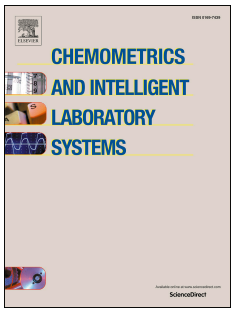
Accepted Manuscript



Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach

Celestino Ordóñez, Manuel Oviedo de la Fuente, Javier Roca-Pardiñas, José Ramón Rodríguez-Pérez

PII: S0169-7439(17)30476-8

DOI: 10.1016/j.chemolab.2017.12.001

Reference: CHEMOM 3553

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 26 July 2017

Revised Date: 23 November 2017

Accepted Date: 1 December 2017

Please cite this article as: C. Ordóñez, M.O. de la Fuente, J. Roca-Pardiñas, José.Ramó. Rodríguez-Pérez, Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: 10.1016/j.chemolab.2017.12.001.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Determining optimum wavelengths for leaf water content estimation from reflectance: a distance correlation approach

Celestino Ordóñez[a], Manuel Oviedo de la Fuente[b,c], Javier Roca-Pardiñas[b,d], José Ramón Rodríguez-Pérez[e]

[a] *Department of Mining Exploitation and Prospecting. University of Oviedo. Escuela Politécnica de Mieres, 33600 Mieres - Spain*
[b] *Technological Institute for Industrial Mathematics (ITMATI)*
[c] *Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela*
[d] *Department of Statistics. University of Vigo*
[e] *Grupo de Investigación en Geomática e Ingeniería Cartográfica (GEOINCA), Escuela Superior y Técnica de Ingeniería Agraria, Universidad de León, Avenida de Astorga, s/n, 24401, Ponferrada (León) Espaa*

Abstract

This paper proposes a method to estimate leaf water content from reflectance in four commercial vineyard varieties by estimating the local maxima of a distance correlation function. First, it applies four different functional regression models to the data and compares the models to test the viability of estimating water content from reflectance. It then applies our methodology to select a small number of wavelengths (optimum wavelengths) from the continuous spectrum, which simplifies the regression problem. Finally, it compares the results to those obtained by means of two different methods: a nonparametric kernel smoothing for variable selection in functional data and a wavelet-based weighted LASSO functional linear regression. Our approach proved to have some advantages over these two testing approaches, mainly mainly in terms of the computing time and the lack of assumption of an underlying model. Finally the paper concludes that estimating water content from a few wavelengths is almost equivalent to doing so using larger wavelength intervals.

Keywords: vineyard, water content, distance correlation, functional analysis, design points

ACCEPTED MANUSCRIPT

1. Introduction

Water availability plays an important role in the production and quality of agricultural plants, especially in multi-annual crops such as vines (Vitis vinifera L.) [1]. One way to estimate vine water content is to measure leaf water content [2]. Another is to use a pressure chamber to measure leaf water potential. [3], but this method is tedious, time consuming and even destructive [4, 5]. Plant water content can alternatively be assessed by remote sensing technologies [6, 7]. Leaf reflectance, i.e., the ratio of incoming radiance reflected from the leaves, may be used to estimate water content in addition to other chemical properties such as chorophyll, carbon or nitrogen content. Absorption of radiation by water in the leaf tends to decrease reflectance. The NIR region of the electromagnetic spectrum [730 - 2300] nm contains several wavelengths strongly influenced by the presence of water, and the state of water in the measured sample [8]. Several methods have been proposed to estimate water content from leaf reflectance: vegetation indices [9, 10, 11], multiple regression models [12, 13, 14] or inversion models [15, 16].

When the reflectance is measured with devices of high radiometric resolution, the data can be considered as curves. This leads some authors to propose the use of functional data regression techniques [17, 18]. However, some people still find functional data analysis too complex and difficult to interpret. They prefer methods that are less mathematically complex and easier to interpret, such as vegetation indices or linear regression models with a small number of covariates, even though the predictive results they provide are worse than those provided by more complex regression models. It is therefore important to develop new methods to drastically reduce the dimension of the problem and thereby facilitate the application of simple and readily interpretable models, which relate response and predictor variables when only a few optimum wavelengths must be considered.

Methods based on linear finite dimensional projections such as Functional Principal Component Regression (FPCR) or Functional Partial Least Squares (FPLS) [19] have been proposed to reduce dimensionality. However, one drawback of these kind of methods is that the output is not directly interpretable in terms of the original variables. Hence the great interest in variable selection methods, especially in those where the output only depends

ACCEPTED MANUSCRIPT

on the data, not on any underlying modeling [20]. A number of variable selection methods have been proposed, among them the Elastic Net [21] or Boosting approaches [22]. The problem of variable selection when the predictor variables are categorical has been addressed in [23]. In this particular case, the effect of one variable can be determined not by one, but by several coefficients. Authors in [24] tackled the problem of consistency in regression models with high dimensionality and proposed a limit in the dimension of the problem compared to the sample size for consistent variable selection. A different solution was proposed in [25], using a wavelet-based LASSO procedure [26]. The regression is performed in the wavelet domain and then, after discarding small coefficients, the inverse wavelet transformation is applied to return to the original domain. More recently, this approach was improved by means of screening and penalty factor weighting schemes [27].

In this work we study the utility of distance correlation [28] as an intrinsic method for variable selection. Neither projection nor transformation of the variables is needed. Moreover, it is unnecessary to assume an a priori regression model; we just look for local maxima of the distance correlation function. The rest of the article is structured as follows: First, we provide a brief summary of the functional parametric [29, 25] and nonparametric regression models [30] used to estimate leaf water content from reflectance. Second, we provide a brief explanation of the three methods used to determine optimum wavelengths: one is based on a nonparametric kernel smoothing [31], another is a wavelet-based weighted LASSO regression [27], and our proposal, which is based on calculating local maxima on a distance correlation function. Third, we apply all the methods explained in the previous section to simulated and real data. Then we analyze the results obtained and extract a set of conclusions that summarize the whole work extracted.

2. Methodology

The following lines summarize the four different functional approaches employed in this work to estimate leaf water content from reflectance. A brief explanation of each model is given below, so we recommend consulting the cited literature for each of the methods. Then, we explain the method proposed to simplify the problem, reducing its dimension to a few dimensions corresponding to a small number of optimum wavelengths. This method is compared with another two approaches for variable selection in functional data regression.

ACCEPTED MANUSCRIPT

2.1. Functional regression models

Consider a sample data $\{X_i, Y_i\}_{i=1}^{n}$ where $X_i = (X_i(t_1), X_i(t_2), \ldots, X_i(t_N))$ and $Y_i \in \mathbb{R}$, n being the sample size and N the number of discrete observation points where the independent variable X_i is observed. In our study, X_i represents the reflectance at wavelengths $(t_1, t_2, \ldots, t_N)$ and Y_i the water content of each vine leaf. We can assume that both variables are related by the model

$$Y_i = r\left(X_i\right) + \varepsilon_i \tag{1}$$

where $r(\cdot)$ is the regression function and ε_i is an error term with zero mean that represents other sources of variability not accounted for in X_i.

When we have a fine grid of data $X_i\left(t\right)$, such when a spectrometer is used to register leaf reflectance, we may formulate the regression problem within the context of functional data analysis [18]. In this case $X_i = X_i(t)$ can be considered a function of $t \in [a, b]$. In functional data analysis we assume the underlying processes generating the data smooth and may therefore be approximated by functions. Techniques commonly used in multivariate statistics, such as principal component analysis, regression, clustering, classification or ANOVA, are also adapted to work with functions instead of vectors. One of the advantages of FDA over classical multivariate statistics is that it allows us to extract additional information contained in the functions and their derivatives [32].

We applied the four functional regression models described in the next section to estimate water content from reflectance.

2.1.1. Functional linear regression (FLR)

Let be $X_i \in \mathcal{L}_2(T)\ \forall t \in [a, b]$, and $Y_i \in \mathcal{R}$, a parametric functional linear model, as formulated in [29], can be written following the model in (1) as follows:

$$Y_i = \alpha + \int_T X_i(t)\beta(t)dt + \varepsilon_i \tag{2}$$

where $\alpha \in \mathbb{R}$ and $\beta\left(t\right) \in \mathcal{L}_2(T)$ are the regression coefficients. In this model $X_i(t)$ and $\beta(t)$ are approximated by means of decomposition in K basis functions

ACCEPTED MANUSCRIPT

$$X_i(t) \approx \sum_{k=1}^{K} a_{ik}\phi_k = \mathbf{a}_i^\top \Phi \quad \text{and} \quad \beta(t) \approx \sum_{k=1}^{K} b_k\theta_k(t) = \mathbf{b}^\top \Theta$$

so

$$\int_T X_i(t)\beta(t)dt \approx \mathbf{a_i}^\top \Phi\Theta^\top \mathbf{b}$$

where $\mathbf{a_i}$ and $\mathbf{b}$ are Kx1 vector of coefficients, and Φ and Θ are the basis functions. The choice of the appropriate basis functions (and the number of basis elements) becomes a crucial step [33]. They are usually polynomial, exponential, B-splines, Fourier functions or wavelets.

The unknowns α and $\mathbf{b}$ are obtained by minimizing the penalized residual sum of squares

$$n^{-1}\sum_{i=1}^{n}\left[Y_i - \alpha - \int_T X_i(t)\beta(t)dt\right]^2 + \lambda\int_T [D^p\beta(t)]^2 dt \tag{3}$$

The second term is a regularization term that penalizes high local variations of the regression coefficients. λ is a positive constant that controls the trade-off between roughness and fidelity to the data, and $D^p(\beta)$ is the derivative of order p. The second derivative is normally used, given that it measures the size of the curvature.

2.1.2. Functional wavelet-based LASSO regression (FWLASSO)

LASSO (Least Absolute Shrinkage and Selection Operator) is a well known technique for shrinkage and variable selection in multiple regression. It basically consists in penalizing the magnitude of the regression coefficients in order to reduce the influence of the small ones as compared with the large ones. Its extension to functional regression leads to an expression similar to Eq. (3), changing the regularization term as follows:

$$\widehat{\beta}(t) = \underset{\beta(t)\in\mathcal{L}_2(T)}{\arg\min}\left(\sum_{i=1}^{n}\left[Y_i - \int_T X_i(t)\beta(t)dt\right]^2 + \lambda\int_T |\beta(t)|\,dt\right) \tag{4}$$

When the penalty parameter λ increases, the range of t values with $\beta(t) = 0$ also increases.

ACCEPTED MANUSCRIPT

As with FLR, the predictors $X_i(t)$ and regression coefficients $\beta(t)$ are approximated using basis functions, such as B-splines [34] or wavelets. In this work, we used wavelet-based LASSO in functional regression following [25] and [27]. The problem is solved in the wavelet domain and then, after selecting the non-null coefficients, these coefficients are mapped back to the original domain. Among other advantages, a wavelet-based LASSO regression performs well when the coefficient function is spiky.
For a primary decomposition level j_0, the wavelet decomposition of the predictors can be represented as

$$X_i(t) = \sum_{k=0}^{2^{j_0}-1} z'_{i,j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} z_{i,j,k}\psi_{j,k}(t)$$

where the wavelet coefficients are defined by

$$z'_{i,j_0,k} = \int_T X_i(t)\phi_{j_0,k}(t)dt, z_{i,j,k} = \int_T X_i(t)\psi_{j,k}(t)dt$$

being

$$\phi_{j,k} = 2^{j/2}\phi(2^j t - k), \psi j, k = 2^{j/2}\psi(2^j t - k)$$

the orthonormal scaling (father wavelet) and mother wavelet basis functions, respectively.

A similar decomposition is performed for $\beta(t)$ (see [25] for a more detailed description of wavelet-based decomposition LASSO in functional regression, and [27] for a variant of this method that includes different prescreening and weighting schemes for the penalty term).

2.1.3. Functional principal components regression (FPCR)

We tested a functional regression model that uses functional principal components (FRPCA) for a basis expansion. We selected the principal components by means of AICc (bias-corrected Akaike Information criterion) [35]. PCA for functional data can be formulated as an eigenanalysis of the empirical variance-covariance function [29]

$$\upsilon\,(s,t) = (n-1)^{-1}\sum_{i=1}^{n-1}\left(X_i(s) - \bar{X}(s)\right)\left(X_i(t) - \bar{X}(t)\right)$$

where $\bar{X}(t) = n^{-1}\sum_{i=1}^{n} X_i(t)$.

ACCEPTED MANUSCRIPT

Given s, the following eigen decomposition problem is formulated

$$\int_T \upsilon(s,t)\,\xi(t)\,dt = \rho\xi(s) \tag{5}$$

where ρ is the eigenvalue and ξ the weight function (loadings in the PCA). $X(t)$ and $\xi(s)$ can be expanded on basis functions as in the previous section

$$X(t) = \mathbf{C}\phi \quad \text{and} \quad \xi(s) = \phi^{\top}\mathbf{d} \tag{6}$$

where the coefficient matrix $\mathbf{C}$ is nxK and $\mathbf{d}$ is a Kx1 vector of coefficients.

Replacing (6) in (5), we obtain the eigenequation in matrix form

$$(n-1)^{-1}\mathbf{C}^{\top}\mathbf{C}\mathbf{d} = \rho\mathbf{d} \tag{7}$$

assuming the basis functions are orthonormal. The resulting functional principal components are used as basis functions to approximate the original functions $X_i(t)$ and solve the functional linear regression problem as in Section 2.1.1.

2.1.4. Nonparametric functional regression (NPFR)

We used a nonparametric functional approach widely described in [30]. In this case, we use a kernel regression estimator to estimate the regression function r in model (1). Specifically, for the Nadaraya-Watson estimator [31], we obtain:

$$\widehat{r}(X) = \frac{\sum_{i=1}^{n} K\left(h^{-1}d\left(X, X_i(t)\right)\right) Y_i}{\sum_{i=1}^{n} K\left(h^{-1}d\left(X, X_i(t)\right)\right)} \tag{8}$$

$K(\cdot)$ is a kernel function, h is the smoothing parameter (bandwidth) and $d(\cdot,\cdot)$ is a semi-metric that measures the proximity between functional objects.

Many kernels are possible, but the usual choices are Epanechnikov, Gaussian, Quartic and Tricube, although the election of the kernel is an unimportant factor.

Regarding the metric for the distance between the covariates, we use the $\mathcal{L}_2$ semi-metric to solve our specific problem

$$d(X_i, X_j) = \int_a^b \left(X_i(t) - X_j(t)\right)^2 dt$$

ACCEPTED MANUSCRIPT

However, other metrics or semi-metrics such as $\mathcal{L}_1$ or $\mathcal{L}_{inf}$ can be used [30].

The bandwidth h is usually determined by cross-validation, although other methods may be used [36].

2.2. Selection of the optimum wavelengths

The hypothesis underlying this procedure is that a few reflectance values corresponding to particular wavelengths contain most of the information concerning the response variable. If we are able to determine these wavelengths, we will have simplified the solution of the problem significantly. Following are three different approaches to determine optimum wavelengths are shown.

2.2.1. Nonparametric variable selection approach (NOVAS)

One of the main interest points in this research was to determine whether solving the regression required considering the whole spectrum or specific bands or wavelengths provided an equivalent response. Reducing the dimension of the problem is advisable for two main reasons: (i) it produces results that are easier to interpret, (ii) it can reduce the computation time.

First, we used the method proposed by [37] to estimate the most predictive design points (NOVAS), the design points were those of the model corresponding to the wavelengths $t_1, \ldots, t_N$ where the reflectance was measured. This is an iterative forward/backward method that consists in estimating a subset $t_1, \ldots, t_r$ ($r << N$), of the measured wavelength spectrum with the greatest predictive influence. This subset is determined by cross-validation minimizing the CV score:

$$CV\left(\mathrm{t},\mathrm{h}\right) = \frac{1}{n}\sum_{i=1}^{i=n}\left\{Y_i - \widehat{g}_{h,-i}\left(X_i\right)\right\}^2 \tag{9}$$

where $h = (h_1, \ldots, h_r)$ represents bandwidths used to estimate the leave-one-out local estimator $\widehat{g}_{\mathrm{h},-i}\left(\mathrm{x}\right)$ of $g\left(\mathrm{x}\right) = E\left(Y \mid X = \mathrm{x}\right)$, defined as:

$$\widehat{g}_{h,-i}(\mathrm{x}) = \overline{Y}_i(\mathrm{x},\mathrm{h}) + \widehat{\gamma}_{\mathrm{h},-i}^{\top}\left\{\mathrm{x} - \overline{X}_i(\mathrm{x},\mathrm{h})\right\} \tag{10}$$

The terms on the right, $\overline{Y}_i(\mathrm{x},\mathrm{h})$ and $\overline{X}_i(\mathrm{x},\mathrm{h})$ are calculated by means of kernel smoothing

$$\overline{Y}_i(\mathrm{x},\mathrm{h}) = \frac{\sum_{j:j\neq i} Y_j K_j(\mathrm{x} \mid \mathrm{h})}{\sum_{j:j\neq i} K_j(\mathrm{x} \mid \mathrm{h})}, \overline{X}_i(\mathrm{x},\mathrm{h}) = \frac{\sum_{j:j\neq i} X_j K_j(\mathrm{x} \mid \mathrm{h})}{\sum_{j:j\neq i} K_j(\mathrm{x} \mid \mathrm{h})}$$

ACCEPTED MANUSCRIPT

and the parameter $\gamma = \widehat{\gamma}_{\mathrm{h},-i}$ minimizes

$$\sum_{j:j\neq i} [Y_j - \overline{Y}_i(\mathrm{x},\mathrm{h}) - \gamma^\top \left\{X_j - \overline{X}_i(\mathrm{x},\mathrm{h})\right\}]^2 K_j(\mathrm{x} \mid \mathrm{h})$$

being

$$K_j(\mathrm{x},\mathrm{h}) = K\left(\sqrt{\sum_{k=1}^{r} \frac{(x_k - X_{jk})^2}{h_k^2}}\right)$$

In each iteration $s = 2, 3, ...,$ the forward addition algorithm looks for $(\mathrm{t}_s, \mathrm{h}_s)$ that maximize $[CV(\mathrm{t}_{s-1}, \mathrm{h}_{s-1}) - CV(\mathrm{t}_s, \mathrm{h}_s)]$, being $\mathrm{t}_s = (\widehat{t}_1, ..., \widehat{t}_{s-1}, \mathrm{t}_s)$ and $\mathrm{h}_s = (h_1, ..., h_s)$. The algorithm stops adding terms to t_s when $PCV_s < PCV_{s-1}$, being $PCV = CV \times (1 + \frac{\delta_0}{\log n})$ and δ_0 a constant.

Similarly, in each iteration of the backward deletion a component t of t_s that minimizes $[CV(\mathrm{t}_{s+1}[-t], \mathrm{h}_s) - CV(\mathrm{t}_{s+1}, \mathrm{h}_{s+1})]$ is removed. The algorithm stops when $PCV_s \leq PCV_{s+1}$.

2.2.2. Wavelet-based LASSO approach

The variable selection in the wavelet-based LASSO for functional linear regression model depends on the L_1-type penalty parameter λ. When $\lambda = 0$, the LASSO model is equivalent to an ordinary functional linear regression model without regularization, which usually produces overfitting and an irregular and wiggly regression coefficient function $\beta(t)$ difficult to interpret. When λ is very large, $\beta(t)$ tends to be null for a wide range of t values, so the final model is simple but inaccurate. Then, it is essential to find an adequate value of λ to reach a trade-off between power of prediction and parsimony. Cross-validation, is a common method for estimating λ.

Other parameters to consider are the type of wavelets and the primary decompositon level j_0, that controls the number of coefficients of the scaling and wavelet coefficients: 2^{j_0} and $N - 2^{j_0}$, respectively. N is the number of sample points t_j along $X_i(t_j)$. In this work we used Daubechies wavelets, while j_0 was determined by k-fold cross-validation.

Daubechies wavelet [38] is a family of asymmetric and orthonormal wavelets. Like the Haar wavelet (the fast ans simplest wavelet family), the Daubechies wavelet conserves the energy of signals and redistributes this energy in a more compact form. However, the scaling signals and wavelets of Daubechies

ACCEPTED MANUSCRIPT

wavelet have slightly longer supports, and they are more localized and smoother than Haar wavelet.

2.2.3. Local maxima distance correlation approach (LMDC)

The two previous variable selection methods have some drawbacks that we wish to avoid. For instance, NOVAS is quite expensive from a computational perspective, while wavelet-based LASSO requires transforming the original variables and assuming a linear model. To avoid this inconveniences, we propose a different method to determine the optimum wavelengths and solve the regression problem. We are interested in a simple, fast and model-free [39] method for variable selection with functional data. Our approach consists in calculating the local maxima of the distance correlation along the wavelength spectrum. Previously, the distance correlation curve is smoothed to avoid non-relevant local maxima.

Distance correlation $\mathcal{R}(X,Y)$ is an extension of the Pearson coefficient correlation [40] for non-linear dependences. Being $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ two random vectors, the distance correlation is defined as

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases} \tag{11}$$

where

$$\mathcal{V}^2(X,Y) = ||f_{X,Y} - f_X f_Y||^2 = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|^{1+p}|s|^{1+q}}$$

is the distance covariance, a measure of the distance between $f_{X,Y}$, the joint characteristic function of random vectors X and Y, and the product $f_X f_Y$ of the characteristics functions of X and Y,respectively. c_p and c_q are constants depending on the dimensions p and q, respectively.

One of the main advantages of distance correlation over the Pearson correlation is that it defines $\mathcal{R}(X,Y)$ in arbitrary finite dimensions of X and Y, and it characterises independence, i.e. $\mathcal{R}(X,Y) = 0 \Leftrightarrow X, Y$ are independent.

The correlation distance is a measure of the degree of correlation between two variables X, Y of arbitrary finite dimensions, so it is potentially a good indicator of the linear or nonlinear correlations between functional and multivariate variables. Accordingly, those variables with high values of the distance correlation may be useful for designing a functional linear or non

ACCEPTED MANUSCRIPT

linear (or additive) model. Recently, [41] provided conditions for the application of the distance correlation to functional spaces. Distance correlation was also applied to choose the most relevant variables in curve classification by "hunting the local maxima" of the covariance function [20]. A local maxima is a point t_j with the highest value in the interval (t_{j-h}, t_{j+h}). The choice of h depends on the nature of the data and the discretization pattern. Using local maxima as covariates reduces the redundancy, given that highly relevant points close to the local maxima are automatically excluded from the model. We tested several models in which the optimal classification rule depends on a small number of variables. These variables corresponded to local maximum of the distance covariance function.

For an observed random sample $(\mathbf{X}, \mathbf{Y}) = (X_k, Y_k), k = 1, ...n$, the empirical distance covariance is a non-negative number defined by

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl} \tag{12}$$

where $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$ and $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$ with $a_{kl} = \|X_k - X_l\|, \quad b_{kl} = \|Y_k - Y_l\|, \quad k, l = 1, \ldots, n$, and the subscript . denotes that the mean is computed for the index it replaces.

Once we calculate the local maxima of the distance correlation has been calculated (i.e., the optimum wavelengths), we check if we may estimate the water content from the reflectance at those wavelengths using a regression model. To this end, we apply a stepwise forward linear or nonlinear regression technique to the data. The reflectance at those wavelengths are the predictors and the leaf water content is the response variable. Then, we count the number of times a wavelength appears in each model and propose the most frequent as optimum wavelengths. The hypothesis underlying this procedure is that the reflectance values corresponding to those particular wavelengths contain most of the information concerning the response variable.

2.3. Algorithm

The algorithm to implement the LMDC approach for variable selection with functional data can be written as follows:

Step 1: Calculate de distance correlation $\mathcal{R}(t) = \{\mathcal{R}(X(t_j), Y)\}_{j=1}^{N}$, using the expression in (11) from the data $\{X_i(t_j), Y_i\}_{i=1}^{n}$.

Step 2: In order to avoid non relevant local maxima, smooth the distance correlation function. Particularly, in this work we fitted a nonparametric regression model

ACCEPTED MANUSCRIPT

$$\mathcal{R}(t) = m(t) + \varepsilon$$

where the function m was approximated using regression splines [42]

$$m(t) = \sum_{k=1}^{K} \gamma_k B_k(t)$$

being $\gamma_1, \ldots, \gamma_k$ unknown coefficients, and $B_1, \ldots, B_K$ are a set of K basis functions of order p (e.g. $p = 2$ for a cubic spline).

Finally the smoothed correlations are obtained as

$$\hat{\mathcal{R}}(t) = \sum_{k=1}^{K} \hat{\beta}_k B_k(t)$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the estimated coefficients obtained from the data.

Step 3: Calculate the local maxima of the smoothed correlation. Specifically, we used the STEM (Smoothing and TEsting of Maxima) algorithm proposed in [43]. Only the significant local maxima for a default level of significance are selected. Denoting the arguments values (argvals) of the local maxima a $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{\tilde{N}}$ ($\tilde{N} < N$), we ordered them from highest to lowest values of distance correlation, that is

$$\hat{\mathcal{R}}(\tilde{t}_1) \geq \hat{\mathcal{R}}(\tilde{t}_2) > \ldots \geq \hat{\mathcal{R}}(\tilde{t}_{\tilde{N}})$$

An adequate election of the number of basis K in Step2 is important in determining of the local maxima. If k is too small, no local maxima will be detected. On the other hand, if K is too big, too many local maxima will be detected.

Step 4: Check if the relationship between the reponse and the predictor variables is linear. That is, check the null hypothesis $H_0 : Y = \langle X, \beta \rangle + \epsilon$, versus a general alternative. To this aim, we apply a test of linearity that uses the Projected Cramer-von Mises statistic, [44]

Step 5: Fit a regression model to the response of interest Y using the vector of covariates $X(\tilde{t}) = \{X(\tilde{t}_1), \ldots, X(\tilde{t}_{\tilde{N}})\}$. This model will depend on the results of the contrast carried out in Step 4. A linear model will be used if the null hypothesis is not rejected and a nonparametric (e.g. generalized additive model) model otherwise.

ACCEPTED MANUSCRIPT

Step 6: Once the type model has been selected, we propose to apply a forward stepwise regression method to determine the significant covariates, taking advantage of the fact that the local maxima have been ordered. This means we start with a model with the first covariate (the one with the highest value of distance correlation), and the rest of the ordered covariates are added to the model in turn. This substantially reduces the computing time.

As a result, the final number of covariates $\tilde{\tilde{N}}$ will fulfil $\tilde{\tilde{N}} \leq \tilde{N} \leq N$. At any rate, the LMDC approach can be combined with other methods with variable selection such as FWLASSO or NOVAS (see next section).

3. Simulation study

We perform simulations to study the performance of our proposal as compared to other approaches. Following [25], each simulated functional predictor $X_i(t), t \in (0,1)$, is a Brownian bridge stochastic process with zero mean and covariance $cov(X(t), X(s)) = s(1-t)$ for $s < t$, with $X(0) = X(1) = 0$ (Figure 1). We specifically generate samples of size $n = 100$ for the model $Y =< X(t)^l, \beta > +\varepsilon$, with $l = 1$ (a linear model) and $l = 2$ (nonlinear model). Two different functional regression coefficients $\beta(t)$ were considered: a bump function (β_1), showing several sharp peaks, and a heavisine signal (β_2) (Figure 2).

To demonstrate the performance of the our method for different noise levels, we set the variance of the error term σ^2 using the signal-to-noise ratio (SNR) of 5% and 50%. The covariates selected using the LMDC approach were also used as predictor variables to construct several different vector regression models that were compared. Specifically these regression models are: linear (LM), support vector machines (SVM), lasso (LASSO), K-nearest neighbor (KNN) and generalized additive model (GAM). The results were also compared with those obtained from different functional regression models: a) functional regression with principal components (FPCR), b) nonparametric functional regression (NPFR), c) nonparametric variable selection (NOVAS), d) functional wavelet-based weighted LASSO (FWLASSO). Among these methods, the first two apply no variable selection. That is to say they work with all the covariates from a functional perspective.

For each considered scenario we checked all the regression procedures (variable selection and pure functional regression models) using a sample of $n_{train} = 100$ curves for estimation process and $n_{test} = 50$ curves for prediction. The curves are discretized: (i) in $N = 128$ equi-spaced points in $t \in [0,1]$,

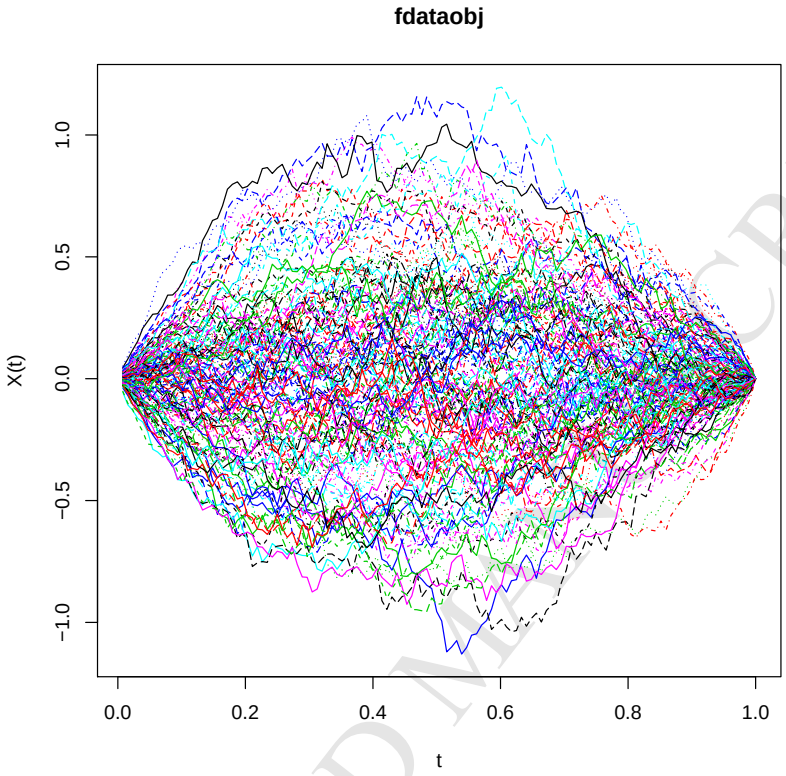


Figure 1: Functional predictors $X_i(t)$, $i = 1, ..., n$

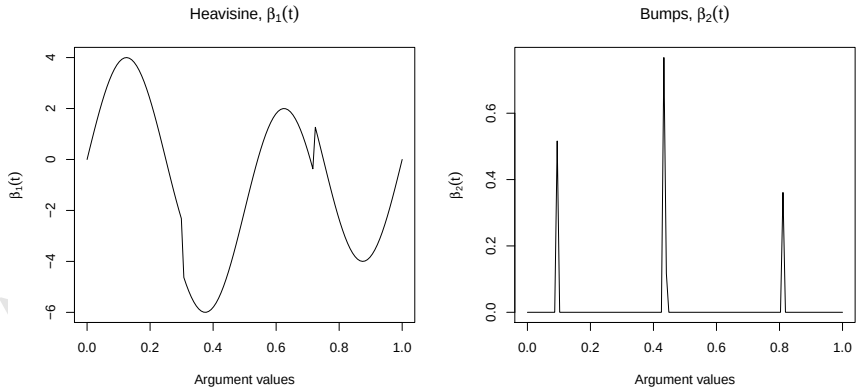


Figure 2: The two functional coefficients $\beta(t)$ simulated. Left: heavisine function; right: bump function.

ACCEPTED MANUSCRIPT

(ii) in $N = 256$. The performance of the methods was compared using the root mean squared error evaluated in a n_{test} independent sample: $RMSE = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{i=n_{test}} (\widehat{Y}_i - Y_i)^2}$.

The algorithm was conducted with the following R packages or scripts: `fda.usc` [45] for functional regression (FLR, FPCR, NPFR); `fda.usc` package for the test of linearity, `mgcv` package [46] for the GAM model; the NOVAS algorithm [37] is available at `https://www.math.univ-toulouse.fr/~ferraty/online-resources.html`; the wavelet-based weighted LASSO for functional regression algorithm (FWLASSO) is available at `http://amstat.tandfonline.com/doi/suppl/10.1080/10618600.2014.925458?scroll=top`.

4. Case study

This work was conducted with four different varieties of grape (Cabernet Sauvignon, Menca, Merlot and Tempranillo) in four vineyards in the village of Cacabelos (León, Spain) that belong to the Bierzo Protected Designation of Origin. All the vineyards shared the same characteristics: row spacing, training system, rootstock and planting year. A total of 162 vines were selected for leaf measurements (47 Cabernet Sauvignon, 45 Mencía, 27 Merlot and 43 Tempranillo vines). Field data collection was carried out on days between berry set and veraison, the time recommended by [47].

We used a FieldSpec 4 portable spectroradiometer (Analytical Spectral Devices, Inc., Boulder, CO, USA) to collect the leaf reflectance data. This spectroradiometer captures spectral data at wavelengths in visible, near-infrared and short-wavelength infrared (the wavelengths ranged from 860 nm to 2500 nm). We also use a plant probe in order to minimize measurement errors associated with stray light. This device consists of a grip to locate the fibre optic cable input to the spectroradiometer, a quartz-halogen bulb, and a quartz window to press the probe against the surface of the leaf [48]. Figure 3 shows how data collection was carried out with a spectroradiometer.

Figure 3: Registering leaf reflectance with a spectroradiometer.

Three mature leaves per vine were measured. The upper face of the leaf was measured three times at three different points (avoiding veins, holes and leafspots) and the mean reflectance value for each leaf was saved. Each measured leaf was cut off, immediately placed in a sealable plastic bag and stored in an insulated cooler. Leaf water content was calculated by the equivalent water thickness (EWT), i.e., the water weight (difference between fresh and dry weight of the leaf) divided by the leaf area.

The spectral reflectance values were pre-processed after modeling, so we actually use the Continuum Removal (CR) as the predictor instead of the reflectance. CR is a transformation of the spectra data used to identify the water absorption features in the leaf spectrum [49]. This transformation normalizes reflectance values to a common baseline and allows us to compare spectra that are either acquired by different instruments or under different light conditions. Figure 4 shows the CR for the vineyard leaves that constitute the sample data.

ACCEPTED MANUSCRIPT

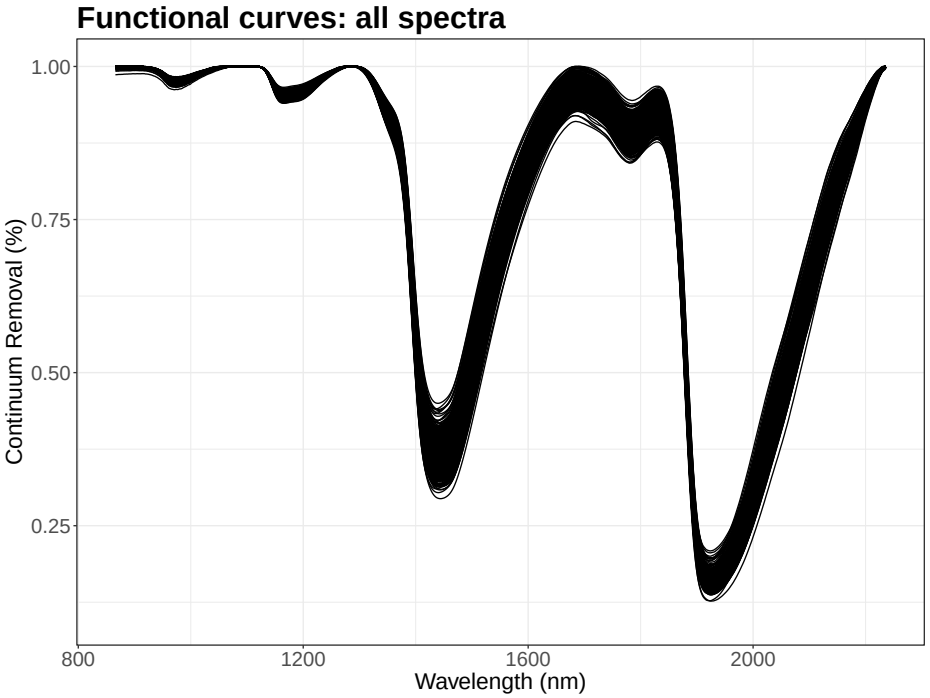


Figure 4: Values of the continuum removal transformation applied to the reflectance measurement for the Tempranillo variety. CR curves for the other varieties are not represented because they overlap previous ones.

The functional regression models described Section 2.1 were applied to a training data of $n_{train} = 200$ samples, leaving the remaining $n_{test} = 85$ to test the models. The analysis was extended to the whole spectrum and studied in four zones: $Z_1 = \{t \in [860 - 1065]\}$, $Z_2 = \{t \in [1114 - 1265]\}$, $Z_3 = \{t \in [1265 - 1668]\}$ and $Z_4 = \{t \in [1830 - 2240]\}$, where reflectance is supposed to be affected by water content. These zones are centered in 970, 1200, 1440 and 1950 nm. This corresponds to the approximate wavelengths at which water has maxima absortion [50].
For each model, the goodness of fit was measured using of the root mean square error.

5. Results

5.1. Simulated data

Table 1 shows the RMSE obtained for the different models tested. As explained in Section 3, we consider two levels of noise (controlled by the

ACCEPTED MANUSCRIPT

SNR), as well as linear and nonlinear (quadratic) regression models. Numbers in bold correspond to the two best models for each column, that is, those with minimum error. As may be seen, the best results were obtained using the LMDC approach for variable selection and the wavelet-based weighted LASSO regression for functional linear models (FWLASSO.mean, FWLASSO.var, means that the penalty term is weighted by the average of the absolute value of the wavelet coefficients or by the variance of these coefficients, respectively). The functional regression model using principal components (FPCR), which implies no variable selection, also performs well. However, when we deal with the quadratic model, models such as LMDE+LM or FWLASSO produce bad results as compared to other like LMDC+GAM or LMDC+NOVAS. The NOVAS method also provides quite good results on its own. Hence the importance of running the linearity test before selecting the appropriate regression model. As expected, the effect of the SNR on model performance presents and increase of the error with the SNR that is especially significant for the linear models.

Table 2 shows the CPU time in seconds for the different regression models tested. As seen, the NOVAS procedure is considerably slower than the rest, and the difference increases with the sample size n and the number of discretization points N. These results correspond to the Heavisine function, but similar results were obtained for the bump function.

ACCEPTED MANUSCRIPT

Table 1: $RMSE$ for the different regression models compared using Heavisine beta parameter. We consider linear and nonlinear regression models, as well as different training-test and signal to noise ratio (SNR). VS indicates whether the method includes variable selection. $RMSE$ is multiplied by a factor $1e+4$ for the Heavisine function and by $1e+8$ for the bump function.

			Linear ($l=1$)				Non-Linear ($l=2$)			
			SNR=0.1		SNR=0.5		SNR=0.1		SNR=0.5	
	Model	VS	$N=128$	$N=256$	$N=128$	$N=256$	$N=128$	$N=256$	$N=128$	$N=256$
Heavisine	LMDC+LM	Y	187	204	725	755	569	598	797	766
	LMDC+SVM	Y	186	208	769	822	611	651	887	844
	LMDC+LASSO	Y	177	193	**706**	**740**	553	581	781	756
	LMDC+kNN	Y	263	261	830	832	199	207	426	413
	LMDC+NOVAS	Y	239	233	784	784	99	**99**	**329**	**325**
	LMDC+GAM	Y	195	214	746	769	**91**	**89**	**331**	**316**
	FWLASSO.mean	Y	178	**173**	762	759	561	576	783	767
	FWLASSO.var	Y	**176**	**173**	762	763	574	575	772	766
	NOVAS	Y	250	243	821	844	**97**	100	339	343
	FPCR	N	**141**	**138**	**682**	**686**	575	608	808	787
	NPFR	N	257	257	828	812	167	191	432	422
Bump	LMDC+LM	Y	247	121	1088	447	612	256	927	329
	LMDC+SVM	Y	257	132	1171	507	661	279	1016	368
	LMDC+LASSO	Y	**238**	119	**1066**	**442**	597	252	909	331
	LMDC+kNN	Y	333	155	1195	505	194	97	493	179
	LMDC+NOVAS	Y	281	126	1115	466	98	**45**	**356**	**137**
	LMDC+GAM	Y	255	125	1118	462	**88**	**43**	**350**	**134**
	FWLASSO.mean	Y	251	**93**	1124	448	593	253	924	329
	FWLASSO.var	Y	253	**92**	1126	450	589	254	923	330
	NOVAS	Y	287	132	1190	500	**91**	49	364	157
	FPCR	N	**249**	94	**1086**	**426**	617	260	942	341
	NPFR	N	329	127	1187	479	166	80	489	176

Table 2: CPU time (*mean*, in seconds) for the different regression models compared using Heavisine beta parameter. We consider linear an nonlinear regression models, as well as different training-test and signal to noise ratio (SNR).

		Linear ($l=1$)				Non-Linear ($l=2$)			
		SNR=0.1		SNR=0.5		SNR=0.1		SNR=0.5	
Model	VS	$N=128$	$N=256$	$N=128$	$N=256$	$N=128$	$N=256$	$N=128$	$N=256$
LMDC+LM	Y	0.800	0.858	0.798	0.844	0.792	0.803	0.819	0.853
LMDC+SVM	Y	1.036	1.160	0.999	1.091	0.914	0.945	0.944	0.983
LMDC+LASSO	Y	0.834	0.883	0.828	0.872	0.826	0.828	0.850	0.886
LMDC+kNN	Y	0.779	0.833	0.776	0.818	0.775	0.778	0.800	0.830
LMDC+NOVAS	Y	1.434	1.491	1.439	1.494	1.476	1.519	1.434	1.496
LMDC+GAM	Y	1.041	1.086	1.003	1.038	1.046	1.102	1.030	1.119
FWLASSO.mean	Y	0.793	1.175	0.956	1.359	1.507	1.545	1.559	1.535
FWLASSO.var	Y	0.782	1.169	0.933	1.354	1.484	1.504	1.547	1.522
NOVAS	Y	4.100	7.634	3.922	7.260	4.628	8.438	3.928	7.436
FPCR	N	0.730	1.088	0.741	1.076	0.739	1.073	0.738	1.082
NPFR	N	0.474	0.624	0.472	0.621	0.478	0.639	0.469	0.625

5.2. Experimental data

5.2.1. Comparing wavelength intervals through distance correlation

Using distance correlation, we compared the information contained in each of the four spectrum zones under study before carrying out the regres-

ACCEPTED MANUSCRIPT

sion analysis.. Previous works in distance correlation can be found in [51], where the author proposes a variable selection algorithm to select an optimal subset of covariates in a functional regression framework, and in [52], where the Minimum Redundance Maximum Relevance (mRMR) procedure is applied to choose the most relevant design points in functional classification setting.

Table 3 shows the distance correlation between the reflectance at different wavelengths and the water content, for the whole spectrum (Z_{1-4}) and for the zones $Z_i, i = 1, ..., 4$. We may see that the maximum value corresponds to Z_3. Moreover, $\mathcal{R}(Z_2, Z_3) = \{0.85, 0.90, 0.77, 0.93\}$ for Cabernet, Mencia, Merlot and Tempranillo varieties, respectively. Then, we conclude that both areas contain or share the same type of information. Consequently, from now on we will compare the results corresponding to Z_3 with those for the whole wavelength interval Z_{1-4}.

Table 3: Distance correlation $\mathcal{R}$ between response (water content) and predictor variables (reflectance) for the different zones under study and four varieties.

Variety	Zone				
	Z_1	Z_2	Z_3	Z_4	Z_{1-4}
Cabernet	0.18	0.22	**0.29**	0.18	0.26
Mencia	0.43	0.46	**0.47**	0.29	0.44
Merlot	0.44	0.35	**0.52**	0.30	0.47
Tempranillo	0.57	0.58	**0.61**	0.40	0.57

5.2.2. Functional regression

Two linear regression models, one using Fourier functions (FLR) and the other using principal components (FPCR) as basis functions, as well as the nonparametric functional model (NPFR) explained in Section 2.1.4 were applied to the data as classical representatives of functional regression. Table 4 shows the RMSE valued for the test sample, for each of the four grape varieties. We may observe that similar results were obtained for the three functional models evaluated. Zone $Z_3(t)$ $t \in [1265, 1668]$ produced even better results than those corresponding to the whole wavelength interval studied $Z_{1-4}(t)$ $t \in [865, 2500]$. The dimension of the problem may be reduced by approximately 1/4 (from 1635 wavelengths to 403). Obviously, this is important because computing time is reduced and the problem is simplified.

ACCEPTED MANUSCRIPT

Table 4: RMSE x 10^5 values obtained applying the three functional regression models for each of the varieties, for zones Z_3 and for the whole spectrum Z_{1-4} to the test sample.

Variety	Zone	FLR	FPCR	NPFR
Cabernet	Z_{1-4}	253	**251**	252
Cabernet	Z_3	256	245	**242**
Mencia	Z_{1-4}	171	**167**	186
Mencia	Z_3	169	**163**	172
Merlot	Z_{1-4}	158	**143**	164
Merlot	Z_3	148	**142**	152
Tempranillo	Z_{1-4}	220	**210**	255
Tempranillo	Z_3	201	**198**	241

We generally obtained the minimum prediction errors for the model using a PC basis rather than a Fourier basis. For the sake of simplicity we will limit the analysis to these basis functions from now on.

Also, as shown in Table 4, in general the nonparametric regression model produces worse results in terms of error than the linear models. This suggests a possible linear relationship between water content and reflectance. Table 5 shows the results of the linearity test (step 4 of our algorithm). In all cases the p-value> 0.05, so the null hypothesis of linearity cannot be rejected.

Table 5: p-values of the linearity test for each of the varieties, for zones Z_3 and for the whole spectrum Z_{1-4}.

Variety	Zone	
	Z_3	Z_{1-4}
Cabernet	0.28	0.56
Mencia	0.38	0.40
Merlot	0.82	0.50
Tempranillo	0.22	0.68

5.3. Selection of the optimum wavelengths

The estimation of the design points, that is, the optimum wavelengths, was carried out following the procedures described in Sections 2.3.1 and 2.3.2. Table 6 shows the RMSE obtained using both methods, for each of the four grape varieties analysed and for zones Z_3 and Z_{1-4}. Those RMSE values

ACCEPTED MANUSCRIPT

correspond to regression models that relate water content (response variable) with the optimum wavelengths (predictor covariates). As may be seen, LMDC has a better performance than NOVAS in determining the optimum wavelength. Moreover, in LMDC is much faster than NOVAS mainly for two reasons: 1) distances are calculated only once in LMDC, while distances are calculated each time a predictor variable is included in the NOVAS procedure, 2) in LMDC+LM, stepwise LM scheme estimates only $\tilde{N}$ regression models to provide the final model with $\tilde{\tilde{N}}$ predictors ($\tilde{\tilde{N}} \leq \tilde{N} << N$), in LMDC+NOVAS, $\tilde{N}$, $\tilde{N}-1$,..., $\tilde{N}-\tilde{\tilde{N}}$ regression models are adjusted, while the NOVAS procedure requires adjusting many more models, specifically N, $N-1$,..., $N-\tilde{\tilde{N}}$. In addition, NOVAS requires performing cross-validation in each step. FWLASSO has also a good performance, but errors are generally greater than those corresponding to the LMDC+LM approach, which, in addition, is a method that is easy to interpret.

Table 6: RMSE x 10^5 of the regression models obtained limiting the covariates to the optimum wavelengths obtained with the functional and distance correlation approaches. Results correspond to the test sample.

Variety	Zone	LMDC				FWLASSO		NOVAS
		LM	LASSO	NOVAS	GAM	mean	var	
Cabernet	Z_{1-4}	**241**	243	246	243	260	260	288
Cabernet	Z_3	243	246	**242**	245	245	245	272
Mencia	Z_{1-4}	172	**170**	172	173	175	185	180
Mencia	Z_3	169	**164**	169	171	171	181	175
Merlot	Z_{1-4}	**145**	149	168	152	174	174	179
Merlot	Z_3	**130**	138	146	**130**	163	163	150
Tempranillo	Z_{1-4}	212	211	246	217	**197**	197	265
Tempranillo	Z_3	222	209	226	224	**190**	190	243

As the optimum wavelengths are not always the same in each regression model, the selection criteria was to retain those appearing in most models. Figure 5 shows, at the top, the distance correlation for a set of 50 functions obtained upon resampling the original dataset, for both Z_{1-4} and Z_3 areas. A histogram of frequencies reflecting the number of times a specific wavelength appears in a regression model is shown at the bottom. The highest frequencies, over 20, correspond to narrow bands around 1326 and 1515. Z_{1-4} also presents a concentration of wavelengths around 2216 nm undiscovered using

ACCEPTED MANUSCRIPT

the NOVAS procedure. These wavelengths are close to those where water absorption is maxima, according to [50].

Similar results were obtained applying the NOVAS procedure. This means that we can limit our study to this small region of the wavelength spectrum. In addition, more than half of the models only have two covariates, so we conclude that an adequate prediction for leaf water content from the reflectances may be obtained at just two optimal wavelengths.

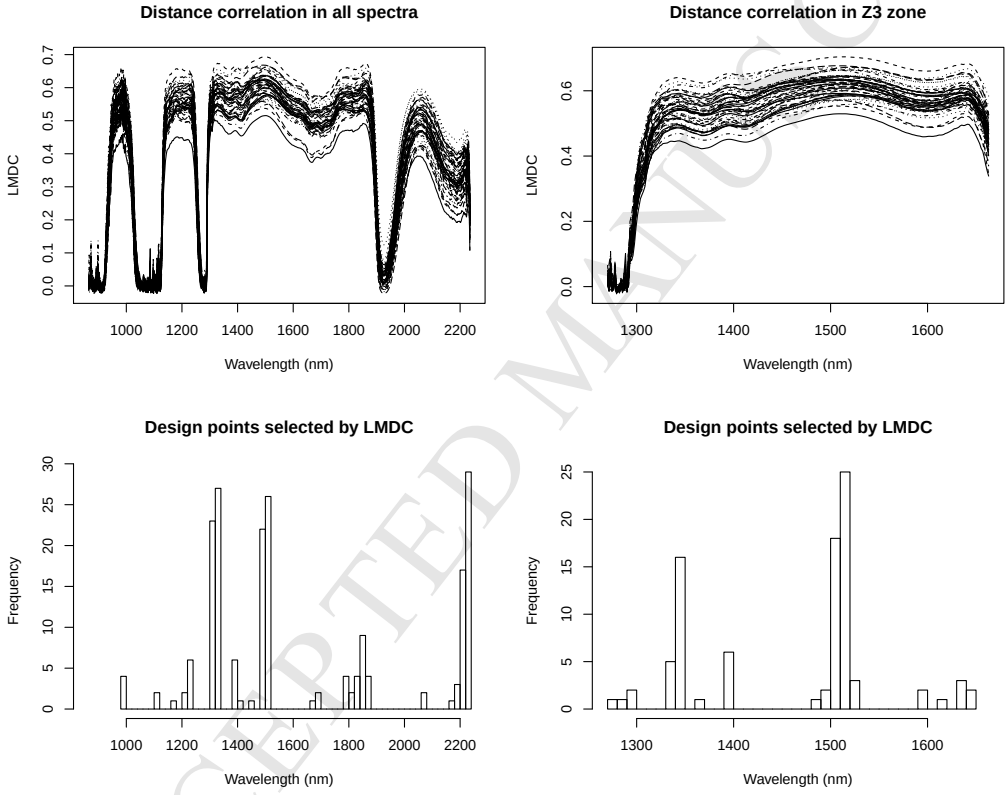


Figure 5: Distance correlation between response for all spectra Z_{1-4} (top left) and limited to Z_3 zone (top right), in $nrep = 50$ resamples. Histogram of design points selected using LMDC for all spectra Z_{1-4}(bottom left) and Z_3 zone (bottom right). The information corresponds to the Tempranillo variety.

Furthermore, if we represent the coefficients $\beta(t)$ of the linear regression model for the 50 samples (Figure 6), we can appreciate that these narrow bands correspond to local maximum and minimum of $\hat{\beta}$ that are significantly

different from zero. Moreover, $\hat{\beta}$ is null from 860 nm to a value around 1250 nm, which is in line with the fact that zone Z_1 provide no significant information regarding leaf water content.

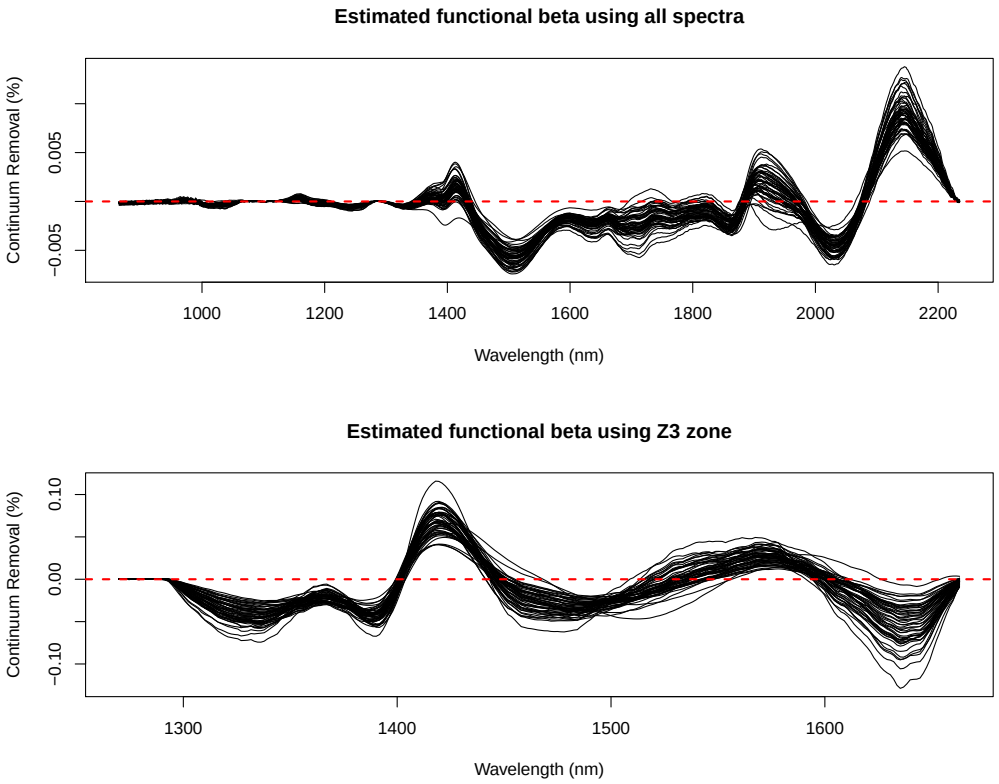


Figure 6: Functional linear regression coefficients for all spectrum under study Z_{1-4} (top) and for Z_3 zone (bottom).

6. Conclusions

Estimating leaf water content from reflectance by means of functional regression has some advantages with respect to other methods that consider reflectance a set of discrete points instead of functions. However, many researchers prefer the latter type of methods, such as vegetation indices, because they consider no advantage in using complex mathematical techniques or simply because they are unfamiliar with them. In this work we show that

ACCEPTED MANUSCRIPT

functional data analysis also provides interesting information that can be useful for these researchers.

We study the utility of the distance correlation statistic in selecting optimum wavelengths to estimate water content from reflectance. First our method (LMDC) was tested with simulated data and compared with other methods, both for a linear and a nonlinear regression models. The results showed that LMDC has some advantages over other methods, mainly because we assume no model and because it requires no transformation of the original covariates. Second, the method was tested in a real dataset of four vineyard varieties. The results obtained show us that, indeed, estimating water content only requires the reflectance at few wavelengths. Particularly, two narrow bands between 1326 nm and 1515 nm, respectively, contain most of the information. Furthermore, when this method for detecting optimum wavelengths was compared with two different functional approaches previously proposed by other authors, one based on a kernel smoothing and the other on a wavelet-based weighted LASSO, we checked that distance correlation produces better results, whether in terms of error estimation, simplicity or computation time. In addition, our proposal does not need to assume any type of regression model, any one may be used depending on the nature of the problem.

In short, determining local maxima in the distance correlation function is an efficient functional variable selection method that can help to construct simple regression models for leaf water content estimation from reflectance.

Acknowledgments

This study was made possible with financial funding from: a) FC-15-GRUPIN14-033 of the Fundación para el Fomento en Asturias de la Investigación Científica Aplicada y la Tecnología (FICYT) (Spain), with FEDER support included, b) Ministry of Economy and Competitiveness (MTM2016-76969P) and European Regional Development Fund (ERDF), c)Grupo de Referencia Competitiva, 2016-2019 (ED431C 2016/040), financiado pola Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia.

References

[1] M.M. Chaves, O. Zarrouk, R. Francisco, J.M. Costa, T. Santos, A.P. Regalado, M.L. Regalado, and C.M. Lopez. Grapevine under deficit ir-

ACCEPTED MANUSCRIPT

rigation - hints from physiological and molecular data. *Annals of Botany*, 10(5):661–676, 2010.

[2] J.A. Kennedy, M.A., Matthews, and A.L. Waterhouse. Effect of maturity and vinewater status on grape skin and wine flavonoids. *American Journal of Enology and Viticulture*, 53:661–676, 2002.

[3] N.C. Turner. Measurement of plant water status by the pressure chamber technique. *Irrigation Science*, 9(4):289–308, 1988.

[4] P. Ceccato, S. Flasse, S. Tarantola, S. Jacquemoud, and J. Grégoire. Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of Environment*, 77:22–33, 2001.

[5] A.E. Strever. Estimating water stress in vitis vinifera l. using field spectrometry: a preliminary study incorporating multispectral vigour classification. In *Proceedings of the conference FRUTIC 05, information and technology for suistanaible fruit and vegetation production*, Montpellier, France, 2005.

[6] D. Moshou, X. Pantazi, D. Kateris, and I. Gravalos. Water stress detection based on optical multisensor fusion with a least squares support vector machine classifier. *Biosystem Engineering*, 117:15–22, 2014.

[7] Z. Oumar and O. Mutanga. Predicting plant water content in eucalyptus grandis forest stands in kwazulu-natal, south africa using field spectra resampled to the sumbandila satellite sensor. *International Journal of Applied Earth Observation and Geoinformation*, 12(3):158–164, 2010.

[8] R. De Bei, D. Cozzolino, W. Sullivan, W. Cynkar, S. Fuentes, R. Dambergs, J. Pech, and S. Tyerman. Non-destructive measurement of grapevine water potential using near infrared spectroscopy. *Australian Journal of Grape and Wine Research*, 17:62–71, 2011.

[9] S.Z. Dobrowski, S.L. Ustin, and J.A. Wolpert. Remote estimation of vine canopy density in vertically shootpositioned vineyards: determining optimal vegetation indices. *Australian Journal of Grape and Wine Research*, 8:117–125, 2002.

[10] D.A. Sims and J.A. Gamon. Estimation of vegetation water content and photosynthetic tissue area from spectral reflectance: A comparison

ACCEPTED MANUSCRIPT

of indices based on liquid water and chlorophyll absorption features. *Remote Sensing of Environment*, 84:526–537, 2003.

[11] L. Serrano, C. González-Flor, and G. Gorchs. Assessment of grape yield and composition using the reflectance based water index in mediterranean rainfed vineyards. *Remote Sensing of Environment*, 118:249–258, 2011.

[12] Y.L. Grossman, S.L. Ustin, S. Jacquemoud, E.W. Sanderson, G. Schmuck, and J. Verdebout. Critique of stepwise multiple linear regression for the extraction of leaf biochemistry information from leaf reflectance data. *Remote Sensing of Environment*, 56:182–193, 1996.

[13] R. F. Kokaly and R.N. Clark. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sensing of Environment*, 67:267–287, 1999.

[14] T. Cheng, B. Rivard, and A. Sánchez-Azofeifa. spectroscopic determination of leaf water content using continuous wavelet analysis. *Remote Sensing of Environment*, 115:659–670, 2011.

[15] P.J. Zarco-Tejada, J.R. Miller, T.L. Noland, G.H. Mohammed, and P.H. Sampson. Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 39:1491–1507, 2001.

[16] P.J. Zarco-Tejada, C.A. Rueda, and S.L. Ustin. Water content estimation in vegetation with modis reflectance data and model inversion methods. *Remote Sensing of Environment*, 85:109–124, 2003.

[17] C. Ordónez, J. Martínez, J.M. Matías, A.N. Reyes, and J. R. Rodríguez-Pérez. Functional statistical techniques applied to vine leaf water content determination. *Mathematical and Computer Modelling*, 52(7-8):1116–1122, 2010.

[18] C. Ordónez, J. R. Rodríguez-Pérez, J.J. Moreira, and E. Sanz. Using hyperspectral spectrometry and functional models to characterize vine-leaf composition. *IEEE Transactions on Geosciences and Remote Sensing*, 51(5):2610–2618, 2013.

ACCEPTED MANUSCRIPT

[19] P.T. Reiss and R.T. Odgen. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996, 2012.

[20] J. R. Berrendero, A. Cuevas, and J. T. Torrecilla. Variable selection in functional data classification: A maxima-hunting proposal. *Statistica Sinica*, 26, 2015.

[21] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 26:301–320, 2005.

[22] P. Buhlmann and B. Yu. Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Assocation*, 26:324–339, 2003.

[23] J. Gertheiss and G. Tutz. Sparse modeling of categorical explanatory variables. *The Annals of Applied Statistics*, 4:324–339, 2010.

[24] L. Comminges and A. S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Applied Statistics*, 40:2667–2696, 2012.

[25] Y. Zhao, R.T. Ogden, and P.T. Reiss. Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617, 2012.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[27] Yihong Zhao and R. Todd Ogden. Wavelet-based weighted lasso and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics*, 24(3):655–675, 2015.

[28] G.J. Székely, M.L. Rizzoand, and N.K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[29] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, 1997.

ACCEPTED MANUSCRIPT

[30] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 1978.

[31] J.S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, 1996.

[32] D.J. Levitin, R.L. Nuzzo, B.W. Vines, and J.O. Ramsay. Introduction to functional data analysis. *Canadian Psychology*, 48(3):135–155, 2007.

[33] M. Febrero-Bande, P. Galeano, and M. González-Manteiga. Functional principal component regression and functional partial least squares regression: An overview and a comparative study. *Int. Statist. Rev.*, 85(1):61–83, 2015.

[34] G:M. James, J. Wang, and J. Zhu. Linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108, 2009.

[35] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[36] M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407, 1996.

[37] F. Ferraty, P. Hall, and P. Vieu. Most predictive design points for functional data predictors. *Biometrika*, 94(4):807–824, 2010.

[38] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.

[39] L. Li, R. D. Cook, and Ch.J. Nachtsheim. Model-free variable selection. *Journal of the Royal Statistical Society. Series B*, 67(2):285–299, 2005.

[40] J.U.Yule and M.G. Kendall. *An Introduction to the Theory of Statistics (Fourteenth Edition)*. Charles Griffin, 1950.

[41] R. Lyons. Distance covariance in metric spaces. *Ann. Probab.*, 41(5):3284–3305, 2013.

[42] F. Ferraty and P. Vieu. *A practial Guide to Splines*. Springer-Verlag, 2006.

ACCEPTED MANUSCRIPT

[43] Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1d. *Annals of statistics*, 39(6):3290, 2011.

[44] E. García-Portugués, Wenceslao W. González-Manteiga, and M. Febrero-Bande. A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23(3):761–778, 2014.

[45] M. Febrero-Bande and M. O. de la Fuente. Statistical computing in functional data analysis: the R package fda.usc. *J. Statist. Software*, 51(4):1–28, 2012.

[46] S. Wood. *mgce: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, 2017. R package version 1.8.22.

[47] A.O. Santos and O. Kaye. Grapevine leaf water potential based upon near infrared spectroscopy. *Scientia Agricola*, 66:287–292, 2009.

[48] I.C. Lau, T.J. Cudahy, G. Heinson, A.J. Mauger, and P.R. James. Practical applications of hyperspectral remote sensing in regolith research. *Advances in Regolith*, 66:249–253, 2003.

[49] R.N. Clark and T.L. Roush. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, (89):6329–6340, 1984.

[50] K. Palmer K. and D. Williams. Optical properties of water in the near infrared. *Journal of the Optical Society of America*, 64:1107–1110, 1974.

[51] M. Febrero-Bande, W. González-Manteiga, and Manuel Oviedo de la Fuente. Variable selection in functional additive regression models. In *Functional Statistics and Related Fields*, pages 113–122. Springer, 2017.

[52] J.R. Berrendero, A. Cuevas, and J.L. Torrecilla. The mrmr variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation*, 86(5):891–907, 2016.

ACCEPTED MANUSCRIPT

HIGHLIGHTS

- We look for optimal wavelengths to estimate leaf water content from reflectance.
- A new method called Local Maximum Distance Correlation (LMDC) is proposed.
- A non-parametric functional approach (NOVAS) is also evaluated.
- LMDC improved NOVAS results in the case study analyzed.

ACCEPTED MANUSCRIPT