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

Girault Gnanguenon Guesse, Patrice Loisel, Bénedicte Fontez, Thierry Simonneau, Nadine Hilgert

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

Girault Gnanguenon Guesse[a], Patrice Loisel[a], Bénedicte Fontez[a,*], Thierry Simonneau[b], Nadine Hilgert[a]

[a]*MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France*
[b]*LEPSE, Université Montpellier, INRAE, Institut Agro, Montpellier, France*

Abstract

Crop production is affected by a complex combination of agri-environmental dynamics, as temperature, irradiance, for example. To learn on these complex influences, the development of sensors in agriculture opens new avenues. This requires renewing statistical approaches to take into account the joint variations of these dynamic variables, which are considered here as functional variables. The objective of the paper is to infer an interpretable model to study the joint influence of two functional inputs on a scalar output. We propose a Sparse and Structured Procedure to Identify Combined Effects of Functional Predictors, denoted SpiceFP. It is based on a transformation of both functional variables into categorical variables by defining joint modalities, from which we derived a collection of multiple regression models, where 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 true ranges of values. A use case in agronomy is also presented.

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

1. Introduction

Nowadays, several fields of activity and in particular, agriculture, are being revolutionized by the emergence of sensor data. With regard to crops, the setting up of harvest can now be monitored with the aim of including/modeling the influence of multiple environmental conditions. Specifically, water scarcity and temperature increase are two major features which have long been analyzed as determining huge variation in crop yield. Their influences are increasing with climate change and are becoming a major concern for the sustainability of agriculture in many parts of the world. However, relationships between climatic conditions and quality of the harvest are still poorly understood and modeling approaches are still lacking. To better use newly available data from sensors, there is a need for methods able to explore which combination of climatic variables influences harvest quality and at which stage of plant development. Such data sets

*Corresponding author
Email addresses: girault.gnanguenon@gmail.com (Girault Gnanguenon Guesse), patrice.loisel@inrae.fr (Patrice Loisel), benedicte.fontez@supagro.fr (Bénedicte Fontez), thierry.simonneau@inrae.fr (Thierry Simonneau), nadine.hilgert@inrae.fr (Nadine Hilgert)

involve multivariate, longitudinal or temporal data which are handled in various ways, including the large family of functional data analysis [7].

One of the main lines of research on functional data concerns their treatment in regression problems. The regression toolbox is used to extract knowledge from input variables (functional data in our context) to predict and/or explain a scalar output or a continuous variable of interest. Applying directly machine learning and/or supervised learning with its usual black box tools (support-vector machine [5], random forest [22], neural networks [29], etc.) is not indicated in our context: all these black-box tools are based on complex combinations of the regressors whose individual effects are difficult to interpret. In [28, 26], the regression models are usually classified into 3 categories according to the role played by functional data. A distinction is made between the "scalar-on-function", "function-on-scalar" and "function-on-function" regressions. In this paper we will focus on the "scalar-on-function" regression where the response variable is a scalar and the regressors are functions. More precisely, regressors are two functions that jointly influence the response variable.

Various methods exist to solve "scalar-on-function" regressions and the reader can refer to [27] for a review. Regression with functional data often uses pre-treatment of the data like interpolation and/or projection on a basis [26]. This kind of regression is easily implemented (see R-package FDA), but again makes the influence of regressors not easy to interpret. By contrast, the work on *"functional linear regression that's interpretable"* [15] and its Bayesian version [12] open a new research area for functional regressions where interpretation is of major interest. Unfortunately, these models do not take into account a possible combined effect of explanatory variables.

The objective of the present paper is to infer an interpretable model to study the joint influence of two functional inputs on a scalar output. Our approach is based on a transformation of the functional data that implies a change from "scalar-on-function" regression to "scalar-on-image" regression, where the "image" is a bivariate representation of both functional datasets. Scalar-on-image regression models aim to control the smoothness of non-zero estimated coefficients. Different approaches are used to solve scalar-on-image regressions, among which Bayesian approaches [18, 11], total variation penalizing approaches [40], neighborhood taken into account in the selection of variables [19] inspired by the Fused Lasso, etc. [17] proposed an approach based on the Gaussian process and compared it to the Fused Lasso. Other studies on scalar-on-image regressions are inspired, used 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 [3], for identifying parsimonious and structured coefficients. The selection of the coefficients is performed using information criteria instead of cross-validation, as proposed in [42].

In the following, we present a Sparse and Structured Procedure to Identify Combined Effects of Functional Predictors, denoted SpiceFP, and its theory. Simulations and a use case based on a real question in agri-environment are also provided.

2. The SpiceFP approach

This section describes the main steps of the approach; first in §2.1, the originality of our approach was to transform both functional variables into categorical variables by defining joint modalities using class intervals (with bins of equal size). Several candidate partitions are defined this way, depending on the choice of the bin size. The functional model is presented in §2.2, from which we derived a linear multiple regression model where the regressors are the frequencies associated to the joint class intervals. As explained in [8], when faced to a high number of discretization points of functional data, the naive approach would be to consider these data as a classical multi-

variate sample having as dimension the number of discretization points of the functional variables. In this case, multivariate statistics meets limits, among which its failure to take into account the very strong colinearity existing between discretized variables. By contrast, colinearity can be considered in the Fused Lasso penalized regression which was therefore retained in our approach. We followed a "scalar on image" regression model, where the "image" is a contingency table of the joint class intervals, for a fixed candidate partition, to which we associated a graph of contiguity constraints defined in subsection §2.3. Identification was performed through a Generalized Fused Lasso (see §2.4) using each candidate contingency table as input variables. The selection of the best candidate and of its relative regression coefficients was achieved by minimizing an information criteria. To combine our exploratory objectives with technical constraints (such as potentially a small amount of data, for example), we have favored an iterative approach (see §2.5) which explores a larger space of solutions, while keeping aware of possible slight overestimation.

2.1. Transformation of both functional variables

Let us consider the samples $(\mathcal{A}_i)_{i=1,\dots,n}$ and $(\mathcal{B}_i)_{i=1,\dots,n}$ of two explanatory functional variables $\mathcal{A}$ and $\mathcal{B}$, associated to a scalar response variable y with samples $(y_i)_{i=1,\dots,n}$, where n is the number of statistical individuals. Both $\mathcal{A}$ and $\mathcal{B}$ are observed on the same set T containing equidistant observation times. It is assumed that they have no missing values. These practical conditions of use can be released with some pre-treatment of the data (like interpolation, smoothing and imputation), see Section 6 for more details.

The requirement for variable transformation is intrinsically linked to the goal of the approach: identify joint class intervals of the explanatory variables that influence the response. The transformation requires the definition of joint class intervals which can be used as linear regressors to predict the response. The steps to achieve this transformation are shown in the Figure 2.1.

Contingency table of the joint class intervals

Let's partition the first explanatory variable $\mathcal{A}$ in $n_{\mathcal{A}}$ class intervals according to a linear scale. This partition generates $n_{\mathcal{A}}+1$ breaks denoted $L_{\mathcal{A}}(v),\ v=1,\dots,n_{\mathcal{A}}+1$. We chose to have equidistant breaks, as defined in equation (2.1) :

$$L_{\mathcal{A}}(v) = \underline{\mathcal{A}} + \frac{v-1}{n_{\mathcal{A}}}\left(\bar{\mathcal{A}} - \underline{\mathcal{A}}\right),\ v=1,\dots,n_{\mathcal{A}}+1 \tag{2.1}$$

with $\underline{\mathcal{A}} \in \mathbb{R}$ and $\bar{\mathcal{A}} \in \mathbb{R}$ the minimum and maximum scalar values in $\mathcal{A}$. The bins used for partitioning all $(\mathcal{A}_i)_{i=1\dots n}$ are $I_{\mathcal{A}}(v) = [L_{\mathcal{A}}(v), L_{\mathcal{A}}(v+1)[, v=1,\dots,n_{\mathcal{A}}$. 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 $L_{\mathcal{B}}(w)$. $\underline{\mathcal{B}} \in \mathbb{R}$ and $\bar{\mathcal{B}} \in \mathbb{R}$ are the minimum and maximum scalar values in $\mathcal{B}$. The bins used for partitioning all $(\mathcal{B}_i)_{i=1,\dots,n}$ are $I_{\mathcal{B}}(w) = [L_{\mathcal{B}}(w), L_{\mathcal{B}}(w+1)[, w=1,\dots,n_{\mathcal{B}}$. The numbers of class intervals $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ have to be set to compute the breaks $L_{\mathcal{A}}(v)$ and $L_{\mathcal{B}}(w)$, $v=1,\dots,n_{\mathcal{A}}$, $w=1,\dots,n_{\mathcal{B}}$.

Let's define a partition vector $u=(n_{\mathcal{A}}, n_{\mathcal{B}})$. For all i, it is then possible to obtain the frequency bivariate histogram of $(\mathcal{A}_i, \mathcal{B}_i)$ as a contingency table C_i^u, of dimension $n_{\mathcal{A}} \times n_{\mathcal{B}}$, whose components $C^u_{i,(v,w)}$ are obtained through (2.2). The new variables are the joint class intervals, noted $I^u_{\mathcal{A}}(v) \times I^u_{\mathcal{B}}(w)$, $v=1,\dots,n_{\mathcal{A}},\ w=1,\dots,n_{\mathcal{B}}$. They are called joint modalities.

$$C^u_{i,(v,w)} = \sum_{t\in T} \mathbb{1}_{\mathcal{A}_i(t)\in I^u_{\mathcal{A}}(v),\ \mathcal{B}_i(t)\in I^u_{\mathcal{B}}(w)} = Card\ \{t\in T | \mathcal{A}_i(t)\in I^u_{\mathcal{A}}(v), \mathcal{B}_i(t)\in I^u_{\mathcal{B}}(w)\}, \tag{2.2}$$

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

$C^u_{i,(v,w)}$ is the number of times that the observations of $\mathcal{A}_i$ and $\mathcal{B}_i$ are at the same time in $I^u_{\mathcal{A}}(v) \times I^u_{\mathcal{B}}(w)$. $C^u_{i,(v,w)}$ can also be interpreted as a discrete approximation of the density of the time spent by the individual i with variable $\mathcal{A}$ around $L_{\mathcal{A}}(v)$ and variable $\mathcal{B}$ around $L_{\mathcal{B}}(w)$.

Part *1* of Figure 2.1 shows the transformation of the functional explanatory variables $\mathcal{A}$ and $\mathcal{B}$ for the fixed $u = (4, 3)$. Note that, for a fixed u, the $\left(I^u_{\mathcal{A}}(v) \times I^u_{\mathcal{B}}(w)\right)_{v,w}$ is a collection of 2D intervals in which the pairs $(\mathcal{A}_i, \mathcal{B}_i)$ will be projected. This is detailed in the following subsection.

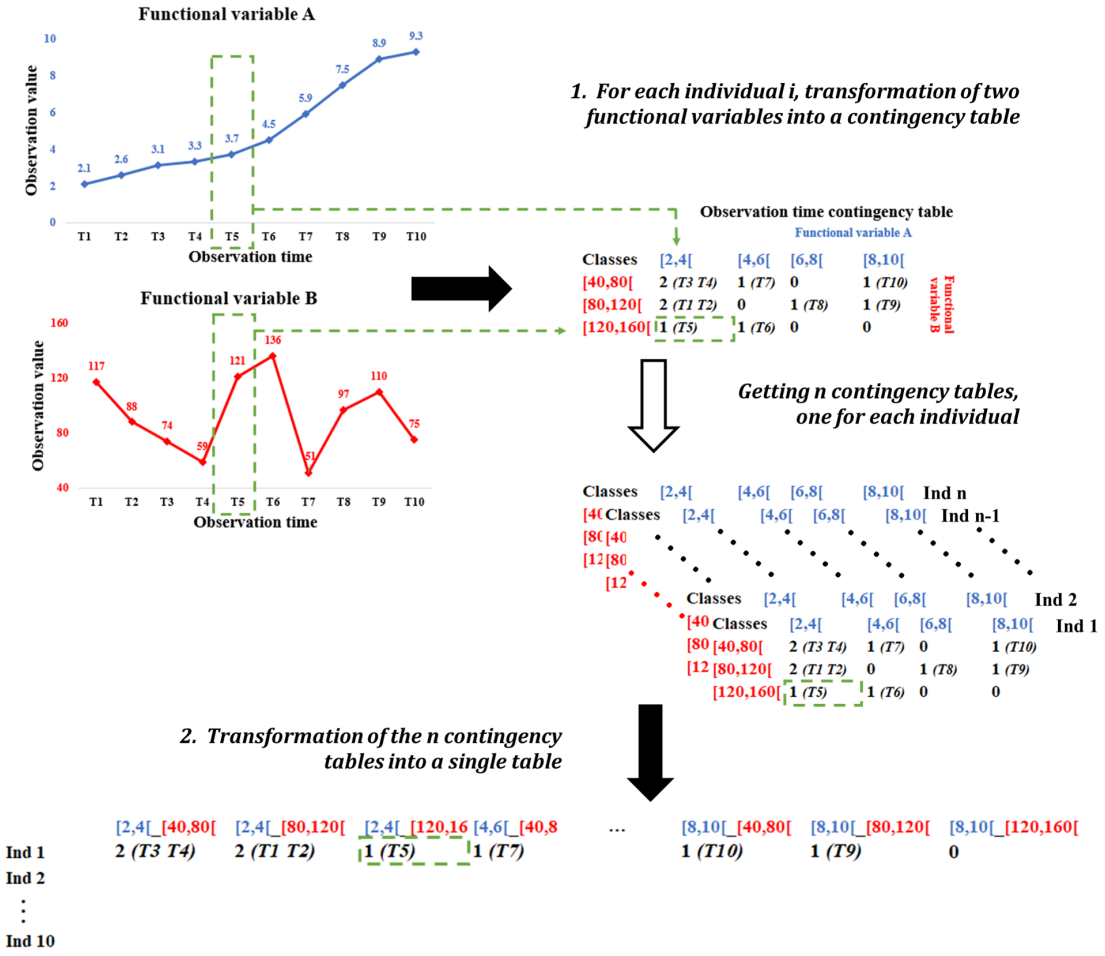


Figure 2.1: Transformation of both functional explanatory variables for the SPICEFP approach

2.2. Functional model

The functional model used in the present work can be formulated as follows:

$$y = \int F_{u,\beta}(\mathcal{A}(t), \mathcal{B}(t))dt + \varepsilon \tag{2.3}$$

where $F_{u,\beta} = \mathcal{I}^u.\beta$ (scalar product of two vectors), $\mathcal{I}^u$ is a vector of components functions $\mathcal{I}^u_{v,w}$ defined by $\mathcal{I}^u_{v,w}(a, b) = \mathbb{1}_{a \in I^u_{\mathcal{A}}(v),\ b \in I^u_{\mathcal{B}}(w)}$ for $a \in [\underline{\mathcal{A}}, \bar{\mathcal{A}}]$ and $b \in [\underline{\mathcal{B}}, \bar{\mathcal{B}}]$, and $\beta = (\beta_{(v,w)})_{v,w}$ is the

vector of unknown coefficients to estimate. ε is an i.i.d. Gaussian error. Through (2.3), we note that the relationship between the variable to be explained y and the predictors $\mathcal{A}(.)$ and $\mathcal{B}(.)$ is stable over time: the function $F_{u,\beta}$ (and thus β) is not time-dependent. It is hence assumed that certain ranges of cross-values have an influence and that only the time spent in these ranges is important.

A discrete version of the model (2.3) can be written as follows (up to a constant), for each i :

$$y_i = \sum_{t\in T} F_{u,\beta}(\mathcal{A}_i(t), \mathcal{B}_i(t)) + \varepsilon_i = \sum_{t\in T} \mathcal{I}^u(\mathcal{A}_i(t), \mathcal{B}_i(t)).\beta + \varepsilon_i,$$

which, combined with (2.2), yields $C^u_{i,(v,w)} = \sum_{t\in T} \mathcal{I}^u_{v,w}(\mathcal{A}_i(t), \mathcal{B}_i(t))$ and:

$$y_i = \sum_{v=1}^{n_\mathcal{A}} \sum_{w=1}^{n_\mathcal{B}} C^u_{i,(v,w)}\ \beta^u_{(v,w)} + \varepsilon_i \tag{2.4}$$

This model is a linear multiple regression model where the regressors are the frequencies associated to the joint class intervals $(I^u_\mathcal{A}(v) \times I^u_\mathcal{B}(w))_{v,w}$. The indicator functions $(\mathcal{I}^u_{v,w})_{v,w}$ define a functional basis.

From the contingency tables C^u_i, we construct the design matrix X^u associated to model (2.4) as follows. After vectorization (stacking column by column) and transposition of the contingency table C^u_i (see part *2* of Figure 2.1), we obtain, for a fixed partition vector $u = (n_\mathcal{A}, n_\mathcal{B})$, a row vector of length $n_\mathcal{A}.n_\mathcal{B}$:

$$X^u_i = {}^tVect(C^u_i)\,,\ X^u_i \in \mathbb{R}^{n_\mathcal{A} n_\mathcal{B}} \tag{2.5}$$

which represents the number of time observations t during which an individual i has been observed in each of the $n_\mathcal{A} \times n_\mathcal{B}$ levels described by the joint class intervals. The n stacked row vectors form the matrix $X^u = {}^t(X^u_1, X^u_2, \ldots, X^u_n) \in \mathbb{R}^{n\times n_\mathcal{A} n_\mathcal{B}}$.

2.3. Creation of a graph of contiguity constraints

To each matrix X^u corresponds a graph $G^u(V^u, E^u)$, which contains the contiguity constraints between modalities of the contingency table. V^u represents the columns (new variables) of the candidate matrix X^u and E^u all the edges connecting two close joint modalities. We used the Rook's case contiguity rule [24] where two joint modalities 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.

2.4. Selection of class intervals and related regression coefficients

The Fused Lasso is a variant of the Lasso introduced in 2005 by [35], in order to take into account the existence of a structure in the variables. In its original form, the Fused Lasso aims not only at parsimony of coefficients but also at parsimony of differences in consecutive coefficients. This version of the Fused Lasso can be interpreted as a one-dimensional Fused Lasso (1D-Fused Lasso). The Generalized Fused Lasso (GFL) [41] aims to promote smoothness over neighboring variables on a general graph $G = (V, E)$ made of V knots and E edges. Each explanatory variable corresponds to a node on the graph and an edge symbolizes the link between a pair of separate nodes in G.

For a fixed partition vector $u = (n_{\mathcal{A}}, n_{\mathcal{B}})$, the GFL criterion to minimize is written:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - X_i^u\beta)^2 + \lambda_p \sum_{j\in V^u}|\beta_j| + \lambda_f \sum_{(j,j')\ \in\ E^u}|\beta_j - \beta_{j'}| \tag{2.6}$$

with respect to β, where:

- $\beta = {}^t(\beta_{(1,1)}, \beta_{(2,1)}, \ldots, \beta_{(1,2)}, \ldots, \beta_{(n_{\mathcal{A}}, n_{\mathcal{B}})}) \in \mathbb{R}^{n_{\mathcal{A}} n_{\mathcal{B}}}$ the unknown coefficients,
- $\lambda_p \geq 0$ and $\lambda_f > 0$ the regularization parameters (of parsimony and fusion) to be optimized,
- for $j = (v, w)$ fixed, the couples (j, j') relative to j and contained in E^u are $(j, j')_1 = ((v, w), (v + 1, w))$ and $(j, j')_2 = ((v, w), (v, w + 1))$. In the following and depending on the context, the index j will refer either to the pair (v, w), or to the j^{th} element of the vector obtained from the matrix stored by columns.

The argmin solution of (2.6), denoted $\widehat{\beta}^u$, is computed as a function of the regularization parameters λ_p and λ_f, for a fixed value of u.

2.4.1. The Generalized Fused Lasso in the Generalized Lasso framework

If differences of contiguous coefficients were not penalized in (2.6) (i.e. if λ_f was zero), then the criterion would reduce to the Lasso criterion presented by [34]. The Lasso minimizes the residual sum of squares subject to the the constraint that the sum of the absolute value of the coefficients is less than a constant. In this case, there is only one regularization parameter to estimate. [6] proposed the Least-Angle Regression (LARS) algorithm able to solve the problem for all $\lambda \in [0, \infty[$, producing a full piece-wise linear solution path. The result of the path algorithm is the finite set of increasing λ values, where each λ delimits a model reduction dimension (the number of non-zero β components). Criterion (2.6) has two regularization parameters λ_p and λ_f to be optimized. The path algorithm is no longer suitable to identify them.

Our proposal is to parameterize (2.6) with the ratio:

$$\gamma = \frac{\lambda_p}{\lambda_f} \tag{2.7}$$

(choice made in the R package genlasso [3] used for implementing SpiceFP). This ratio represents a balance between parsimony and fusion. Then, for a fixed value of γ, criterion (2.6) can be equivalently rewritten as:

$$\frac{1}{2}||y - X^u\beta||_2^2 + \lambda||D^{u,\gamma}\beta||_1, \tag{2.8}$$

where $y = {}^t(y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is the response vector and $D^{u,\gamma}$ is a specified penalty matrix (see below). This model corresponds to the Generalized Lasso model, introduced by [36] as an encapsulation of statistical models using the L_1 norm to impose additional constraints. Through this new parametrization, the value of $\lambda = \lambda_f$ can be optimized with the path algorithm and a corresponding $\widehat{\beta}^{u,\gamma}(\lambda)$ can be estimated. Finally, there are as many pairs $(\lambda, \widehat{\beta}^{u,\gamma}(\lambda))$ solutions as there are u and γ parameters set. So, several models are available and we have to select one of them in order to deduce the optimal pair. Selection of the best model is done with an information criteria which requires an estimation of the degree of freedom for each model. Parameter λ_p will be deduced from γ through $\lambda_p = \gamma\lambda_f$.

In our context, the penalty matrix $D^{u,\gamma}$ is a row-binding of two sub-matrices, namely:

- $D^{u,\gamma,p} \in \mathbb{R}^{n_{\mathcal{A}}n_{\mathcal{B}} \times n_{\mathcal{A}}n_{\mathcal{B}}}$: the penalty sub-matrix associated to the regularization of parsimony $\left(\lambda_p \sum_{j \in V^u} |\beta_j|\right)$
- $D^{u,f} \in \mathbb{R}^{2n_{\mathcal{A}}n_{\mathcal{B}} \times n_{\mathcal{A}}n_{\mathcal{B}}}$: the sub-matrix associated to the regularization of the fusion according to the two dimensions $\left(\lambda_f \sum_{(j,j') \in E^u} |\beta_j - \beta_{j'}|\right)$. Affecting two dimensions, $D^{u,f}$ can be subdivided into $D^{u,f1}$ and $D^{u,f2}$.

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

$$
\begin{aligned}
D^{u,f1}_{(v,w)(v',w')} &= \begin{cases} 1 & \text{if } (v',w') = (v+1,w) \\ -1 & \text{if } (v',w') = (v,w) \text{ and } v < n_{\mathcal{A}} \\ 0 & \text{if not} \end{cases} \\
D^{u,f2}_{(v,w)(v',w')} &= \begin{cases} 1 & \text{if } (v',w') = (v,w+1) \\ -1 & \text{if } (v',w') = (v,w) \text{ and } w < n_{\mathcal{B}} \\ 0 & \text{if not} \end{cases} \\
D^{u,\gamma,p} &= \gamma.\mathbb{I}_{n_{\mathcal{A}}.n_{\mathcal{B}}}
\end{aligned}
\tag{2.9}
$$

where $\gamma \geq 0$ and $\mathbb{I}_{n_{\mathcal{A}}n_{\mathcal{B}}}$ is an identity matrix.

The Generalized Fused Lasso presented in criterion (2.6) is a 2 Dimensional - Sparse Fused Lasso (2d-SFL).

2.4.2. Degrees of freedom of the Generalized Fused Lasso fit

Our approach follows that of [37], who established the calculation of the degree of freedom for any Lasso-type regression written as a generalized Lasso problem as presented in (2.8).

We first introduce some notations. For any penalty matrix $D \in \mathbb{R}^{m \times p}$ involved in a generalized Lasso problem of type (2.8), let $\mathcal{S}$ be the active set corresponding to a particular solution $\hat{\beta}$, it is defined as:

$$\mathcal{S} = \{r \in \{1, \ldots, m\} : (D\hat{\beta})_r \neq 0\} = support(D\hat{\beta}).$$

Let $D_{-\mathcal{S}}$ be the matrix D from which were removed the rows indexed by $\mathcal{S}$. And let $null(D_{-\mathcal{S}})$ be the null space or kernel of $D_{-\mathcal{S}}$.

The calculation of the degrees of freedom of the generalized Lasso fit is stated in [37, Theorem 3] and reminded here :

Theorem (Generalized lasso degrees of freedom - Tibshirani and Taylor (2012)) *Assume that $y \in \mathbb{R}^n$ follows a normal distribution ($y \sim N(\mu, \sigma^2 I)$ with given (unknown) mean vector $\mu \in \mathbb{R}^n$ and marginal variance σ^2). For any fixed and nonrandom predictor matrix $X \in \mathbb{R}^{n \times p}$, penalty matrix $D \in \mathbb{R}^{m \times p}$ and $\lambda \geq 0$, the degree of freedom of the generalized Lasso fit can be expressed as*

$$df(X\hat{\beta}) = E[dim(X(null(D_{-\mathcal{S}})))] \tag{2.10}$$

with $\mathcal{S} = \mathcal{S}(y)$ the active set corresponding to any generalized Lasso solution $\hat{\beta}(y)$ at y.

The notation $X(V)$ represents the image space of a subspace V by X. It is the space generated by the columns of the X matrix projected on V. The assumptions required for this Theorem are those usually made in regression estimations (Gaussian i.i.d. errors). No assumptions are made on matrix D nor on X. This result can thus be applied to the 2d-Sparse Fused Lasso constraint. To compute $df(X^u\hat{\beta}^{u,\gamma})$, we need $\beta^{u,\gamma}$, X^u, $D^{u,\gamma}$, λ defined in (2.8) and γ in (2.7). Note that, in our context, $m = 3n_{\mathcal{A}}n_{\mathcal{B}}$ the number of rows in $D^{u,\gamma}$ and $p = n_{\mathcal{A}}n_{\mathcal{B}}$ the number of columns in X^u.

With fused lasso constraints, we are interested in sets of parameters that share the same value. By this, we refer to connected components, which are illustrated on Figure 2.4.2 and defined as follows:

Definition 2.1. A **connected component** *cc* is a set of indexes of non zero coefficients $\beta^{u,\gamma}_{v,w}$ that are linked together via the $D^{u,\gamma}$ matrix (2.9) and that all share the same real value.

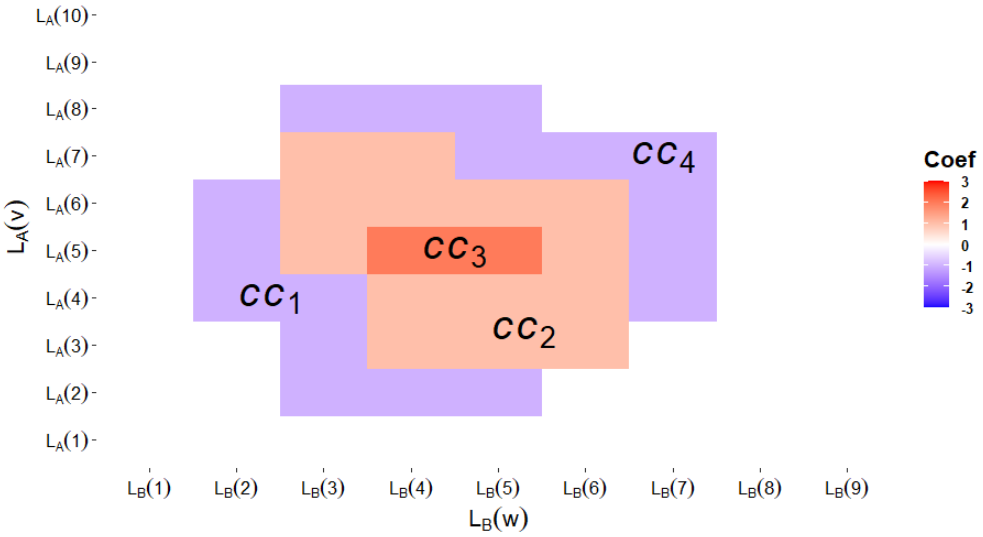


Figure 2.2: Example of $\beta^{u,\gamma}$ coefficient values with 4 connected components (here $u = (9, 8)$).

Two coefficients $\beta^{u,\gamma}_{v,w}$ and $\beta^{u,\gamma}_{v',w'}$ are linked via the $D^{u,\gamma}$ matrix if and only if $|v' - v| = 1$ or (exclusive) $|w' - w| = 1$ (i.e., if and only if $|v' - v| + |w' - w| = 1$). Let us consider that the estimated coefficient $\widehat{\beta}^{u,\gamma}$ contains $Q^{u,\gamma}$ connected components denoted cc_q, $q = 1, 2, \ldots, Q^{u,\gamma}$. In order to identify the respective coefficients involved in each non-zero connected component, we introduce the matrix Θ through Definition 2.2.

Definition 2.2. Let $(cc_q)_{q=1,\ldots,Q}$ be a set of connected components. The **connected component membership matrix** is a binary matrix $\Theta \in \mathbb{R}^{n_{\mathcal{A}}n_{\mathcal{B}} \times Q}$ whose q-th column $\Theta^{(q)}$ indicates the membership or not of the $\hat{\beta}^{u,\gamma}$ components to the connected component cc_q, as follows:

$$\forall j = 1, \ldots, n_{\mathcal{A}}n_{\mathcal{B}},\ \Theta^{(q)}_j = \begin{cases} 0 & \text{if} \quad j \notin cc_q \\ 1 & \text{if} \quad j \in cc_q \end{cases},\ q = 1, \ldots, Q. \tag{2.11}$$

Let's now compute the null space of $D_{-\mathcal{S}}$, denoted $null(D_{-\mathcal{S}})$:

$$null(D_{-\mathcal{S}}) = \{\Theta \in \mathbb{R}^{n_{\mathcal{A}}n_{\mathcal{B}} \times Q} | D_{-\mathcal{S}}\Theta^{(q)} = 0_{(m-s)}, q = 1, \ldots, Q\} \tag{2.12}$$

where $D_{-\mathcal{S}} \in \mathbb{R}^{(m-s)\times n_{\mathcal{A}} n_{\mathcal{B}}}$, $D_{-\mathcal{S}}\Theta^{(q)} \in \mathbb{R}^{m-s}$ and $s = Card(S)$.

As the $D^{u,\gamma}$ matrix has a simple structure, adapting Theorem 3 of [37] in the context of 2d-Sparse Fused Lasso is equivalent to looking for the components of $\widehat{\beta}^{u,\gamma}$ which have different values. This is the subject of the next corollary.

Corollary 2.3. *The degree of freedom $\widehat{df}(X\widehat{\beta}^{u,\gamma})$ associated to the criterion (2.8) in the context of 2d-Sparse Fused Lasso is equal to the number of connected components $Q^{u,\gamma}$.*

Proof:
$$\begin{aligned}\widehat{df}(X\widehat{\beta}^{u,\gamma}) &= dim\left(X^u(null(D^{u,\gamma}_{-\mathcal{S}}))\right) = dim\left(X^u(\Theta)\right)\\ &= dim\left(Vect\left\{X^u\Theta^{(q)},\ q=1,\dots,Q^{u,\gamma}\right\}\right) = rank([X^u\Theta^{(1)},\dots,X^u\Theta^{(Q^{u,\gamma})}]) = Q^{u,\gamma}\end{aligned}$$

where we omitted Θ dependencies on u and γ to lighten the notations in the corollary.

2.4.3. Choice of the best candidate matrix and selection of its variables

SpiceFP requires the construction of different candidate explanatory matrices X^u from both functional variables and partition's vector u. Constructing a GFL for a matrix of predictors associated to a fixed u requires identifying the optimal values of the penalty parameters: λ and γ in (2.8). In penalized regressions, cross-validation is often used to optimize regularization parameters, but it is time consuming. We suggest using an information criterion to achieve the same purpose [9]. To that aim, we computed an adapted information criterion for each model indexed by u, λ and γ. The best model is obtained by minimizing the information criterion chosen, which yields the best partition $\widehat{u}$ and allows to select the best variables from $X^{\widehat{u}}$. The variable selection is done through the selection of $\widehat{\gamma}$ and $\widehat{\lambda}$ and the associated non-zero $\widehat{\beta}$ components are deduced.

There exist various information criteria including Akaike Information Criterion (AIC) [2] and Bayesian Information Criterion (BIC) [31]. These criteria penalize log-likelihood by the number of model parameters. The BIC also penalizes log-likelihood by the sample size. A well defined penalization is essential to compare regression models involving different explanatory matrices. These information criteria require computing the degree of freedom $Q^{u,\gamma}$ of the GFL model, obtained in Corollary 2.3.

For each u and γ defined on respective grids (given by the users) and λ taken from the set $(\lambda_e)_{e=1,\dots,N_\lambda}$ delivered by the path algorithm, we considered the following information criteria:

- $AIC^{u,\gamma}_e = -2\log(L^{u,\gamma}_e) + 2Q^{u,\gamma}_e$
- $BIC^{u,\gamma}_e = -2\log(L^{u,\gamma}_e) + \log(n)Q^{u,\gamma}_e$

with $L^{u,\gamma}_e$ the likelihood function of the following model: $y = X^u\beta^{u,\gamma}_e + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\ \sigma^2 I)$, associated with criterion (2.8). We have:

$$-2\log(L^{u,\gamma}_e) = 2n\log(\sigma) + n\log(2\pi) + \frac{1}{\sigma^2}||y - X^u\beta^{u,\gamma}_e||^2_2$$

where the variance of the residuals σ^2 is unknown. As mentioned by [14], the same variance estimator must be used in the calculation of the criteria for all the constructed models, and at each iteration (see bellow). We thus decided to estimate σ^2 by the variance of the response variable: $\widehat{\sigma}^2 = \frac{1}{n-1}||y-\overline{y}||^2_2$. It's a biased estimator of σ^2, but this bias remains fixed for all models compared [14]. Such an estimator leads to an overestimation of the variance, which penalizes the introduction of new coefficients in the model. This bias can be partly offset by an iterative approach. That's why we have implemented an iterative algorithm where residuals are used as new response variable and so on, until shutdown conditions are verified.

2.5. SpiceFP: an iterative approach

To capture all potential non-zero coefficients, one should take a fine partition with values of $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ large enough. This could imply a very low or even zero number of points in the joint class intervals (making the method ineffective) and prohibitively long computation times [20]. As a trade off between thinness and work-ability we chose to develop an iterative approach to explore a large space of solutions (that allows addition of different thinness of partition).

Data required for the use of SpiceFP are the functional explanatory variables $\mathcal{A}_i$ and $\mathcal{B}_i$ discretized on a grid T and a response variable y_i with $i = 1, \ldots, n$. Other elements are also required at the input of the algorithm: Γ, a set of positive reals representing γ ratios of regularization parameters, $\mathcal{U}_{\mathcal{A}}$ and $\mathcal{U}_{\mathcal{B}}$ the sets of numbers of class intervals n_A and n_B, n_λ the selected number of pairs (among N_λ) (λ, $\widehat{\beta}^{u,\gamma}(\lambda)$), the information criterion to be used, and K the maximum number of iterations to explore. The n_λ values of λ are chosen equally spaced on the log scale (see Genlasso package [3]).

SpiceFP constructs for each couple (u, γ), a matrix of explanatory variables X^u and a penalty matrix $D^{u,\gamma}$. This first step (line 1 to 13, algorithm 1) is performed only once and the set of candidate explanatory matrices remains unchanged. The second step (line 14 to 34, algorithm 1) of the approach is iterative. For each candidate matrix a solution path (λ, $\widehat{\beta}^{u,\gamma}(\lambda)$) is obtained. Once the criterion associated with each model is computed, the optimal triplet $(\widehat{u}, \widehat{\gamma}, \widehat{e})$ is the argument that minimizes $Crit_e^{u,\gamma}$, where e is the index of the solution path-coefficients $(\lambda_e, \hat{\beta}_e^{u,\gamma})$. $X^{\widehat{u}}$, $D^{\widehat{u},\widehat{\gamma}}$ and $\widehat{\beta}_{\widehat{e}}^{\widehat{u},\widehat{\gamma}}$ respectively represent the optimal matrices of the explanatory, penalty and coefficient variables. At each iteration k, we denote $u^k = \widehat{u}$ and $\beta^k = \widehat{\beta}_{\widehat{e}}^{\widehat{u},\widehat{\gamma}}$. SpiceFP then checks if the selected coefficients according to the criterion $Crit_e^{u,\gamma}$ correspond to a zero vector or if the maximum number of iterations K is reached. The algorithm is stopped when at least one of these conditions is verified. When none of these conditions are verified, the residuals of the optimal model are computed and used as a response variable at the next iteration of the algorithm.

The third step (line 35, algorithm 1) gives the final prediction as the sum of all the predictions obtained at each iteration: $\sum_{k=1}^{k_*} X^{u^k}\beta^k$. The function corresponding to the final prediction can be written as follows :

$$F_* = \sum_{k=1}^{k_*} F_{u^k,\beta^k} = \sum_{k=1}^{k_*} \mathcal{I}^{u^k}.\beta^k \tag{2.13}$$

hence the predicted value is given by: $E(y_i) = \sum_{t\in T} F_*(\mathcal{A}_i(t), \mathcal{B}_i(t))$. We remark that each vector β^k at each iteration may have different length $dim(u^k)$.

2.6. Adaptation of SpiceFP to the partitioning of functional variables according to a non-linear scale

A linear partitioning of type (2.1) of the functional variables is not always suitable. Other types of partitioning may have to be chosen. Assuming for example that functional variable $\mathcal{B}$ requires a partitioning according to a logarithmic scale, the following breaks can be used instead of Equation (2.1):

$$L_{\mathcal{B}}(w) = \underline{\mathcal{B}} + \frac{1}{\alpha_{\mathcal{B}}}\left(\left(1 + \alpha_{\mathcal{B}}(\bar{\mathcal{B}} - \underline{\mathcal{B}})\right)^{\frac{w-1}{n_{\mathcal{B}}}} - 1\right), w = 1, \ldots, n_{\mathcal{B}} + 1, \tag{2.14}$$

where parameters $\alpha_{\mathcal{B}} > 0$ and $n_{\mathcal{B}}$ have to be set to determine $L_{\mathcal{B}}(w)$. For a fixed $n_{\mathcal{B}}$, high $\alpha_{\mathcal{B}}$ value is related to high proportion of breaks close to $\underline{\mathcal{B}}$ and vice versa. So, two partitioning parameters

Input : K; $\mathcal{U}_{\mathcal{A}}$;$\mathcal{U}_{\mathcal{B}}$; Γ; n_λ; $Crit \in \{AIC, BIC\}$

Output: $k_*, (u^k)_{k \leq k_*}$; $(\beta^k)_{k \leq k_*}$

Data : $\{y_i, i = 1, \ldots, n\}$; $\{\mathcal{A}_i(t)$; $\mathcal{B}_i(t)$, $i = 1, \ldots, n$, $t \in T\}$

foreach $u \in \mathcal{U}_{\mathcal{A}}$. $\mathcal{U}_{\mathcal{B}}$ **do**
  for $i \leftarrow 1$ **to** n **do**
    for $v \leftarrow 1$ **to** $n_{\mathcal{A}}$ **do**
      for $w \leftarrow 1$ **to** $n_{\mathcal{B}}$ **do**
        $C^u_{i,(v,w)} = Card\ \{t \in T | \mathcal{A}_i(t) \in I^u_{\mathcal{A}}(v), \mathcal{B}_i(t) \in I^u_{\mathcal{B}}(w)\}$
      end
    end
    $X^u[i,] = {}^tVect(C^u_i)$
  end
  foreach $\gamma \in \Gamma$ **do**
    Construct $D^{u,\gamma}$ as shown in the equation (2.9)
  end
end
for $k \leftarrow 1$ **to** K **do**
  foreach $u \in \mathcal{U}_{\mathcal{A}}$. $\mathcal{U}_{\mathcal{B}}$ **do**
    Center y and each joint modality in X^u
    foreach $\gamma \in \Gamma$ **do**
      Find the solution path $\widehat{\beta}^{u,\gamma}(\lambda) = \underset{\beta \in \mathbb{R}^{n_{\mathcal{A}} n_{\mathcal{B}}}}{\operatorname{argmin}} \frac{1}{2}||y - X^u\beta||_2^2 + \lambda||D^{u,\gamma}\beta||_1$
      Select n_λ equally spaced couples $(\lambda_e, \widehat{\beta}^{u,\gamma}_e)$ on the log scale with respect to λ over the solution path and compute $Crit^{u,\gamma}_e$ for $e = 1 \ldots n_\lambda$:
      $AIC^{u,\gamma}_e = \frac{1}{\sigma^2}||y - X^u\widehat{\beta}^{u,\gamma}_e||_2^2 + 2Q^{u,\gamma}_e$ or
      $BIC^{u,\gamma}_e = \frac{1}{\sigma^2}||y - X^u\widehat{\beta}^{u,\gamma}_e||_2^2 + \log(n)Q^{u,\gamma}_e$
    end
  end
  $(\widehat{u}, \widehat{\gamma}, \widehat{e}) \leftarrow \underset{u \in \mathcal{U}_{\mathcal{A}}.\mathcal{U}_{\mathcal{B}},\ \gamma \in \Gamma,\ 1 \leq e \leq n_\lambda}{\operatorname{argmin}} Crit^{u,\gamma}_e$
  $u^k = \widehat{u},\ \beta^k = \widehat{\beta}^{\widehat{u},\widehat{\gamma}}_{\widehat{e}}$
  if $\beta^k \neq 0_{n_{\mathcal{A}} n_{\mathcal{B}}}$ **then**
    $k_* = k$
    if $k < K$ **then**
      $y \leftarrow\ y$ - $X^{u^k}\beta^k$
    end
  else
    Leave the loop in k
  end
end
Construct the result: $\sum_{k=1}^{k_*} X^{u^k}\beta^k$

Algorithm 1: SpiceFP algorithm

($n_{\mathcal{B}}$ and $\alpha_{\mathcal{B}}$) should be optimized for partitioning $\mathcal{B}$. In this case, the partition vector is written $u = (n_{\mathcal{A}}, n_{\mathcal{B}}, \alpha_{\mathcal{B}})$. Let $\mathcal{V}_{\mathcal{B}}$ be the set containing the possible values of $\alpha_{\mathcal{B}}$. $\mathcal{V}_{\mathcal{B}}$ is an additional input to the algorithm. The only change in the core of the algorithm is to replace $u \in \mathcal{U}_{\mathcal{A}}.\ \mathcal{U}_{\mathcal{B}}$ by $u \in \mathcal{U}_{\mathcal{A}}.\ \mathcal{U}_{\mathcal{B}}.\ \mathcal{V}_{\mathcal{B}}$ (lines 1 to 15).

Various types of breaks (chosen between Equations (2.1), (2.14) or user-defined) can be used for partitioning explanatory functional variables and the SpiceFP algorithm can be easily adapted.

3. Use Case: Grapevine dataset

3.1. Data presentation

Data were collected during an experiment conducted in a vineyard of the Institut Agro campus at Montpellier in 2014 (Syrah vines). The aim was to study the influence of the micro-climate (temperature, solar irradiation) at the grape level on the anthocyanin content of the berries. Experts in viticulture assume that the accumulation of chemical compounds affecting the quality of the grape berry is jointly influenced by these initial explanatory variables. This assumption is reinforced by results of [33], which underlined that the anthocyanin composition of Merlot grapes was influenced by a complex combined effect of berry temperature and solar irradiation.

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 basis of the plant, and eight under control conditions (without open-top chambers). The chambers were made up of 2 translucent polycarbonate panels placed on the ground at about 10cm below the bunches on each side of the vines and inclined to form a mini two-pitched greenhouse roof open at its ridge. 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 at bunch level was recorded through the measurement of temperature and irradiance. According to [38], solar radiation can be characterized by three different quantifiers, including Photosynthetic Photon Flux Density (PPFD), measured in $10^{-6} mol\ m^{-2}\ s^{-1}$. It corresponds to the number of incident photons useful for photosynthesis, received per unit of time on a horizontal surface unit. Rows were roughly oriented south-north, and irradiance was separately measured on bunches located on the east and west side of the row. This measurement system therefore enabled the observation of different modalities of the couple (temperature, irradiance) affecting the grape berries. Temperature and Irradiance were recorded every twelve minutes throughout the maturation period when anthocyanins are known to accumulate.

So the experimental design contained: rows (1, 2, 3) $\times$ 2 sun exposure orientations (East, West) $\times$ 16 vine stocks = 96 "statistical individuals". Anthocyanin contents were measured weekly via the Ferari Index FI_i for each individual i. This is a non-destructive measure of anthocyanin content of the bunches [1]. The objective of our study is to understand how the couple (Temperature, Irradiance) acts on a weekly variation of the Ferari Index ΔFI.

Temperature and Irradiance are variables of different natures. Temperature is a variable whose variations are regular enough to be partitioned according to Equation (2.1). The Irradiance variable is partitioned according to (2.14), as explained in the following subsection.

3.2. Partitioning of the Irradiance variable

Observed on a one-day scale, PPFD 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 not partitioned according to a linear scale

as proposed by Equation (2.1), but rather according to a logarithmic scale as in Equation (2.14). The use of logarithmic transformation has consistently been used in the development of models involving solar radiation [30], [23], [4]. ¨

3.3. Objective

The objective of our study was to identify the ranges of temperature and irradiance that jointly influence or not the accumulation of anthocyanins between sunrise and noon. This study is an appropriate application framework to test the SpiceFP method.

Our study is composed of two parts. First a simulation part is presented in section 4. Temperature and Irradiance, measured during the week of July 17 to 24, 2014, are the input variables of a model which simulates output variables to be predicted, using two different known β: one with two distinct patches of coefficients and another with a concentric gradient of coefficients (Table 2, column 1). This simulation study made it possible to evaluate the functioning and performance of the SpiceFP algorithm. Second, in section 5, the algorithm was tested on a complete dataset, obtained from July 24 to August 1, 2014, to understand the effect of Temperature and Irradiance interaction on ΔFI. Dataset and script are available online at https://forgemia.inra.fr/exploratory-penalized-regression/paper-script-and-data.git.

4. Simulation study

We present in this section simulations that help to better understand the SpiceFP characteristics. Remember that the approach must be able to identify an optimal partition or joint class intervals used as linear regressors.

4.1. Simulation design and SpiceFP setting

In order to carry out the simulations, a few steps were required:

- We considered the observations of temperature ($\mathcal{A}$) and irradiance ($\mathcal{B}$) in the Vine dataset obtained between sunrise and noon during the week of July 17 to 24, 2014.
- We then arbitrarily set a partition vector $u^0 = (n_{\mathcal{A}} = 17, n_{\mathcal{B}} = 20, \alpha_{\mathcal{B}} = 0.05)$.
- From $\mathcal{A}$, $\mathcal{B}$, and u^0 we constructed X^{u^0} using equations (2.2) and (2.5). X^{u^0} was computed based on observed data set in order to respect realistic frequencies for the joint class intervals.
- Based on dimensions of X^{u^0}, we drew coefficients β^{u^0} from a random distribution (values are available on git@forgemia.inra.fr) and computed the response variable of the simulation:

$$Y = X^{u^0}\beta^{u^0} + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I). \tag{4.1}$$

After that, we used the SpiceFP approach to make the estimation.

In this simulation study, we chose to simulate two different coefficient vectors. They are as followed:

- Coefficients β^{u_0} for simulation 1 is made of two distinct patches (column 1, row 1 of the Table 2). Two noise levels were used and presented in table 1.
- Coefficients β^{u_0} for simulation 2 is made of a concentric gradient of coefficients (column 1, row 2 of the Table 2). Two noise levels were also used and presented in Table 1.

Table 1: Noise simulation design

	Low noise	High noise
Simulation 1	$\sigma_\varepsilon = 1.50$ $(\sigma^2_\varepsilon/\sigma^2_Y = 0.03)$	$\sigma_\varepsilon = 2.5$ $(\sigma^2_\varepsilon/\sigma^2_Y = 0.08)$
Simulation 2	$\sigma_\varepsilon = 0.25$ $(\sigma^2_\varepsilon/\sigma^2_Y = 0.01)$	$\sigma_\varepsilon = 1 \quad (\sigma^2_\varepsilon/\sigma^2_Y = 0.14)$

We thus generated four simulation datasets (1 matrix $X^{u_0} \times 2$ simulated coefficient vectors $\times$ 2 noise levels).

The inputs required by the algorithm were as follows:

- $\mathcal{U}_\mathcal{A} = \{15, 16, \ldots, 20\}$
- $\mathcal{U}_\mathcal{B} = \{18, 19, \ldots, 22\}$
- $\mathcal{V}_\mathcal{B} = \{0.015, 0.05, 0.135, 0.405, 1.215\}$
- $K = 2$
- $\Gamma = \{0.0001, 0.05, 0.15, 0.45, 2, 8\}$
- $n_\lambda = 100$

4.2. Simulation results

The results of the simulations are presented in Tables 2 to 5 (Tables 4 and 5 are in the appendix). Table 2 is related to the estimated coefficient vectors and Tables 4 and 5 present respectively histograms of the residuals and scatter plots showing the quality of the estimates. For each of the four response variables (one per row), three estimations are provided (columns 3 to 5, Table 2), computed as follows:

- column 3: at the first iteration, an exploratory matrix X^{u^1} indexed by a partition vector u^1 is identified. The estimated coefficient at iteration 1 of SpiceFP is noted β^1 (see algorithm 1, line 25). The estimated response $\widehat{Y}^1 = X^{u^1}\beta^1$ and the residuals $\varepsilon^1 = Y - \widehat{Y}^1$ are computed. Figures in column 3 provide the visualization of β^1. For a suitable visualization, the vector of coefficients is transformed into a matrix of dimension $n_\mathcal{A} \times n_\mathcal{B}$, where $n_\mathcal{A}$ et $n_\mathcal{B}$ are the numbers of class intervals associated to u^1.

- column 4: at the second iteration, an exploratory matrix X^{u^2} indexed by a partition vector u^2 is identified. The estimated coefficient at iteration 2 of SpiceFP is noted β^2. The response variable used at the second iteration in the model is ε^1, the residuals obtained at iteration 1. The estimated response $\widehat{Y}^2 = X^{u^1}\beta^1 + X^{u^2}\beta^2$ and the residuals $\varepsilon^2 = \varepsilon^1 - X^{u^2}\beta^2$ are computed. Figures in column 4 give the visualization of $F(a, b) = F^1(a, b) + F^2(a, b)$ where F is defined as in Equation (2.3). In the case of the visualizations in Table 4, 500 equally spaced values of $a \in [\underline{\mathcal{A}}, \bar{\mathcal{A}}]$ and 1000 equally spaced values $b \in [\underline{\mathcal{B}}, \bar{\mathcal{B}}]$ were used.

- column 5: the results presented in this column are slightly out of the scope of the SpiceFP approach and concerns only the outputs of iteration 1. The idea here is to consider an average of the 1% best models. The best models are defined in the sense of the information criterion. Best coefficients relative to the different partitions available in these 1% best models are identified. Let's assume that there are n_m. They are then indexed by $u^{1,(1)}, u^{1,(2)}, \ldots, u^{1,(n_m)}$. The coefficients of interest can be noted $\beta^{1,(1)}, \beta^{1,(2)}, \ldots, \beta^{1,(n_m)}$. The estimated response

is $\widehat{Y}^{1,(1:n_m)} = \frac{1}{n_m} \sum_{m=1}^{n_m} X^{u^{1,(m)}} \beta^{1,(m)}$. All the selected models have the same weight in the computation of this average. The residuals can be obtained by $\varepsilon^{1,(1:n_m)} = Y - \widehat{Y}^{1,(1:n_m)}$. Figures in column 5 show the visualization of $\frac{1}{n_m} \sum_{m=1}^{n_m} F^{1,(m)}(a,b)$ where $F^{1,(m)}$ is the function related to model coefficients indexed by $u^{1,(m)}$.

Black color refers to never-observed joint modalities whereas white color refers to joint modalities with null estimated coefficients. The histograms of residuals related to each estimation are presented in Table 4. On top of these histograms, the ratio between the variance of residuals and the variance of response variable are provided. Table 5 shows the goodness of fit of each estimation, by presenting the slope of the regression "predicted versus simulated variables". The closer to 1 the slope is and the closer to 0 the ratio of variances is, the better the estimate is.

From the estimates provided by the algorithm (Table 2), we notice that SpiceFP effectively identifies the simulated zones of influence and assigns the right color to the coefficients: the graphics show two distinct areas, one with positive coefficients (red area), the other with negative coefficients (blue area). The approach tends to assign the same value (same colour) to groups of estimated coefficients, although with a gradient within the group. This behavior can be explained on the one hand by the Generalized Fused Lasso which penalizes the difference between two related coefficients and on the other hand by the variance estimate used for computing the information criterion. This variance was overestimated, which penalized the introduction of new coefficients into the model. With respect to the noise level contained in the response variable, we note that the more noisy Y is, the more false positives are observed. Looking at the differences in the outputs of iterations 1 and 2, we observe that coefficients are added to those previously obtained. These new coefficients lead to improved estimation when the noise is low. When the noise is high, the β estimate of the second iteration has non-zero elements in some areas that are not relevant: for example for high temperature and irradiance values. The average of the 1% best models underestimates, in both simulations, the amplitude of the β values but, in simulation 2, it restores the β support much better than the estimators obtained in iterations 1 or 2.

5. Modeling the evolution of a grape berry quality index

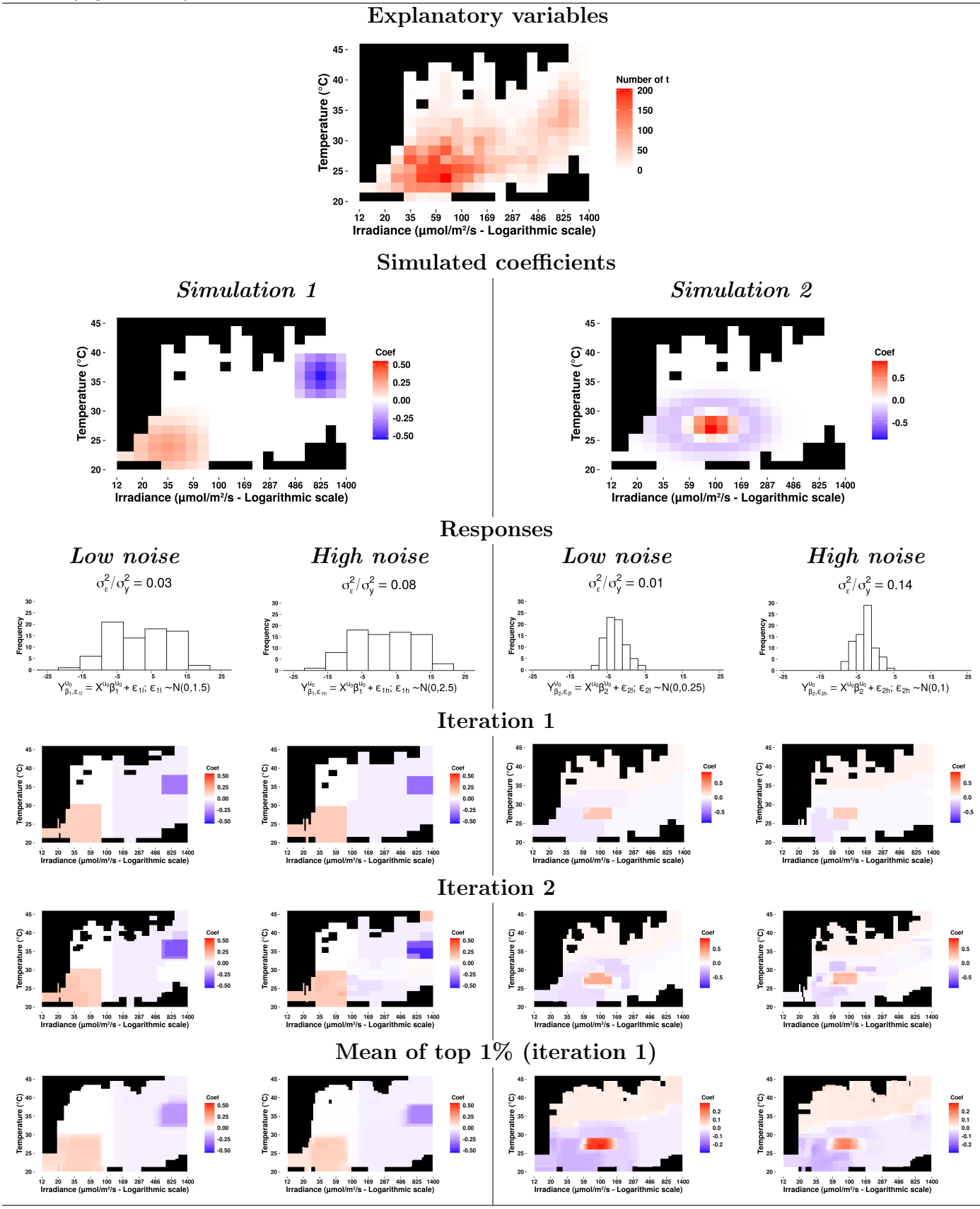
5.1. Methodology used for data analysis

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

Most of the photosynthetic functioning of the grapevine happens in the morning. The time period between sunrise and noon is denoted T_1 below. The following variables and parameters are the input objects of SpiceFP.

- $y_i = \Delta FI_i$, $\mathcal{A}_i(t)$; $\mathcal{B}_i(t)$; $i = 1, \ldots, n_1$; $t \in T_1$;
- $\mathcal{U}_\mathcal{A} = \mathcal{U}_\mathcal{B} = \{10, 11, 12, \ldots, 29, 30\}$
- 10 values were chosen between 0.0025 and 1 for α based on an exponential function $\alpha \in \{0.0025, 0.0048, 0.0094, 0.0183, 0.0357, 0.0695, 0.1353, 0.2636, 0.5134, 1.0000\}$
- $K = 3$

Table 2: Simulation results: estimation with the SPICEFP algorithm of two coefficient functions β with two types of noise (high and low). ■: Joint modalities that have never been observed.



- $\Gamma = \{0.001, 0.01, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6\}$
- $n_\lambda = 20$

5.2. Results

The results are presented in Table 3. The first column shows the estimated coefficients, the second column the histograms of residuals and the last column the scatter plots related to the goodness of fit. For this data set, SpiceFP stopped at the third iteration with a shutdown rule of zero coefficients.

The first row of Table 3 contains the results observed at the first iteration of SpiceFP. In terms of model quality, the slope is 0.558 and the residuals follow a normal distribution centered in 0. The visualization of the coefficients indicates conditions (irradiance $< 100\ \mu mol\ m^{-2}\ s^{-1}$, temperature from 15°C to 33°C) that affect negatively the Ferari Index. The second row of Table 3 is relative to the results of the second iteration. Due to the fact that the coefficients retained in iteration 3 are all zero, the result of the approach is therefore that of row 2. The slope of model goodness of fit is 0.678 and the residuals follow a normal distribution centered in 0. The visualization of the coefficients indicates, in addition to what is observed at iteration 1, a small positive influence (compared to the amplitude of the negative effect) of the temperature below 30°C for an irradiance higher than $100\ \mu mol\ m^{-2}\ s^{-1}$.

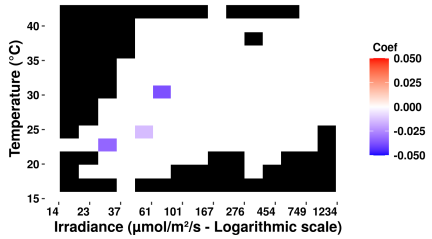
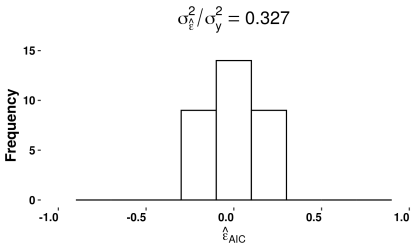
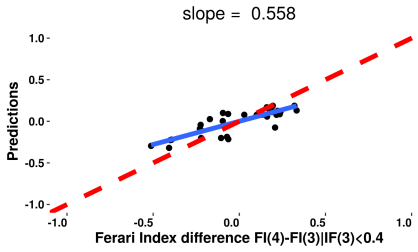
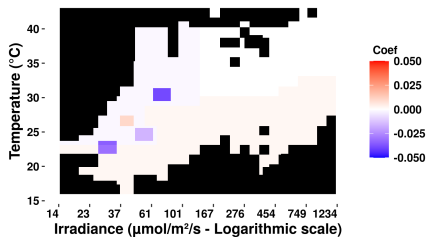
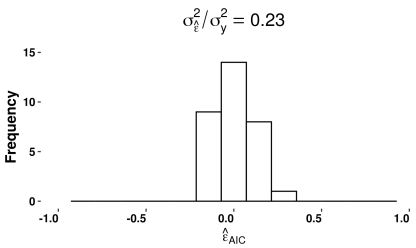
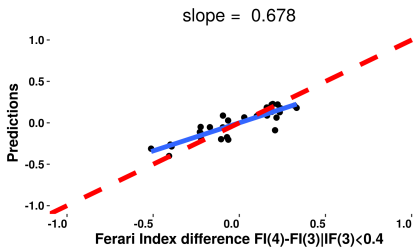
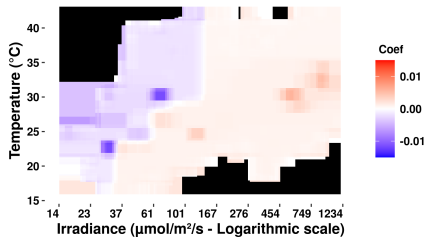
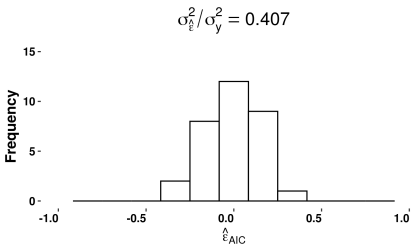
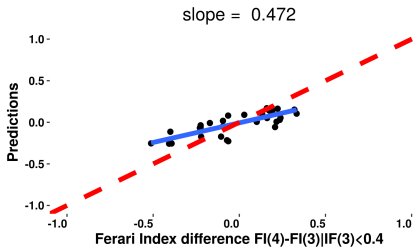
When focusing on the average of the 1% best models (presented in the third row), we remark that the quality of this model (as indicated by the slope : 0.472) is not better than that of the models chosen by SpiceFP at iteration 1 or 2. However, it should be noted that the model obtained has more fused coefficients, which allows the identification of a border zone between the positive and negative zones of influence.

As an interpretation, we can note that, in the morning (sunrise to noon), for low irradiance values ($< 100\ \mu mol\ m^{-2}\ s^{-1}$), there is a range of temperature values that are 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 below 30°C is suitable for increasing the Ferari index. The average of the coefficients shows, within each of the non-zero zones of influence, a relative variability of the coefficients, suggesting the importance of some temperature and irradiance amplitudes. The slope coefficients of the goodness of fit of the models are away from 1. It should be reminded that the response variable is studied with respect to the variations of temperature and irradiance only between sunrise and twelve. There is also a lot of information hidden in the residuals.

6. Discussion

The SpiceFP approach is a scalar-on-function approach. The response variable is real and predictors are functional variables. The approach is based on a transformation of functional variables, which yields a contingency table. To construct this table, it is assumed that no data are missing and that the times of observation are identical for both functional variables. The method takes into account the multicolinearity resulting from the auto-correlations existing in the processes. Moreover, the constructed candidate explanatory matrices are not nested but they covered the same domain. That's why model and variable selection methods (Fused Lasso, information criterion selection) had to be adapted and generalized to the framework of SpiceFP. In the implementation of the approach, all the candidate explanatory matrices are evaluated at each iteration. This is not due to a re-sampling as is the case for bootstrap or bagging methods, but to the evaluation of different partitions.

Table 3: Visualization of the combined effects of irradiance and temperature on the Ferari index (From sunrise to twelve, ■: joint modalities that have never been observed). Rows 1 and 2 present the results of iterations 1 and 2, respectively. The third row presents the average of the 1% best models obtained at iteration 1.

Estimation of coefficients	*Histogram of residuals*	*Goodness of fit*
Iteration 1		
		
Iteration 2		
		
Mean of top 1% (iteration 1)		
		

6.1. SpiceFP: a functional approach

In recent years, several studies in frequentist and Bayesian statistics, parametric as well as non-parametric, have focused on functional data analysis. The solutions developed can be used to achieve a wide range of goals (dimension reduction, regression, clustering, classification, etc.). They provide models that are predictive but often difficult to interpret. This lack of interpretability is partly due to the pre-processing step required on the functional data, which must be projected into bases of functions (splines, kernels...). The transformation of functional variables is a fundamental step, no matter if the variable is a response variable or a predictor.

The SpiceFP approach is primarily explanatory and not necessarily predictive. The first step is a transformation of the predictors into categorical variables. The choice of this transformation was motivated by the potential to interpret the results while considering the hypothesis of a joint influence of the predictors on the response variable. Each partition is a collection of 2D intervals, see §2.1. We constructed linear regressors on the basis of indicator functions associated to these 2D intervals, see Equation (2.4). This basis of indicator functions is not common compared to polynomial basis, Fourier basis, wavelet basis, etc., [39]. In a previous work [21], the authors have focused on multidimensional penalized signal regression, a single surface was estimated with smooth regression coefficient using B-spline tensor products. In our case, the indicator functions facilitate the interpretation of the results, provided that both functional predictors (T in grape

berry dataset) are discretized over *the same set of equidistant observation times*. This constraint can be released with usual pretreatment such as:

- imputation of missing data [32, 16]
- interpolation, smoothing [25] or restriction of the functional variable to an identical set of observation times. Indeed, functions have uncountable supports.

The use of 2D class intervals implies also the assumption that the structure of the underlying process does not change over the observation variable (time in the grape berry dataset). This is what we can call *hypothesis of stationarity*. For example, in our use case on grapevine, this hypothesis of stationarity requires to work at the scale of a week but also to split the day (working with observations obtained in the morning (sunrise to noon)). When analyzing the data in section 5, we assumed that the underlying process is time-invariant in the mornings of the week under consideration.

The method is design-dependent: it 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 there is potentially something happening. There is also a limitation due to the curse of dimensionality [10]: "If the number n of observations remains fixed while the dimension p of the observations increases, the observations get rapidly very isolated and local methods cannot work.". The more fluctuations in many directions, the more data will be needed. Our approach will work best if the shapes are simple.

6.2. SpiceFP: an iterative approach

The SpiceFP approach can be summarized as follows: first, the functional predictors are transformed into a set of candidate matrices X^u (and $D^{u,\gamma}$) for each (u, γ). Then we iterate the following steps: estimation of the coefficients for each of the candidate matrices, evaluation and selection based on the chosen validation information criteria with respect to (u, γ) and then checking of the shutdown conditions. The number of iterations performed is controlled by the shutdown conditions. The residuals obtained at iteration k are used as response variable at iteration $k + 1$. This iterative procedure has been implemented in order to:

- *extract knowledge still available in the residuals with the risk of overestimating the number of non-zero coefficients*: This risk is linked to several factors such as:
  1. *the noise level in the response variable*: the higher the noise level in the response variable, the higher the risk of model misspecification in the SpiceFP approach [13]. This situation is well illustrated by the results of the simulations at iteration 2 (row 1 vs row 2 and row 3 vs row 4, column 4, Table 2).
  2. *the shape of the area of influence to be identified*: the number of connected components to cover the shape increases when the shape is not a combination of rectangles. For a shape composed of rectangles of different sizes, the iterative approach allows changing the scale of the grid to get a refined estimation.

  This tolerance to the overestimation of the coefficients is made in return for the identification of easily interpretable areas of influence. This contributes to the explanatory character of the approach. The number of non-zero coefficients is also related to the value of the γ parameter presented in the Equation (2.7) which controls the ratio between parsimony and fusion of the coefficients. The lower the parameter γ (conditionally to the data), the larger the estimated area of influence. This increases the risk of getting unnecessary /untrue non-zero coefficients.

- *Explore the space of observations by using different partitions/pavements (fine scale, large scale):* from the second iteration onwards, the candidate matrix selected covers the same space as the previous one, with the difference that it can cover it with different widths. Thus new areas of influence can be identified, or existing ones reinforced, etc. There is therefore no independence between the different candidate matrices because they are all constructed from the observed predictors. Even if the response of the next iteration corresponds to the residuals of the previous iteration, this approach is not orthogonal.

Eventually, the shutdown condition of SpiceFP is rather natural. The approach stops when the vector of coefficients at one iteration is null: in this case, the residuals of the model are equivalent to the response variable. This criterion is internal to the approach.

7. Acknowledgments

The authors dedicate this work to their colleague Eric Lebon, who passed away. 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. We are also grateful to Nicolas Verzelen, researcher at INRAE, for his help and constructive discussions.

References

[1] 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(4):1053–1061.

[2] Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY.

[3] Arnold, T. B. and Tibshirani, R. J. (2019). *genlasso: Path algorithm for generalized lasso problems*. Package version 1.4 for R version 3.6.1.

[4] 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):1–7.

[5] Crambes, C., Gannoun, A., and Henchiri, Y. (2013). Support vector machine quantile regression approach for functional data: Simulation and application studies. *Journal of Multivariate Analysis*, 121:50 – 68.

[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499.

[7] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[8] Ferraty, F. and Vieu, P. (2011). Richesse et complexité des données fonctionnelles. *Revue MODULAD*, 43:25–43.

[9] Garcia, M. G., Medeiros, M. C., and Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, 33(3):679 – 693.

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

[11] 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(1):46–64. PMID: 24729670.

[12] Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2019). Bayesian Functional Linear Regression with Sparse Step Functions. *Bayesian Analysis*, 14(1):111–135.

[13] Gustafson, P. (2002). On the simultaneous effects of model misspecification and errors in variables. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(3):463–474.

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

[15] James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Ann. Statist.*, 37(5A):2083–2108.

[16] Josse, J. and Husson, F. (2016). missmda: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, Articles*, 70(1):1–31.

[17] Kang, J., Reich, B. J., and Staicu, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184.

[18] 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(2):687–713.

[19] 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(3):321–333.

[20] Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 353–360, New York, NY, USA. Omnipress.

[21] Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

[22] Möller, A., Tutz, G., and Gertheiss, J. (2016). Random forests for functional covariates. *Journal of Chemometrics*, 30(12):715–725.

[23] O'Connor, T. G. (1995). Acacia karroo invasion of grassland: Environmental and biotic effects influencing seedling emergence and establishment. *Oecologia*, 103(2):214–223.

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

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

[26] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

[27] Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249.

[28] Reiss, P. T., Huang, L., and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The international journal of biostatistics*, 6 1:Article 28.

[29] Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, 64:183 – 210. Trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks 2004.

[30] 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(3):887–895.

[31] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

[32] Stekhoven, D. J. and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

[33] 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(3):235–247.

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[35] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

[36] Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

[37] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232.

[38] Varlet-Grancher, C., Gosse, G., Chartier, M., Sinoquet, H., Bonhomme, R., and Allirand, J. (1989). Mise au point : rayonnement solaire absorbé ou intercepté par un couvert végétal. *Agronomie*, 9(5):419–439.

[39] Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

[40] Wang, X., Zhu, H., and for the Alzheimer's Disease Neuroimaging Initiative (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112:519:1156–1168.

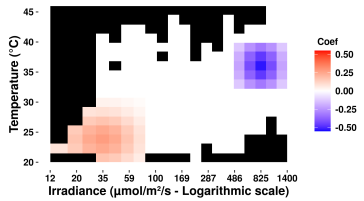
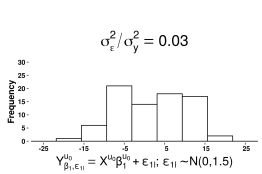
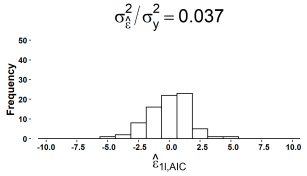
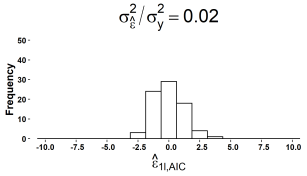
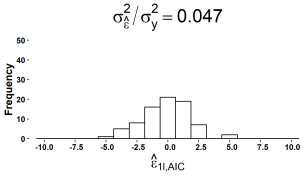
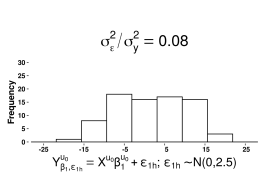
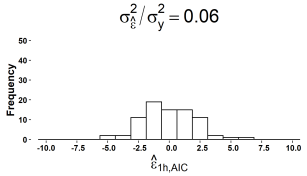
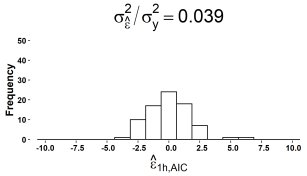
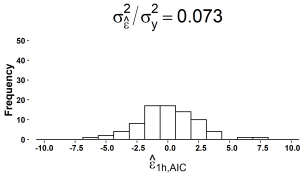
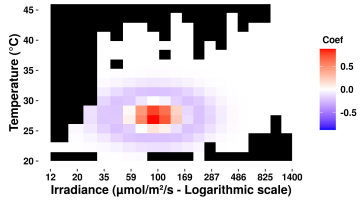
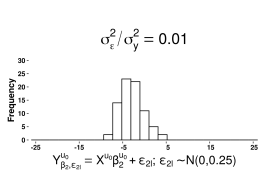
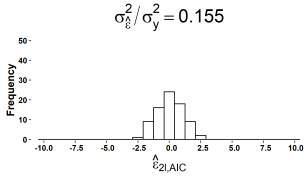
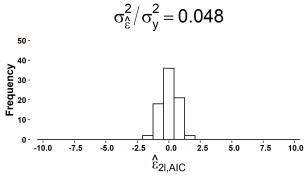
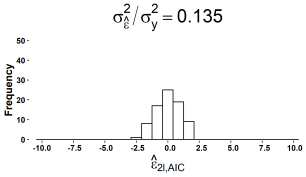
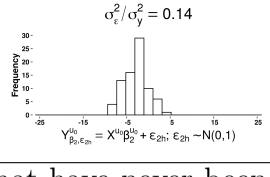
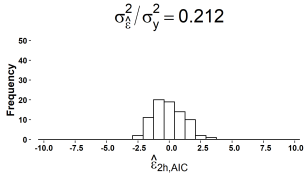
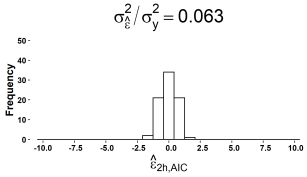
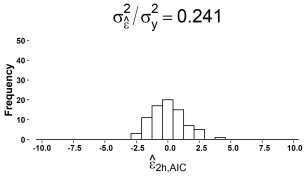
[41] Xin, B., Kawahara, Y., Wang, Y., and Gao, W. (2014). Efficient generalized fused lasso and its application to the diagnosis of alzheimer's disease. *Proceedings of the National Conference on Artificial Intelligence*, 3:2163–2169.

[42] Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76 2:463–483.

Appendix

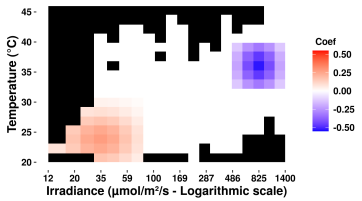
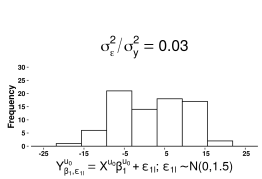
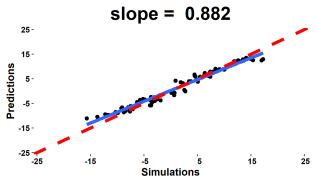
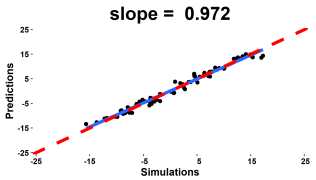
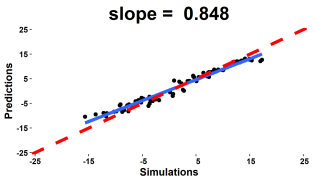
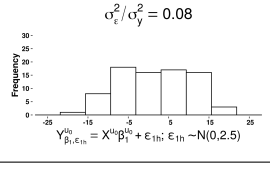
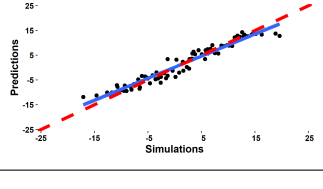
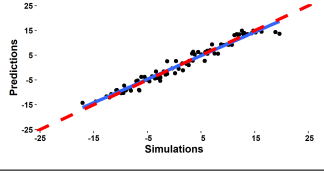
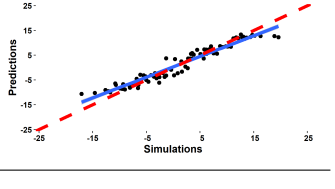
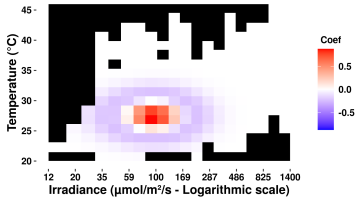
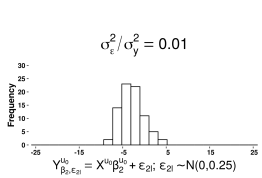
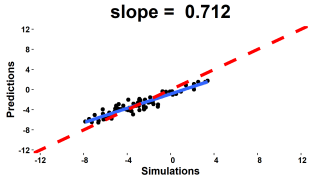
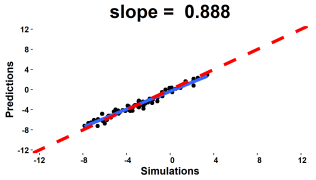
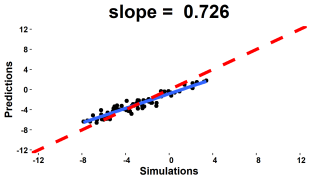
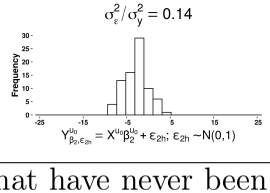
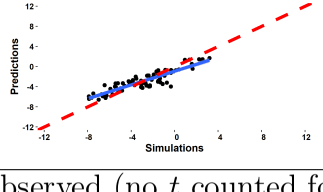
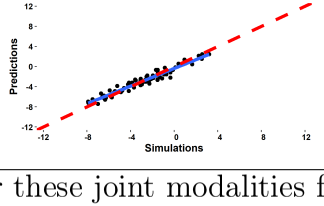
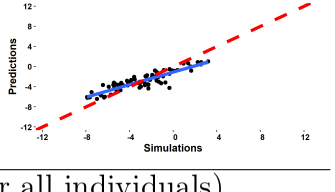
Simulation's residual histogram and quality of the estimate

Table 4: Histogram of residuals

Simulations	Estimations			
	Responses	iter 1	iter 2	Mean of top 1% iter 1
				
				
				
				

■: Joint modalities that have never been observed (no t counted for these joint modalities for all individuals)

Table 5: Quality of the estimate

Simulations	Responses	Estimations iter 1	Estimations iter 2	Estimations Mean of top 1% iter 1
				
				
				
				

■: Joint modalities that have never been observed (no t counted for these joint modalities for all individuals)