Chemometrics and Intelligent Laboratory Systems 173 (2018) 41–50

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

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), Campus Vida, Santiago de Compostela, Spain
[c] MODESTYA Group, Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Campus Vida, Santiago de Compostela, Spain
[d] Department of Statistics, University of Vigo, Spain
[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), Spain

ARTICLE INFO

Keywords:
Vineyard
Water content
Distance correlation
Functional analysis
Design points

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 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.

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 chlorophyll, 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 [9–11], multiple regression models [12–14] 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.

[*] Corresponding author. Edif. Instituto Investigaciones Tecnológicas, planta -1, Rúa de Constantino Candeira s/n, 15782 Campus Vida, Santiago de Compostela, Spain.
E-mail address: manuel.oviedo@usc.es (M. Oviedo de la Fuente).

https://doi.org/10.1016/j.chemolab.2017.12.001
Received 26 July 2017; Received in revised form 23 November 2017; Accepted 1 December 2017
Available online 6 December 2017

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 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 Ref. [23]. In this particular case, the effect of one variable can be determined not by one, but by several coefficients. Authors in Ref. [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 Ref. [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 [25, 29] 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.

2.1. Functional regression models

Consider a sample data $\{X_i, Y_i\}_{i=1}^n$ where $X_i = (X_i(t_1), X_i(t_2), \dots, X_i(t_N))$ 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 $(t_1, t_2, \dots, t_N)$ and Y_i the water content of each vine leaf. We can assume that both variables are related by the model

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

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 $X_i(t)$, 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 $X_i \in \mathscr{L}_2(T)\ \forall t \in [a, b]$, and $Y_i \in \mathbb{R}$, a parametric functional linear model, as formulated in Ref. [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 $\beta(t) \in \mathscr{L}_2(T)$ are the regression coefficients. In this model $X_i(t)$ and $\beta(t)$ are approximated by means of decomposition in K basis functions

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

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:

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

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

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