

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

Girault GNANGUENON GUESSE, Patrice LOISEL, Bénédicte FONTEZ, Thierry SIMONNEAU, Nadine HILGERT

MISTEA and LEPSE (Montpellier University)

30th International Biometric Conference



l'institut Agro
agriculture • alimentation • environnement



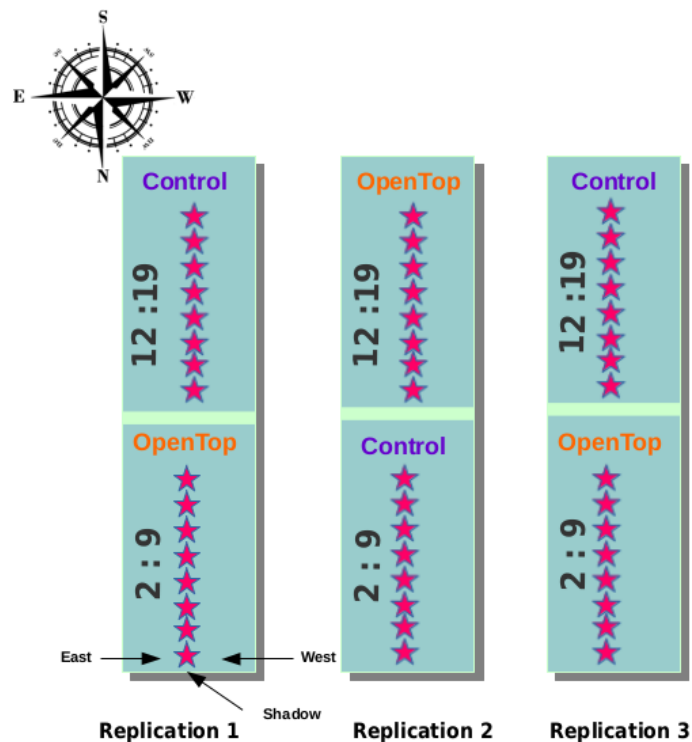
Outline

- 1 Introduction
- 2 Data and questions
- 3 The SPICEFP) approach
- 4 Simulation study
- 5 Use case : vine data set
- 6 Conclusion
- 7 Bibliography

- **Digital agriculture:** agricultural revolution made by the use of sensors to assist in decision
- **Data collected from sensors:** massive, uncertain, correlated, probably rich in information
- **Agronomic research:** need to understand and predict with more accuracy complex systems
- **Statistical research:** need to re-examine existing tools in order to take into account some specific data
- **Case of the INNOVINE project (LEPSE & SPO):** link the micro-environmental and compositional characteristics of the berry

Design of experiments

- Experiment on the vineyard of the INRAE/Institut Agro campus at Montpellier in 2014 (Syrah vines)
- 3 Replications (1/2/3) \times 3 Orientations (E/O/W) \times 16 Stumps (2 - 9 / 12 - 19) = 144 statistical individuals



Data

- **Variable response:**

Polyphenols contents
measured weekly via the
Ferari Index at a date d for
each individual i noted
 $FI_i(d)$

- **Explanatory variables**
(every 12 minutes on grape
berries):

- **Temperature**

$\mathcal{A} = \{\mathcal{A}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$

- **Irradiance** $\mathcal{B} = \{\mathcal{B}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$

Data

- **Variable response:**
Polyphenols contents measured weekly via the Ferari Index at a date d for each individual i noted $FI_i(d)$
- **Explanatory variables** (every 12 minutes on grape berries):
 - **Temperature**
 $\mathcal{A} = \{\mathcal{A}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$
 - **Irradiance** $\mathcal{B} = \{\mathcal{B}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$

Questions

- **Agronomy:** how to extract knowledge (from data) to explain the influence of climate on the development of a berry component quality ?
- **Statistics:** how to identify an operator \mathcal{F} such that:

$$\begin{aligned}\Delta FI_i(d_1, d_2) &= FI_i(d_2) - FI_i(d_1) \\ &= \mathcal{F}(\mathcal{A}_i(t), \mathcal{B}_i(t); t \in T)\end{aligned}$$

Data

- **Variable response:**
Polyphenols contents measured weekly via the Ferari Index at a date d for each individual i noted $FI_i(d)$
- **Explanatory variables** (every 12 minutes on grape berries):
 - **Temperature**
 $\mathcal{A} = \{\mathcal{A}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$
 - **Irradiance** $\mathcal{B} = \{\mathcal{B}_i(t) : t \in T; i = 1, \dots, n; \underline{T} = d_1; \bar{T} = d_2\}$

Questions

- **Agronomy:** how to extract knowledge (from data) to explain the influence of climate on the development of a berry component quality ?
- **Statistics:** how to identify an operator \mathcal{F} such that:

$$\begin{aligned}\Delta FI_i(d_1, d_2) &= FI_i(d_2) - FI_i(d_1) \\ &= \mathcal{F}(\mathcal{A}_i(t), \mathcal{B}_i(t); t \in T)\end{aligned}$$

3 assumptions : \mathcal{A} and \mathcal{B} **jointly influence** the response variable;
no missing data and same equidistant observation times for both functional variables

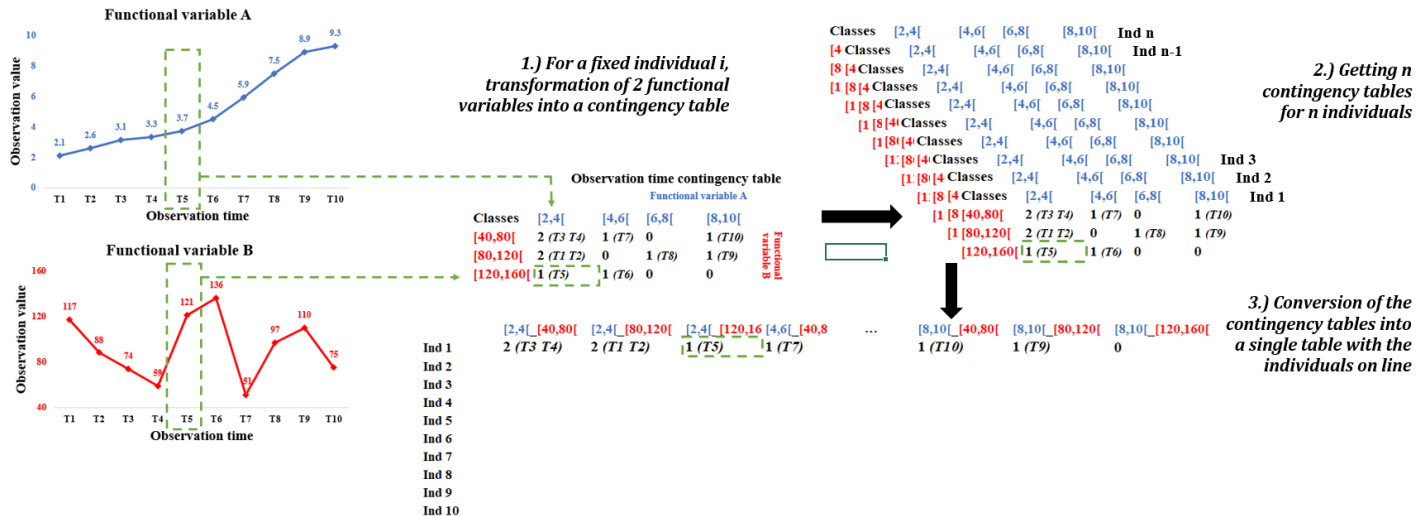
SPICEFP - From functions to contingency tables based on a partition vector $u = (n_A, n_B)$ (step 1)

Three steps are involved to implement SPICEFP :

- Transformation of both functional variables into a candidate explanatory matrix

Transformation of functional explanatory variables for SPICEFP modeling

Illustration for 2 functional variables and $n=10$ individuals



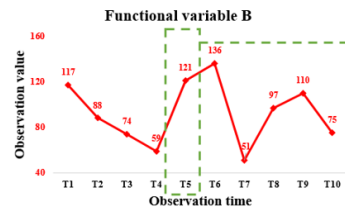
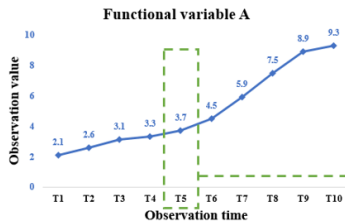
SPICEFP - From contingency tables to graphs (step 2)

using a contiguity matrix E^u (edges matrix)

- Creation of a graph of contiguity constraints: definition of an edge matrix to represent the contiguity constraints between modalities of the contingency table.

Transformation of functional explanatory variables for SPICEFP modeling

Illustration for 2 functional variables and $n=10$ individuals



1.) For a fixed individual i , transformation of 2 functional variables into a contingency table

	Functional variable A			
Classes	[2,4]	[4,6]	[6,8]	[8,10]
[40,80]	2 (T3 T4)	1 (T7)	0	1 (T10)
[80,120]	2 (T1 T2)	0	1 (T8)	1 (T9)
[120,160]	1 (T5)	1 (T6)	0	0

Functional variable B

	[2,4]	[4,6]	[6,8]	[8,10]	...	[8,10]	[8,10]	[8,10]
[40,80]	2 (T3 T4)	2 (T1 T2)	1 (T5)	1 (T7)	...	1 (T10)	1 (T9)	0

- Ind 1
- Ind 2
- Ind 3
- Ind 4
- Ind 5
- Ind 6
- Ind 7
- Ind 8
- Ind 9
- Ind 10

	[4,6]	[40,80]		
[2,4]	[80,120]	[4,6]	[80,120]	[6,8]
	[4,6]	[120,160]		

Classes	[2,4]	[4,6]	[6,8]	[8,10]	Ind n
[4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	Ind n-1
[1 [8 [4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	
[1 [8 [4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	
[1 [8 [4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	
[1 [8 [4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	Ind 3
[1 [8 [4 Classes	[2,4]	[4,6]	[6,8]	[8,10]	Ind 2
[1 [8 [40,80]	2 (T3 T4)	1 (T7)	0	1 (T10)	Ind 1
[1 [80,120]	2 (T1 T2)	0	1 (T8)	1 (T9)	
[1 [120,160]	1 (T5)	1 (T6)	0	0	

2.) Getting n contingency tables for n individuals

3.) Conversion of the contingency tables into a single table with the individuals on line

4.) Associate to this unique table, a neighbourhood matrix giving information on the closeness of the classes according to the 2 functional variables

SPICEFP : Identification of the best class intervals and related regression coefficients (step 3)

- 1 for $u = (n_{\mathcal{A}}, n_{\mathcal{B}})$, perform the Generalized Fused Lasso [Tibshirani and Taylor, 2011]:

$$\beta^u = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^n (y_i - X_i^u \beta)^2 + \lambda_p \sum_j |\beta_j| + \lambda_f \sum_{(j,j') \in E^u} |\beta_j - \beta_{j'}| \quad (1)$$

with: $\lambda_p \geq 0$; $\lambda_f \geq 0$ the regularization parameters and the couples (j, j') contained in E^u the edges matrix

SPICEFP : Identification of the best class intervals and related regression coefficients (step 3)

- 1 for $u = (n_{\mathcal{A}}, n_{\mathcal{B}})$, perform the Generalized Fused Lasso [Tibshirani and Taylor, 2011]:

$$\beta^u = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^n (y_i - X_i^u \beta)^2 + \lambda_p \sum_j |\beta_j| + \lambda_f \sum_{(j,j') \in E^u} |\beta_j - \beta_{j'}| \quad (1)$$

with: $\lambda_p \geq 0$; $\lambda_f \geq 0$ the regularization parameters and the couples (j, j') contained in E^u the edges matrix

- 2 choose the best candidate matrix (adapted information criterion) :

$$df(X\hat{\beta}^u) = \mathbf{Number\ of\ fused\ groups}$$

[Tibshirani and Taylor, 2012]

- 3 compute the residuals of the best model: $\varepsilon = y - X\hat{\beta}^{u*}$
- 4 Check the shutdown conditions and if they are not verified, go back to step 3 and replace y by ε

Methodology

- ① Simulations based on observed functional data
- ② Set a partition vector u^0 and Construct X^{u^0}
- ③ Simulate the coefficients β_1 and β_2
- ④ Compute $y_1 = X^{u^0} \beta_1 + \varepsilon_1$ et $y_2 = X^{u^0} \beta_2 + \varepsilon_2$
- ⑤ Estimate β_1 and β_2 using SPICEFP

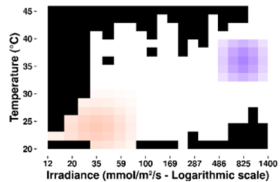
Simulation study (methodology and results)

Methodology

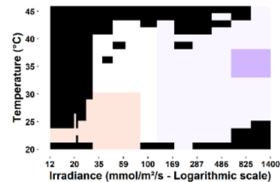
- 1 Simulations based on observed functional data
- 2 Set a partition vector u^0 and Construct X^{u^0}
- 3 Simulate the coefficients β_1 and β_2
- 4 Compute $y_1 = X^{u^0} \beta_1 + \varepsilon_1$ et $y_2 = X^{u^0} \beta_2 + \varepsilon_2$
- 5 Estimate β_1 and β_2 using SPICEFP

Results

Coefficients



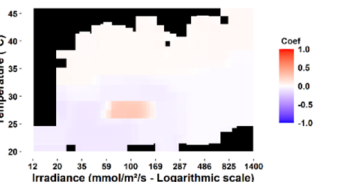
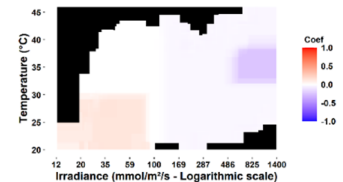
Iteration 1



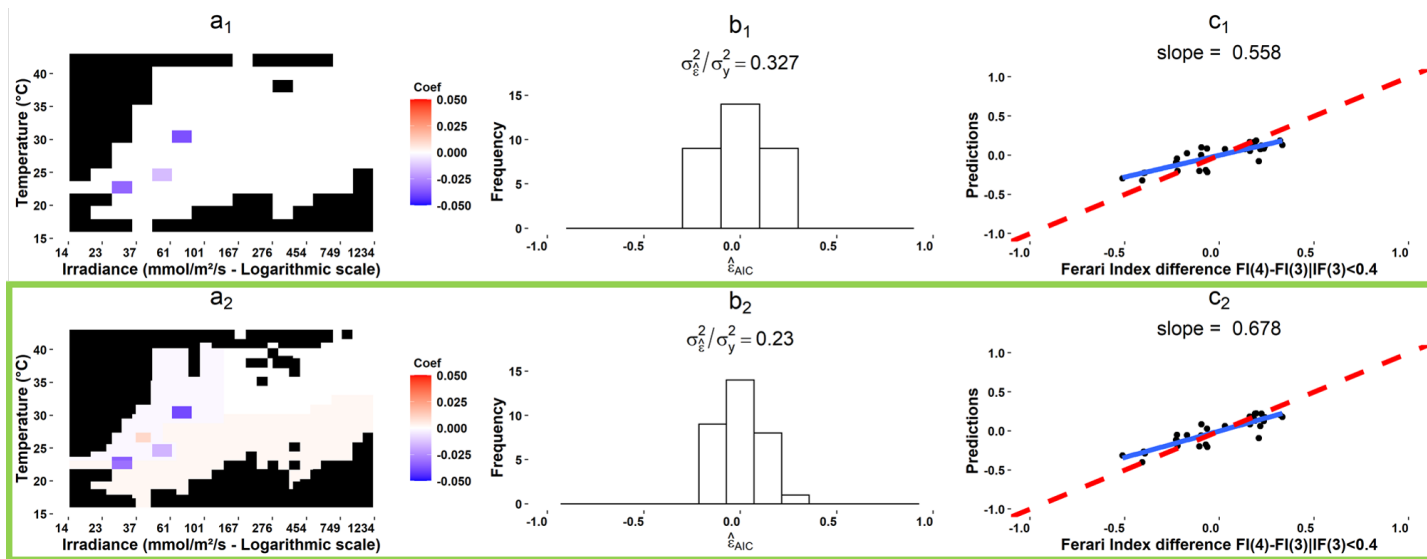
Iteration 2



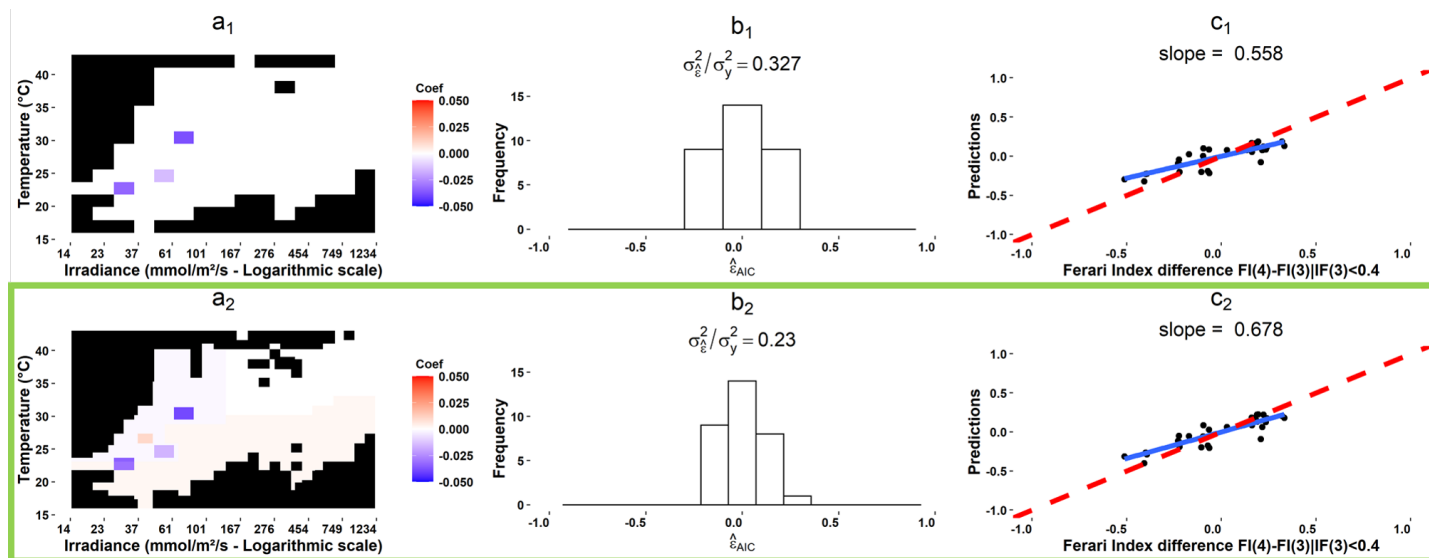
Top 1% Iteration 1



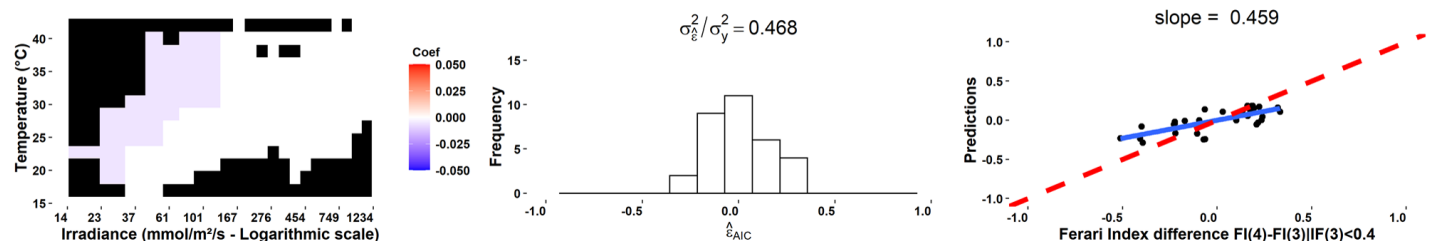
Vine data analysis (3rd week of ripening; sunrise to 12 a.m.)



Vine data analysis (3rd week of ripening; sunrise to 12 a.m.)







When need more fusion (therefore less parsimony):



Conclusion

- To use SPICEFP, it is assumed no missing data and same equidistant observation times for both functional variables.
- Usual pre-treatments (imputation, interpolation) can be used to validate these hypotheses
- SPICEFP approach allows to identify the best joint class intervals of two functional variables and assign regression coefficients to them (scalar-on functions approach)
- SPICEFP transform the scalar-on functions problem into scalar-on image [Goldsmith et al., 2014, Wang et al., 2017]
- Slight modifications will allow to go to a third dimension

Bibliography

-  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.
-  Tibshirani, R. J. and Taylor, J. (2011).
The solution path of the generalized lasso.
The Annals of Statistics, 39(3):1335–1371.
-  Tibshirani, R. J. and Taylor, J. (2012).
Degrees of freedom in lasso problems.
Ann. Statist., 40(2):1198–1232.
-  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.