, and Wang, 2013; Grollemund et al., 2019; Centofanti et al., 2022). But they are not suitable to study the combined influence of two variables when we are more concerned with the time spent in specific combinations of values (to be identified) taken by both variables rather than with the dates at which these values were observed.

We propose a Sparse and Structured Procedure to Identify Combined Effects of Formatted temporal Predictors, denoted SPICEFP. The temporal predictors are formatted based on the assumption that the relationship between the variable to be explained and the predictors is stable over time, i.e. we assume the same additive relation between output and input variables during the whole period/time of observation. We also assume sparsity, i.e. that only some ranges of cross-values of these variables have an impact on the scalar output and only the time spent in these ranges is important. This type of relationship is often observed in agronomy, particularly in the ripening of grapes. The literature (Fernandes de Oliveira et al., 2015; Spayd et al., 2002; Downey, Dokoozlian, and Krstic, 2006) suggests that high temperatures over a long period have a negative impact on the biosynthesis pathway and that, conversely, low temperatures associated with high irradiation favour the accumulation of anthocyanins (Cohen, Tarara, and Kennedy, 2008).Thus, the temporal data set can be viewed as repeated measurements of two explanatory variables. We performed a transformation of the temporal data and proposed a model that is close to a sparse scalar-on-image type regression, where the *so-called* image is a bivariate count of cross-values of the two explanatory variables.

Scalar-on-image regression models aim to control the smoothness of non-zero estimated

coefficients. This is consistent with usual biological processes that adapt or react progressively (up to a point) to environmental conditions. Different approaches are used to control the smoothness, among which Bayesian approaches (Li et al., 2015; Goldsmith, Huang, and Crainiceanu, 2014), total variation penalizing approaches (Wang and Zhu, 2017), neighborhood taken into account in the selection of variables (Li et al., 2020) inspired by the Fused Lasso, etc. Kang, Reich, and Staicu (2018) proposed an approach based on the Gaussian process and compared it to the Fused Lasso. Other studies on scalar-on-image regressions are inspired by, used in or compared to models involving different L_1 regularization. Following this trend, we chose to use the fused Lasso, and more specifically its implementation via the *genlasso* package (Arnold and Tibshirani, 2019), for identifying parsimonious and structured coefficients. The selection of the coefficients is performed using information criteria instead of cross-validation, as proposed in Zhou and Li (2014).

This paper is organised as follows: a motivating example in Section 2, the SpiceFP model in Section 3 followed by the SpiceFP selection approach in Section 4, then to illustrate the interest of SpiceFP, simulations in Section 5 followed by the motivating example results in Section 6, eventually a discussion in Section 7.

2. Motivating example

Experts in viticulture assume that the accumulation of chemical compounds affecting the quality of the grape berry is jointly influenced by micro-climate variables. This assumption is reinforced by results of Tarara et al. (2008), which underlined that the anthocyanin composition, a major criterion that determines technological maturity at harvest in red grapes, was influenced by a complex combined effect of berry temperature and solar irradiation. These results motivated an experimental design of INNOVINE, which enabled the decoupling of temperature and irradiance values and also the observation of the climatic conditions

expected with global warming. This experiment conducted in Montpellier in 2014 (Syrah vines) stimulated the development of the SpiceFP procedure.

The challenge was to identify precisely the ranges of temperature and irradiance values that jointly influence or not the accumulation of anthocyanins between sunrise and noon thanks to the use of sensors (non destructive high-throughput measurements).

The experimental plot was made of three rows of vines within the vineyard, each with eight vines equipped with open-top chambers to warm the base of the plant (Sadras, Bubner, and Moran, 2012), and eight under control conditions (without open-top chambers). The greenhouse effect created during the day in the chambers generated a flow of warm air that escaped through the open top, raising the temperature of the bunches by 2 to 3 °C, mimicking global warming.

The microclimate was recorded through the measurement of temperature and irradiance. Irradiance was separately measured on bunches located on the east and west side of the row. The temperature sensors were positioned at the bunch scale, two sensors by vine: one in the east, the other in the west. Temperature and irradiance were recorded every twelve minutes throughout the maturation period when anthocyanins are known to accumulate. Anthocyanin contents were measured weekly via the Ferari Index (FI) obtained from the Multiplex optic sensor (Ben Ghozlen et al., 2010; Bramley et al., 2011). This is a non-destructive measure of anthocyanin content in berries at the bunch scale (Agati et al., 2007).

One originality of our approach was to transform both explanatory temporal variables into categorical variables by defining joint modalities using class intervals (with bins of equal size). Temperature and Irradiance are variables of different natures. Temperature is a variable whose variations are regular enough to be partitioned according to a linear scale. Observed on a one-day scale, Irradiance increases exponentially from sunrise to a daily peak (observation time t_{max}), decreases until sunset, and remains almost constant until the

next sunrise. Irradiance primarily influences plant photosynthesis in a nonlinear way with a maximal reached at high irradiance. Therefore, the Irradiance variable was partitioned according to a logarithmic scale. The logarithmic transformation has consistently been used in the development of models involving solar radiation (Salminen et al., 1983; Bergqvist, Dokoozlian, and Ebisuda, 2001).

3. The SpiceFP modelling

3.1 *Transformation of both temporal variables: a contingency table for a given partition*

Suppose that we observe n triplets $(\mathcal{A}_i(.), \mathcal{B}_i(.), y_i)_{i=1,\dots,n}$, where n is the number of statistical individuals, $\mathcal{A}_i$ and $\mathcal{B}_i$ are the explanatory temporal variables and y_i is the response variable. Both $\mathcal{A}$ and $\mathcal{B}$ are supposed to be observed on the same set T of fixed equidistant observation times, with no missing values. These practical conditions of use can be relaxed with pre-processing of the data e.g., interpolation, smoothing and imputation; see Section 7 for more details.

The values taken by the temporal variables $\mathcal{A}$ and $\mathcal{B}$ are partitioned into, respectively, $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ class intervals. The partition for $\mathcal{A}$ generates $n_{\mathcal{A}}+1$ breaks denoted $L_{\mathcal{A}}(v, n_{\mathcal{A}}),\ v = 1, \dots, n_{\mathcal{A}}+1$. We chose to have equidistant breaks, as defined in Equation (3.1):

$$L_{\mathcal{A}}(v, n_{\mathcal{A}}) = \underline{\mathcal{A}} + \frac{v-1}{n_{\mathcal{A}}}\left(\overline{\mathcal{A}} - \underline{\mathcal{A}}\right),\ v = 1, \dots, n_{\mathcal{A}}+1, \tag{3.1}$$

with $\underline{\mathcal{A}} \in \mathbb{R}$ and $\overline{\mathcal{A}} \in \mathbb{R}$ the minimum and maximum observed values of $\mathcal{A}$ taking into account all individuals. The bins used for partitioning all $(\mathcal{A}_i)_{i=1\dots n}$ are $I_{\mathcal{A}}(v, n_{\mathcal{A}}) = [L_{\mathcal{A}}(v, n_{\mathcal{A}}), L_{\mathcal{A}}(v+1, n_{\mathcal{A}})], v = 1, \dots, n_{\mathcal{A}}$. The partition is the same for all i, $i = 1, \dots, n$. Using the same approach for partitioning the second explanatory variable $\mathcal{B}$, we obtain $n_{\mathcal{B}}+1$ breaks $L_{\mathcal{B}}(w, n_{\mathcal{B}})$ and corresponding $I_{\mathcal{B}}(w, n_{\mathcal{B}}) = [L_{\mathcal{B}}(w, n_{\mathcal{B}}), L_{\mathcal{B}}(w+1, n_{\mathcal{B}})], w = 1, \dots, n_{\mathcal{B}}$. The numbers of class intervals $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ have to be set before computing the breaks.

For all individual i, we construct the contingency table $\mathbf{C}_i^{n_{\mathcal{A}} n_{\mathcal{B}}}$, of dimension $n_{\mathcal{A}} \times n_{\mathcal{B}}$,

whose component in row v, and column w is a frequency count obtained through:

$$C_{i,(v,w)}^{n_{\mathcal{A}}n_{\mathcal{B}}} = \text{Card } \{t \in T | \mathcal{A}_i(t) \in I_{\mathcal{A}}(v, n_{\mathcal{A}}), \mathcal{B}_i(t) \in I_{\mathcal{B}}(w, n_{\mathcal{B}})\}, \tag{3.2}$$

for all $v = 1, \ldots, n_{\mathcal{A}}$, $w = 1, \ldots, n_{\mathcal{B}}$ and each $(n_{\mathcal{A}}, n_{\mathcal{B}})$, with: $\sum_{v=1}^{n_{\mathcal{A}}} \sum_{w=1}^{n_{\mathcal{B}}} C_{i,(v,w)}^{n_{\mathcal{A}}n_{\mathcal{B}}} = Card(T)$.

The frequency $C_{i,(v,w)}^{n_{\mathcal{A}}n_{\mathcal{B}}}$ is the number of times that the observations of $\mathcal{A}_i$ and $\mathcal{B}_i$ are at the same time in $I_{\mathcal{A}}(v, n_{\mathcal{A}})$ and $I_{\mathcal{B}}(w, n_{\mathcal{B}})$ respectively.

Figure 1 shows the transformation of the temporal explanatory variables $\mathcal{A}$ and $\mathcal{B}$ for the partition $(n_{\mathcal{A}}, n_{\mathcal{B}}) = (4, 3)$.

[Figure 1 about here.]

3.2 *Regression model for a given partition*

The SpiceFP model is defined, for each partition $(n_{\mathcal{A}}, n_{\mathcal{B}})$ and each individual i, by:

$$y_i = \sum_{v=1}^{n_{\mathcal{A}}} \sum_{w=1}^{n_{\mathcal{B}}} C_{i,(v,w)}^{n_{\mathcal{A}}n_{\mathcal{B}}} \, \beta_{(v,w)} + \varepsilon_i, \tag{3.3}$$

where $C_{i,(v,w)}^{n_{\mathcal{A}}n_{\mathcal{B}}}$ is given in Equation (3.2), $\beta_{(v,w)}$ is the coefficient of regression of the joint class $(I_{\mathcal{A}}(v, n_{\mathcal{A}}) \times I_{\mathcal{B}}(w, n_{\mathcal{B}}))$ and $\varepsilon_i \sim N(0, \sigma^2)$ is an independent, identically distributed (i.i.d) Gaussian error. This model is a linear multiple regression model where the regressors are the frequencies of the contingency table $\mathbf{C}_i^{n_{\mathcal{A}}n_{\mathcal{B}}}$.

We remark that the i.i.d. Gaussian error assumption is usual for a continuous variable and reasonable for the motivating example where the individuals are observations from different vines on the same plot (same soil, same genetic and so on) and with no spatial correlations observed between them.

From the contingency tables, we construct the design matrix $\mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}$ associated to model (3.3) as follows. After vectorization (stacking column by column) and transposition of the contingency table $\mathbf{C}_i^{n_{\mathcal{A}}n_{\mathcal{B}}}$ (see Matrix $\mathbf{X}$ in Figure 1), we obtain, for a partition $(n_{\mathcal{A}}, n_{\mathcal{B}})$, a row vector $X_i \in \mathbb{R}^{n_{\mathcal{A}}n_{\mathcal{B}}}$ of length $n_{\mathcal{A}}n_{\mathcal{B}}$:

$$X_i = \text{Vect}(\mathbf{C}_i^{n_{\mathcal{A}}n_{\mathcal{B}}})^T, \tag{3.4}$$

which represents the number of observation times during which an individual i has been observed in each of the $n_{\mathcal{A}} \times n_{\mathcal{B}}$ joint classes of the contingency table. The n stacked row vectors form the matrix $\mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})} = (X_1, X_2, \ldots, X_n)^T \in \mathbb{R}^{n \times n_{\mathcal{A}}n_{\mathcal{B}}}$. The regression parameter $\boldsymbol{\beta}_{(n_{\mathcal{A}}n_{\mathcal{B}})}$ is a vector of length $n_{\mathcal{A}}n_{\mathcal{B}}$ obtained by stacking all the coefficients $\beta_{(v,w)}$ in the same order as the vectorization of the contingency table. The dimensions of the matrix $\mathbf{X}$ and of the parameter vector $\boldsymbol{\beta}$ are given in brackets as subscripts as a reminder of their dependencies on the partition dimension $(n_{\mathcal{A}}, n_{\mathcal{B}})$.

The process of transformation generates missing data for some joint classes which are not observed in the dataset. The corresponding columns of $\mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}$ and coefficients $\beta_{(v,w)}$ are therefore removed from the model. The position and number of columns and coefficients to be removed depends on the size of the partition $(n_{\mathcal{A}}, n_{\mathcal{B}})$. A higher dimension results in more missing data for joint classes.

4. The SpiceFP feature selection approach

4.1 *Generalized Fused Lasso to select variables in the regression model (3.3) for a given partition*

For a given partition, the objective of the approach is to infer an interpretable and biologically realistic model thanks to sparsity and smoothness constraints on the parameter $\boldsymbol{\beta}_{(n_{\mathcal{A}}n_{\mathcal{B}})}$.

This objective can be addressed with the Generalized Lasso model introduced by Tibshirani and Taylor (2011) as an encapsulation of statistical models using the L_1 norm to impose additional constraints. The following criterion has to be minimized:

$$\frac{1}{2}||\boldsymbol{y} - \mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}\boldsymbol{\beta}_{(n_{\mathcal{A}}n_{\mathcal{B}})}||_2^2 + \lambda||D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)\boldsymbol{\beta}_{(n_{\mathcal{A}}n_{\mathcal{B}})}||_1, \tag{4.1}$$

where $||.||_q$ designs the L_q norm $(q = 1, 2)$, $\mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}$ was defined in equation 3.4, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$ is the response vector.

The parameter γ represents a balance between sparsity and fusion (sparsity is controlled

by the product of γ by λ). For a fixed value of γ, we can define a penalty matrix $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ as per below. The value of λ controls the degree of fusion and can be optimized with the path algorithm and for a given $\gamma \geqslant 0$ a corresponding $\widehat{\boldsymbol{\beta}}^{\gamma}_{(n_{\mathcal{A}} n_{\mathcal{B}})}(\lambda)$ can be estimated. The penalty matrix we suggest has the following shape:

$$D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma) = \begin{pmatrix} D^{f1}(n_{\mathcal{A}}, n_{\mathcal{B}}) \\ D^{f2}(n_{\mathcal{A}}, n_{\mathcal{B}}) \\ D^{p}(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma) \end{pmatrix} \in \mathbb{R}^{(3 n_{\mathcal{A}} n_{\mathcal{B}} - n_{\mathcal{A}} - n_{\mathcal{B}}) \times n_{\mathcal{A}} n_{\mathcal{B}}} \text{ with:}$$

$$D^{f1}_{(v,w)(v',w')}(n_{\mathcal{A}}, n_{\mathcal{B}}) = \begin{cases} 1 & \text{if} \quad (v', w') = (v+1, w) \\ -1 & \text{if} \quad (v', w') = (v, w) \text{ and } v < n_{\mathcal{A}} \\ 0 & \text{if} \quad \text{not} \end{cases},$$

$$D^{f2}_{(v,w)(v',w')}(n_{\mathcal{A}}, n_{\mathcal{B}}) = \begin{cases} 1 & \text{if} \quad (v', w') = (v, w+1) \\ -1 & \text{if} \quad (v', w') = (v, w) \text{ and } w < n_{\mathcal{B}} \\ 0 & \text{if} \quad \text{not} \end{cases}, \tag{4.2}$$

$$D^{p}_{(v,w)(v',w')}(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma) = \begin{cases} \gamma & \text{if} \quad (v', w') = (v, w) \\ 0 & \text{if} \quad \text{not} \end{cases}.$$

The sub-matrix $D^{p}(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ $(= \gamma\ \mathbb{I}_{n_{\mathcal{A}}.n_{\mathcal{B}}})$ induces sparsity in the coefficients, but when $\gamma = 0$, the model penalty becomes a pure fusion penalty. The sub-matrices $D^{f1}(n_{\mathcal{A}}, n_{\mathcal{B}})$ and $D^{f2}(n_{\mathcal{A}}, n_{\mathcal{B}})$ are associated to the constraints of fusion. They are defined using the joint classes of the given partition. These sub-matrices correspond to the Rook's case contiguity rule (Plant, 2012) where two joint classes are said to be close if the bins following the variable $\mathcal{A}$ (indexed by v) or (exclusive) the bins following the variable $\mathcal{B}$ (indexed by w) are consecutive, as shown in the following diagram:

$$\begin{array}{ccccc} & & (\text{v+1,w}) & & \\ & & \Uparrow & & \\ (\text{v,w-1}) & \Longleftarrow & (\text{v,w}) & \Longrightarrow & (\text{v,w+1}) \\ & & \Downarrow & & \\ & & (\text{v-1,w}) & & \end{array}$$

The constraints induced by the Rook's rule in $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ penalize jumps in β values in the 2D space of the parameter.

We used the R package *genlasso* (Arnold and Tibshirani, 2019) for implementing the estimation of the SpiceFP model (3.3) under the generalized constraints induced by the penalty matrix $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$.

4.2 *The SpiceFP feature selection: the best partition and its best regressors*

SpiceFP needs, as input values, the numbers of breaks $(n_{\mathcal{A}}, n_{\mathcal{B}})$ to define several candidate partitions for the observed variables ($\mathcal{A}$ and $\mathcal{B}$). It realises the construction of a design matrix $\mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}$ hereafter called candidate matrix, for each candidate partition. The candidate matrices lead to different regressors in terms of number and in terms of definition, as the joint classes are not the same. In penalized regressions, cross-validation is often used and implemented to optimize regularization parameters, but it is time consuming, needs a sufficiently large sample size (to avoid a missing value effect) and the running time will be multiply by the number of candidate matrices (as we need to optimize also the partition). In SpiceFP, candidate matrices and regularization parameters are selected at the same time, using information criteria by default. It requires estimating the degree of freedom $Q(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda)$ for each model, see Tibshirani and Taylor (2012). A short summary of their general theorem and its application to the SpiceFP context is given in the Web Appendix (see SECTION Supplementary Materials).

There exist various information criteria including Akaike Information Criterion (AIC)

(Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz, 1978). These criteria penalize the log-likelihood by the number of model parameters. The BIC also penalizes the log-likelihood by the sample size. Given a partition $(n_{\mathcal{A}}, n_{\mathcal{B}})$ and for each value of γ taken on a predefined grid (given by the users) and λ taken from the set $(\lambda_e)_{e=1,\dots,N_\lambda}$ delivered by the path algorithm, we consider the following information criteria:

$$AIC(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda) = -2\log(L(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda)) + 2\ Q(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda) \quad \text{and}$$

$$BIC(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda) = -2\log(L(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda)) + \log(n)\ Q(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda),$$

with $L(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda)$ the likelihood function of the following model: $\boldsymbol{y} = \mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}\boldsymbol{\beta}_{(n_{\mathcal{A}}n_{\mathcal{B}})} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\ \sigma^2\mathbb{I}_{n\times n})$, associated with the criterion (4.1). We have:

$$-2\log L(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda) = n\log(\sigma^2) + n\log(2\pi) + \frac{1}{\sigma^2}||\boldsymbol{y} - \mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}\widehat{\boldsymbol{\beta}}^{\gamma}_{(n_{\mathcal{A}}n_{\mathcal{B}})}(\lambda)||_2^2.$$

Computing the chosen AIC or BIC requires to know the variance σ^2, for all the models. In a simpler context (with a design matrix that does not vary with the partition $(n_{\mathcal{A}}, n_{\mathcal{B}})$) Hirose, Tateishi, and Konishi (2013) suggest taking the unbiased estimator of the error variance of the most complex model. In our context, it is not trivial to define a single most complex or full model for all possible partitions, i.e. contingency tables $\mathbf{C}^{n_{\mathcal{A}}n_{\mathcal{B}}}$. The dimension $n_{\mathcal{A}}n_{\mathcal{B}}$ could theoretically be infinite with therefore a predictor dimension theoretically infinite. Moreover, in our context, the models are not nested, as each "sub-model" is in fact a different contingency table with different joint classes. No sub-model can therefore be defined as a subset of the variables of a complete model. Application of the standard practice from the literature on the tuning parameter selection (Wang, Li, and Tsai, 2007; Wang, Li, and Leng, 2009), even in high dimensional sparse linear regression (Wang and Zhu, 2011), is not straightforward in our situation. We have therefore chosen to estimate σ^2 by the variance of the response variable: $\widehat{\sigma}^2 = \frac{1}{n-1}||\boldsymbol{y} - \overline{\boldsymbol{y}}||_2^2$. It is a biased estimator of σ^2, but this bias remains fixed for all models compared. Such an estimator may lead to overestimate the variance, which penalizes the introduction of new coefficients in the model. The selection

of the best model (best partition with its best regressors) is done by computing the chosen Akaike criterion (AIC or BIC) for each combination of $(n_{\mathcal{A}}, n_{\mathcal{B}})$, γ and λ. The optimal values of $\widehat{n_{\mathcal{A}}}$, $\widehat{n_{\mathcal{B}}}$, $\widehat{\gamma}$ and $\widehat{\lambda}$ are those which minimize the criterion. The estimator $\widehat{\boldsymbol{\beta}}^{\widehat{\gamma}}_{(\widehat{n_{\mathcal{A}}}\widehat{n_{\mathcal{B}}})}(\widehat{\lambda})$ of $\boldsymbol{\beta}$ is deduced.

5. Simulation study

5.1 *Simulation design and SpiceFP setting*

We performed a simulation study to evaluate the finite sample behavior of the SpiceFP algorithm in three challenging situations, where the input variables are the measures of Temperature and log-Irradiance in the Vine dataset, obtained between sunrise and noon during the week of July 17 to 24, 2014. Only 79 statistical individuals were used from the dataset, due to missing data. In the following, we refer to $\mathcal{A}$ as the Temperature variable and $\mathcal{B}$ as the log-Irradiance variable.

Simulation 1 (two patches). To create a diffuse effect, we used a very small scale $(n_{\mathcal{A}} = 30, n_{\mathcal{B}} = 30)$ to construct a design matrix $\boldsymbol{X}_{79,900}$ with the observed frequencies of Temperature and log-Irradiance, using equations (3.2) and (3.4). We chose to simulate two different patches of coefficients $\beta_{v,w}$ (one with positive values and the other with negative values) positioned at two different locations. It should be noted that negative values (around -0.3) are 2 to 3 times higher than positive values (around 0.1). They are presented in row 2 of the Table 1, data is available for consultation (see SECTION Data Availability below).

The response variable of the simulation is computed using the following model:

$$\boldsymbol{y} = \boldsymbol{X}_{79,900}\,\boldsymbol{\beta}_{900} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0,\, 1.5\,\mathbb{I}_{79\times 79}). \tag{5.1}$$

Simulation 2 (two patches, one with higher impact). We multiplied by 10 all positive values of the parameter vector $\boldsymbol{\beta}_{900}$ in (5.1).

Simulation 3 (one patch). We set all positive values of the parameter vector $\boldsymbol{\beta}_{900}$ in (5.1) to 0.

Simulation 4 (two rough patches). We divided by 2 the dimensions ($n_{\mathcal{A}} = 15, n_{\mathcal{B}} = 15$), by averaging the values of beta coefficients, to simulate

$$\boldsymbol{y} = \boldsymbol{X}_{79,225}\,\boldsymbol{\beta}_{225} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0,\ 1.5\,\mathbb{I}_{79\times 79}). \tag{5.2}$$

We thus generated 100 datasets for each simulation. Estimation was done with the SpiceFP algorithm implemented in the SpiceFP R-package (Gnanguenon Guesse et al., 2023), and the following input parameters: $(n_{\mathcal{A}}, n_{\mathcal{B}}) \in \{10, 15, 17, 20, 23, 25, 30\}^2$, $\gamma \in \{0, 1/9, 1/3, 1, 3, 9\}$ and $n_\lambda = 20$. For simulation 4, we reduce the possible matrix candidates to $(n_{\mathcal{A}}, n_{\mathcal{B}}) \in \{10, 15, 20, 25, 30\}^2$. The running time with 4 cores of `spicefp()` function, for each dataset (simulation 1, 2 and 3), was roughly 3 hours (so 300 hours for each simulation), while it reduces to 1 hour per dataset with 4 cores for simulation 4.

[Table 1 about here.]

5.2 *Simulation results*

For each of the four simulations, an illustrative map is drawn on Table 1. A summary of the four simulations is given in Table 2. The main objective of SpiceFP is more exploratory than predictive, and consists in identifying possible combinations of Temperature and log-Irradiance that have an impact (i.e. with non-zero coefficients). We used the usual classification metrics, where sensitivity is defined as the rate of non-zero coefficients correctly identified by the model among the true non-zero coefficients. Conversely, the specificity is the rate of zero-value coefficients correctly identified by the model among the true zero-value coefficients. Prior to computing this, we removed the non affected combinations due to missing values (it modified slightly the results).

In addition, to evaluate the overall goodness of fit we computed the $R^2_{adjusted} = 1 - ||\boldsymbol{y} - \mathbf{X}_{(n,n_{\mathcal{A}}n_{\mathcal{B}})}\widehat{\boldsymbol{\beta}}^{\gamma}_{(n_{\mathcal{A}}n_{\mathcal{B}})}(\lambda)||_2^2/||\boldsymbol{y} - \overline{\boldsymbol{y}}||_2^2 * (n-1)/(n - Q(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma, \lambda))$ as it's a simple to calculate

and well-known criterion that takes model complexity and overfitting into account for model comparisons.

A good specificity was observed for all simulations. Over the 100 datasets of each simulation, less than 10 showed a bad specificity. Simulation 3 exhibited the lowest median value for specificity combined with the lowest $R^2_{adjusted}$. The increase in identification errors is understandable since, in simulation 3, the data sets contained a large number of intensively observed Temperature-log Irradiance combinations with zero-value coefficients. For all simulations, the sensitivity was quite good. Lower values were observed for simulation 2, where SpiceFP often failed to identify the patch of negative coefficients. Again, it was understandable as those coefficients were smaller (in absolute value) and near 0 compared to the patch of positive coefficients. Finally, the degree of freedom is quite small (around 3) for all simulations and minimum (1-2) for simulation 2. The sparsity looked high but it did not seem to result exclusively from the lasso penalty. A pure fusion penalty led to quite the same results (the median values for the df were respectively 4, 3, 4 and 3.5 for the simulations 1 to 4) but with a lower goodness of fit. The use of AIC with σ^2 estimated by the sample variance of Y could be an explanation as well as the configuration of the simulated parameters (two distinct patches with homogeneous values). Other criteria are implemented in the package SpiceFP, such as Mallow's C_p, the generalized cross validation (Wang, Li, and Tsai, 2007, GCV) and were tested. In addition, we tried to compute the AIC with the residual sum of squares divided by n for σ^2. All of these additional criteria led to overfitted models (median values of df around 140 and $R^2_{adjusted}$ always equal to 1), as with the use of the GCV (implemented in the SpiceFP package).

Finally, in order to produce a suitable visualization in Table 1, the vector of coefficients was projected onto a fine-mesh matrix of combinations of Temperature-log Irradiance values (we kept the SpiceFP suggested matrix size of 900×900).

[Table 2 about here.]

6. Results obtained on the motivating example: a step toward modelling the evolution of a grape berry quality index

6.1 *Methodology used for data analysis*

We focus in this section on the modelling of the Ferari Index variation, denoted here as ΔFI, from July 24 to August 1, 2014. We selected the $n_1 = 32$ individuals (bunches) which have the highest contribution to the final Ferari Index during this week of study, with an initial index around 0.2 at the beginning of the week.

We focused our modelling on the morning when grapevine achieves most of its photosynthesis. The time period between sunrise and noon is denoted T_1 below. The input variables of SpiceFP are $y_i = \Delta FI_i$, $\mathcal{A}_i(t)$ (temperature values) and $\mathcal{B}_i(t)$ (irradiance values) for $i = 1, \ldots, n_1$ and $t \in T_1$. Estimation was done with the following parameters: $(n_{\mathcal{A}}, n_{\mathcal{B}}) \in \{10, 11, 12, \ldots, 29, 30\}^2$, $\gamma \in \{0, 1/9, 1/3, 1, 3, 9\}$ and $n_\lambda = 20$.

6.2 *Results*

The results are presented in Table 3. The first column shows the estimated coefficients and the second column the response variable and the contingency table.

In addition to the result of the AIC best model, we computed a post process of the SpiceFP results by averaging the coefficients of the 1% best models, where these were defined here in the sense of the set of models with the lowest 1% of AIC values. Coefficients available in these 1% best models were defined on different partitions and were all projected onto the Temperature-log Irradiance fine-mesh matrix of size 900×900 before averaging. The map for the 1% best models in Table 3 is the representation of the function $\mathcal{F}$ defined by: $\mathcal{F}(v, w) = \dfrac{1}{n_m} \displaystyle\sum_{m=1}^{n_m} \beta^{(m)}_{(v,w)}$, where $n_m = 529$ is the number of 1% best models and $v, w \in \{1, 2, \ldots, 900\}^2$.

In terms of goodness of fit, the AIC best model has: slope ($\widehat{y}$ as a function of y) = 0.49,

$R^2_{adjusted} = 0.60$, $df = 4$, $\widehat{\gamma} = 0$, $\widehat{\lambda} = 0.84$, $\widehat{n}_{\mathcal{A}}$=14, $\widehat{n}_{\mathcal{B}} = 28$. That is, the best model according to AIC corresponds to a pure fused penalty ($\widehat{\gamma} = 0$) with an estimated coefficient fusion of $\widehat{\lambda}$ = 0.84. This model was able to explain 60% of the differences between Ferari Index weeks, which is quite satisfactory for a field experiment. The residuals distribution is unimodal and shows asymmetry (same as $\mathbf{Y}$). The visualization of the coefficients indicates conditions (Irradiance < 100 $\mu mol\ m^{-2}\ s^{-1}$, Temperature from 20°C to 33°C) that affected the Ferari Index negatively. The unique positive coefficient value is 0.001, and practically speaking can be set to zero given this is more realistic from a biological point of view than a positive value with an effect over the whole Temperature-log Irradiance range (see Table 3). In fact, the result of the AIC best model looked very similar to the results obtained in simulation 2. We might assume that we missed a patch of coefficients of smaller values in the upper right side (high Temperature and Irradiance) where frequency of observation is high. This hypothesis is reinforced by the result of the 1% best models which allowed identification of more coefficients in this localization. The average estimation from the 1% best models suggests a border zone between the zones of positive and negative influences. This border zone looks sensible from a biological point of view.

Finally, in the morning (sunrise to noon), a large range of temperature values with low irradiance values (Irradiance < 100 $\mu mol\ m^{-2}\ s^{-1}$) seemed not suitable for an increase of the Ferari Index. On the contrary, a combination of irradiance values above 150 $\mu mol\ m^{-2}\ s^{-1}$ and temperature values below 30°C seemed suitable for increasing the Ferari Index. The average of the coefficients suggested a possible gradation in the impact of some Temperature and log Irradiance combinations.

[Table 3 about here.]

6.3 *A first step toward modelling*

From the map representation of the estimated value for the regression parameter (Table 3) we observed that high temperatures combined with low morning irradiances had a negative impact that delayed anthocyanin accumulation and technological maturity of red grapes. This negative impact is observed for temperature levels that increase with the irradiance level, drawing an oblique separation line in the Temperature-log Irradiance fine-mesh matrix in both figures on the left of Table 3. These results show the importance of combinations of Temperature - log Irradiance values and are consistent with the literature (Fernandes de Oliveira et al., 2015; Spayd et al., 2002; Downey, Dokoozlian, and Krstic, 2006) and (Cohen, Tarara, and Kennedy, 2008). It highlights the interest for further studies to analyze the possible decoupling of Temperature and log Irradiance. Following the SpicеFP parameter estimation results, we may investigate the following agronomic model:

$$\begin{aligned}\Delta FI_i = \beta_- \text{ Card}\{t|\mathcal{A}_i(t) > S(\mathcal{B}_i(t)) \text{ and } B_0 < \mathcal{B}_i(t) < B_1\} \\ + \beta_+ \text{ Card}\{t|(\mathcal{A}_i(t) < S(\mathcal{B}_i(t)) \text{ and } \mathcal{B}_i(t) > B_0) \text{ or } \mathcal{B}_i(t) > B_1\},\end{aligned}$$

where ΔFI_i is the increase in Ferari Index during the selected week and S is the function of the straight line: $S(\mathcal{B}) = A_0 + \theta(\mathcal{B} - B_0)$ and $\beta_- < 0 < \beta_+$. The model is illustrated in the following figure:

[Figure 2 about here.]

It is a first step toward a model for anthocyanin accumulation that determines technological maturity at harvest. Establishment of this kind of simple model could be useful for example to define a degree-lux day for maturity similar to the degree day for growth. The degree day is a transformation suitable to adjust plant growth from different environments, to make growth comparable.

7. Discussion

The SpiceFP approach is dedicated to temporal biological or physical process which ended in a final point with a limited value, such as crop production and harvest quality. The aim of SpiceFP is to extract ranges of values from at least two temporal variables which impact on the final point, thus increasing our knowledge on the process by enlightening how and when the two temporal variables combine and influence the final point.

To achieve this goal, SpiceFP first transforms the temporal variables into 2D class intervals. This transformation makes sense if the underlying process, namely the relationship between the outcome and the input values, does not change over time. This is often the case in agriculture where, for example, phenological stages can be estimated from cumulative degree days. In the motivating example on grapevines quality, this assumption of stability required to work at the scale of a week. When analyzing the data in Section 6, we defined the scale of time (week) that ensures that the underlying biological process remains the same. On this scale of time, SpiceFP searched for reference ranges of combined values, favorable or unfavorable for the increase of the Ferari Index (which is correlated with anthocyanin concentrations and is an indirect indicator of the grapevines quality).

The transformation of temporal variables yields a contingency table. It is assumed that no data are missing and that the observation times are identical for both temporal variables. These constraints can be released with usual pretreatment such as: imputation of missing data (Stekhoven and Bühlmann, 2011; Josse and Husson, 2016), interpolation or smoothing (Ramsay, Hooker, and Graves, 2009).

The underlying assumption of normality could also be released/relaxed in a future version of SpiceFP. The generalized linear models with lasso or elastic net regularization (Friedman, Tibshirani, and Hastie, 2010; Tay, Narasimhan, and Hastie, 2023) could be implemented using the already available R package glmnet.

The principal output of the SpiceFP algorithm is the map of the β estimated values (see Table 1). A common drawback to all data driven approaches is the fact that the results are design-dependent: the SpiceFP algorithm will not be able to properly estimate the β coefficient in an area with little or no data. It is therefore necessary to have data of $(\mathcal{A}, \mathcal{B})$ in areas where combinations of temporal variables has a potential interest. Note also that components in the contingency table (3.2) that contain no observed values of the pair $(\mathcal{A}_i, \mathcal{B}_i)$ are considered missing values. The number of missing values and their location in the contingency table change according to the dimensions $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$. To gain in precision, we would tend to take larger values of $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$, which implies more components without observations, and therefore more missing values. There is also a limitation due to the curse of dimensionality (Giraud, 2014): if the number n of observations remains fixed while the dimension $p = n_{\mathcal{A}} n_{\mathcal{B}}$ of the variables increases, the observations get rapidly very isolated and local methods cannot work. The more fluctuations in many directions, the more data will be needed.

SpiceFP algorithm is sensitive to some parameters like the variance estimator of the residuals. Our suggestion is to use the sample variance estimator which clearly is an overestimation, but simple, easy to compute and identical for all contingency tables we used as candidate models in the SpiceFP algorithm. Another possibility would be to let users give a range of values to check the sensitivity. Occasionally, we observed that many models have AIC or BIC values close to the best model. In practice, we recommend to check not only the best model, but the top percentage e.g., 1% of best models, to visually inspect the stability of the results. A last suggestion to improve this work could be the use of adaptive penalty matrices (Zhang et al., 2023). A first step could be to define a different matrix $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ for each input variable (see Tibshirani, 2014 for suggestions of penalty matrices), a second step could be a data driven local weight/penalty.

For the simulations and motivating example, we chose the Rooks' case contiguity rule with a $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ matrix which penalizes jumps in the parameter bivariate discretized function (by cancelling the first derivative in both direction). More regular solutions can be implemented in the SpiceFP package, for example by replacing the $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ matrix by a $D'(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ matrix which penalizes jumps in the derivative (by cancelling the second derivatives). For example, by replacing in both direction the Rooks' neighborhood definition $(+1, -1, 0)$ by $(-1, +2, -1)$. Regularity can be further enhanced by replacing $D(n_{\mathcal{A}}, n_{\mathcal{B}}, \gamma)$ by a matrix penalizing jumps in higher-order derivatives.

Other approaches than GFL deal with variable selection and may be explored as an alternative in the future: for example, the square-root lasso (Belloni, Chernozhukov, and Wang, 2011) and the quantile universal threshold (Giacobino et al., 2017). The square root does not require the estimation of the variance σ^2 but is not implemented for generalized lasso with contiguity constraints. It may very well control the FDR (False Discovery Rate) but less the TPR (True Positive Rate), according to the simulation results presented in (Giacobino et al., 2017).

In life sciences, exploratory experiments are commonly carried out to suggest new hypothesis. More in-depth analyses will be conducted at a later step to test whether these hypotheses are valid or not. In the same spirit, the SpiceFP approach provides an exploratory analysis which is used for acquiring information. To our knowledge, there is a real need in agronomy for the development of exploratory tools of this type, able to take advantage of temporal data measured by sensors. Models can be proposed on the basis of the β estimate map results and validated subsequently by further experiments.

Acknowledgements

The authors dedicate this work to their late colleague Eric Lebon. He participated in the very fruitful discussions that initiated the present work, and provided us with the data.

Data were collected during the Innovine project, which was funded by the Seventh Framework Program of the European Community (FP7/2007-2013), under Grant Agreement No. FP7-311775. The present work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

The data were processed with SpiceFP package and functions developped by Girault Gnanguenon Guesse during his PhD thesis. We thank him for his invaluable help and support. We are also grateful to Nicolas Verzelen (INRAE, Montpellier), for helpful discussions.

The authors would like to thank the editors, the Co-Editor and both anonymous referees for their insightful remarks that lead us to significantly improve the presentation of the paper.

Data Availability

The data underlying this article are available in this paper and in its online Supplementary Materials.

References

Agati, G., Meyer, S., Matteini, P., and Cerovic, Z. G. (2007). Assessment of anthocyanins in grape (Vitis vinifera L.) berries using a noninvasive chlorophyll fluorescence method. *Journal of agricultural and food chemistry* **55**, 1053–1061.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, 199–213.

Arnold, T. B. and Tibshirani, R. J. (2019). genlasso: Path algorithm for generalized lasso problems. `https://cran.r-project.org/web/packages/genlasso/index.html`. Package version 1.4 for R version 3.6.1.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

Ben Ghozlen, N., Cerovic, Z. G., Germain, C., Toutain, S., and Latouche, G. (2010). Non-Destructive Optical Monitoring of Grape Maturation by Proximal Sensing. *Sensors* **10**, 10040–10068.

Bergqvist, J., Dokoozlian, N., and Ebisuda, N. (2001). Sunlight Exposure and Temperature Effects on Berry Growth and Composition of Cabernet Sauvignon and Grenache in the Central San Joaquin Valley of California. *American Journal of Enology and Viticulture* **52**, 1–7.

Bramley, R. et al. (2011). On-the-go sensing of grape berry anthocyanins during commercial harvest: development and prospects. *Australian Journal of Grape and Wine Research* **17**, 316–326.

Centofanti, F., Fontana, M., Lepore, A., and Vantini, S. (2022). Smooth LASSO estimator for the Function-on-Function linear regression model. *Comput. Stat. Data Anal.* **176**.

Cohen, S., Tarara, J., and Kennedy, J. (2008). Assessing the impact of temperature on grape phenolic metabolism. *Analytica Chimica Acta* **621**, 57–67.

Dai, Z. et al. (2017). Mathematic model for simulating anthocyanin composition during grape ripening: Another way of phenotyping. *Acta Horticulturae* **1160**, 375–380.

Downey, M., Dokoozlian, N., and Krstic, M. (2006). Cultural practice and environmental impacts on the flavonoid composition of grapes and wine : A review of recent research. *American Journal of Enology and Viticulture* **57**, 257–268.

Fan, Y. and Tang, C. Y. (2012). Tuning Parameter Selection in High Dimensional Penalized Likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75**, 531–552.

Fernandes de Oliveira, A., Mercenaro, L., Del Caro, A., Luca, P., and Nieddu, G. (2015). Distinctive anthocyanin accumulation responses to temperature and natural uv radiation of two field-grown vitis vinifera l. cultivars. *Molecules* **20(2)**, 2061–2080.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22.

Giacobino, C., Sardy, S., Diaz-Rodriguez, J., and Hengartner, N. (2017). Quantile universal threshold. *Electronic Journal of Statistics* **11**, 4701–4722.

Giraud, C. (2014). Introduction to High-Dimensional Statistics. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Gnanguenon Guesse, G., Loisel, P., Fontez, B., and Hilgert, N. (2023). SpiceFP: Sparse Method to Identify Joint Effects of Functional Predictors. `https://cran.r-project.org/package=SpiceFP`. Package version 0.1.2 for R version 3.6.1.

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014). Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection. *Journal of Computational and Graphical Statistics* **23**. PMID: 24729670, 46–64.

Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2019). Bayesian Functional Linear Regression with Sparse Step Functions. *Bayesian Analysis* **14**, 111 –135.

Hirose, K., Tateishi, S., and Konishi, S. (2013). Tuning parameter selection in sparse regression modeling. *Computational Statistics & Data Analysis* **59**, 28 –40.

Innovine (n.d.). European project. `https://cordis.europa.eu/project/id/311775`.

Josse, J. and Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software, Articles* **70**, 1–31.

Kang, J., Reich, B. J., and Staicu, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105**, 165–184.

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.* **9**, 687–713.

Li, Y., Sun, H., Deng, X., Zhang, C., Wang, H.-P. B., and Jin, R. (2020). Manufacturing quality prediction using smooth spatial variable selection estimator with applications in aerosol jet® printed electronics manufacturing. *IISE Transactions* **52**, 321–333.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37**, 246 –270.

Plant, R. (2012). Spatial Data Analysis in Ecology and Agriculture Using R. Taylor & Francis.

Ramsay, J., Hooker, G., and Graves, S. (2009). Functional Data Analysis with R and MATLAB. Use R! Springer New York.

Sadras, V., Bubner, R., and Moran, M. (2012). A large-scale, open-top system to increase temperature in realistic vineyard conditions. *Agric. Forest Meteorol.* **154-155**, 187–194.

Salminen, R., Hari, P., Kellomaki, S., Korpilahti, E., Kotiranta, M., and Sievanen, R. (1983). A Measuring System for Estimating the Frequency Distribution of Irradiance Within Plant Canopies. *Journal of Applied Ecology* **20**, 887–895.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* **6**, 461–464.

Spayd, S., Tarara, J., Mee, D., and Ferguson, J. (2002). Separation of sunlight and temperature effects on the composition of vitis vinifera cv. merlot berries. *American Journal of Enology and Viticulture* **53**, 171–182.

Stekhoven, D. J. and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118.

Tarara, J. M., Lee, J., Spayd, S. E., and Scagel, C. F. (2008). Berry Temperature and Solar Radiation Alter Acylation, Proportion, and Concentration of Anthocyanin in Merlot Grapes. *American Journal of Enology and Viticulture* **59**, 235–247.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software* **106**, 1–31.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42**, 285–323.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39**, 1335–1371.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40**, 1198–1232.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis* **102**, 1141–1151.

Wang, X. and Zhu, H. (2017). Generalized Scalar-on-Image Regression Models via Total Variation. *Journal of the American Statistical Association* **112:519**, 1156–1168.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594.

Zhang, X., Huang, F., Hui, F. K., and Haberman, S. (2023). Cause-of-death mortality forecasting using adaptive penalized tensor decompositions. *Insurance: Mathematics and Economics* **111**, 193–213.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **76**, 463–483.

Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional Linear Model with Zero-value Coefficient Function at Sub-regions. *Statistica Sinica* **23**, 25–50.

Supplementary Materials

Web Appendix, Tables, and Figures, and data and code (self-contained R project to apply SpiceFP to the motivating example and to analyze the results) referenced in Section 6 are available with this paper at the Biometrics website on Oxford Academic.

The motivating example dataset is available online in the SpiceFP R-package on CRAN https://forgemia.inra.fr/exploratory-penalized-regression/biometrics_practice.git `git@forgemia.inra.fr:exploratory-penalized-regression/biometrics_practice.git` . OU plus simple :

The data underlying this article are available in its online Supplementary Materials and in the ForgeMia (GitLab of the INRAE department MathNum) at (mettre en référence) Authors; Year; Dataset title; Data repository or archive; Version (if any); Persistent identifier (e.g. DOI)`https://forgemia.inra.fr/exploratory-penalized-regression/biometrics_practice.git`, open access.

8. Tables and Figures

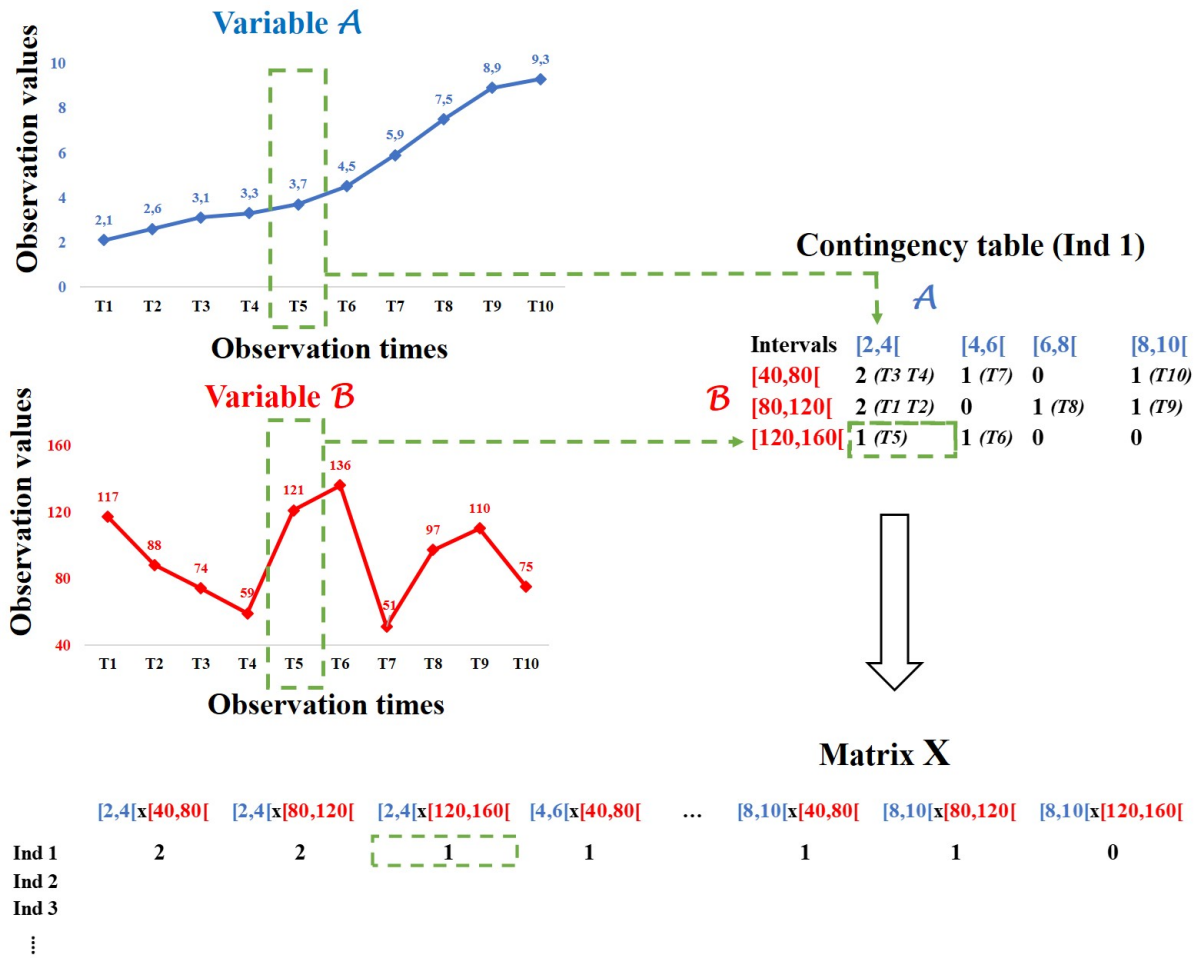


Figure 1. Transformation of both temporal explanatory variables for the SpiceFP approach

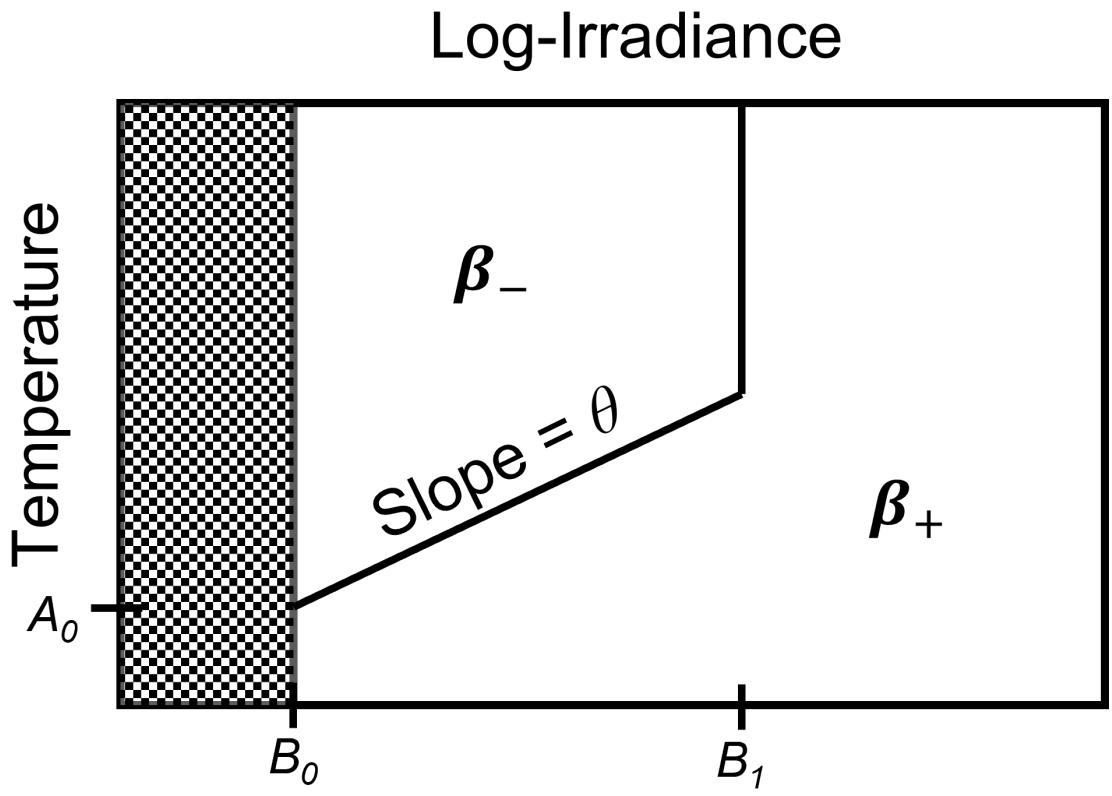


Figure 2. Model based on the oblique separation line

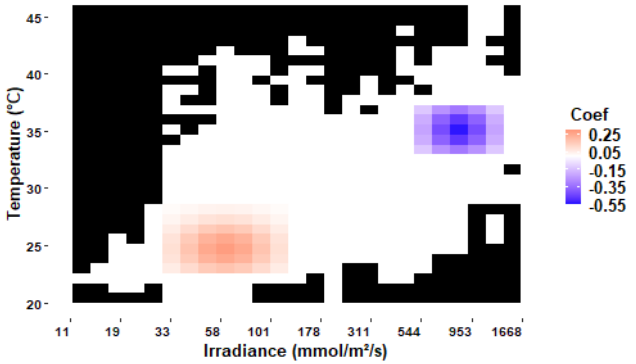
Table 1

Simulation results observed on a fine-mesh (900×900) *matrix: estimation of the coefficients* $\beta_{(v,w)}$ *was done with the* SpiceFP *package. Simulation 1: two patches (generated with a* 30×30 *contingency table). Simulation 2: simulation 1 where the true positive coefficient values* $\times 10$. *Simulation 3: simulation 1 where the true positive coefficient values were set to 0. Simulation 4: simulation 1 but generated with a* 15×15 *contingency table (true coefficient values were averaged).* ■*: Non affected value.*

Fine-mesh matrix of the contingency table 30×30

Simulation 1 True coefficients $\beta_{(v,w)}$

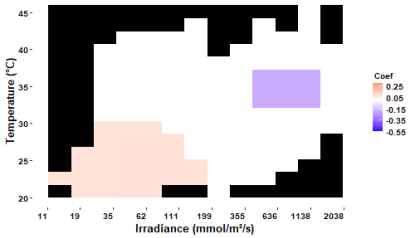
simulated dimension 30×30, df=23



Parameter estimation done on the first dataset of each simulation

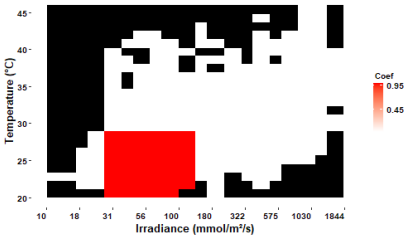
Simulation 1

selected dimensions: 15×15, $df = 3$



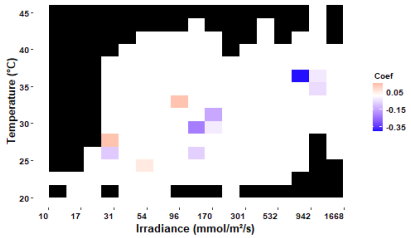
Simulation 2

selected dimensions: 23×23, $df = 1$



Simulation 3

selected dimensions: 15×20, $df = 11$



Simulation 4

selected dimensions: 10×10, $df = 4$

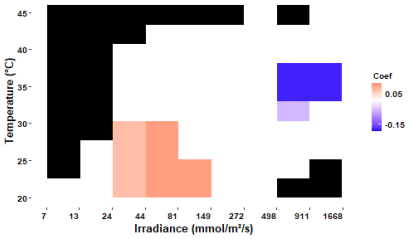


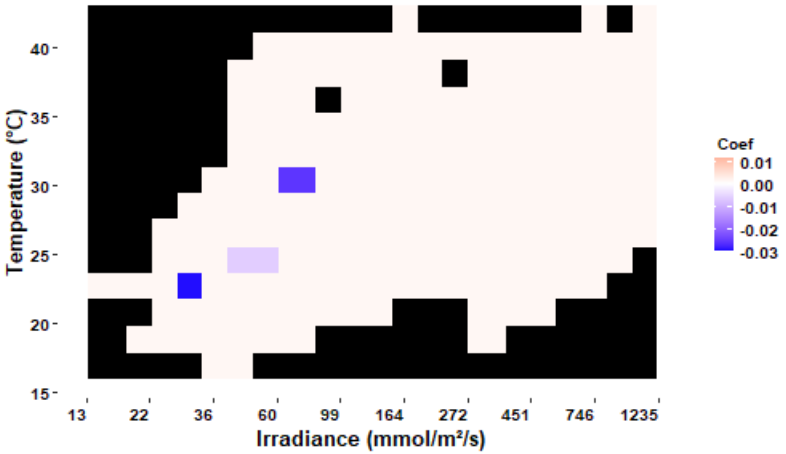
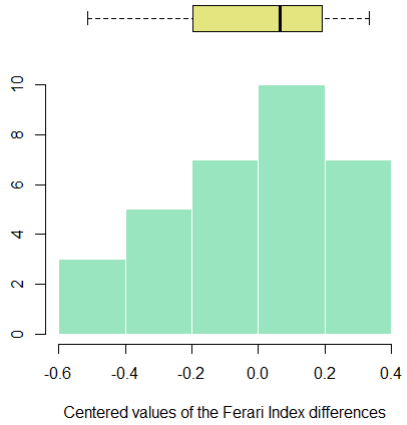
Table 2

Summary of the simulation results (median and min-max) in terms of overall goodness of fit and identification of the domain of non-zero coefficients evaluated on the fine-mesh matrix 900×900*.* $\mathcal{A}$ *and* $\mathcal{B}$ *are respectively for Temperature and Irradiance.*

	Goodness of fit			Non null coefficients domain			
	$R^2_{adjusted}$	Slope	RMSE	*df*	Specificity	Sensitivity	$n_{\mathcal{A}}$	$n_{\mathcal{B}}$
Simulation 1	0.91 0.86-0.95	0.81 0.70-0.90	3.27 1.87-5.21	3 2-5	0.89 0.00-1.00	0.92 0.41-1.00	15 10-30	18 10-30
Simulation 2	0.97 0.96-0.98	0.86 0.83-0.91	34.51 21.62-52.69	1 1-2	0.92 0.66-0.92	0.67 0.60-0.76	23 17-23	15 10-23
Simulation 3	0.46 0.28-0.57	0.40 0.22-0.52	2.10 1.42-3.03	3 1-11	0.70 0.00-0.98	0.94 0.00-1.00	20 10-30	18 10-30
Simulation 4	0.89 0.84-0.94	0.80 0.68-0.88	3.35 1.95-5.29	3.5 2-5	0.88 0.00-1.00	0.92 0.30-1.00	15 10-30	15 10-30

Table 3

Estimation of the combined effects of log Irradiance and Temperature values, observed from sunrise to noon, on the Ferari Index increment during one week. ■*: Non affected values. Row 1 presents the results of the best model. The second row presents the average of the 1% best models. Both results are represented on a fine-mesh matrix* (900×900)

Parameter Estimation from Best Model(s) (AIC)	***Variables***
AIC best model: $\widehat{\gamma} = 0$ *(pure fused),* $df = 4$	*Ferari Index* $(\mathbf{Y})$
	
Average estimation from 1% best models	*Fine-mesh matrix of the contingency table* 30×30