Catena 208 (2022) 105718

Contents lists available at ScienceDirect

Catena

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

Correlation of banana productivity levels and soil morphological properties using regularized optimal scaling regression

Barlin O. Olivares [a,*], Julio Calero [b], Juan C. Rey [c,d], Deyanira Lobo [d], Blanca B. Landa [e], José A. Gómez [e]

[a] Universidad de Córdoba. Campus Rabanales, Programa de Doctorado en Ingeniería Agraria, Alimentaria, Forestal y del Desarrollo Rural Sostenible. Carretera Nacional IV, km 396, 14014 Córdoba, España
[b] Universidad de Jaén, Departamento de Geología, Facultad de Ciencias Experimentales, Campus Universitario Las Lagunillas, 23071 Jaén, España
[c] Instituto Nacional de Investigaciones Agrícolas (INIA-CENIAP), Av. Universidad vía El Limón, 02105 Maracay, Venezuela
[d] Universidad Central de Venezuela, Facultad de Agronomía. Av. Universidad, Maracay, Venezuela
[e] Instituto de Agricultura Sostenible CSIC, Avenida Menéndez Pidal s/n, 14004 Córdoba, España

ARTICLE INFO

Keywords:
Biological activity
Sustainability
Qualitative soil indicators
Dry consistence
Soil structure
Texture

ABSTRACT

Soil morphological properties described in the field, such as texture, consistence or structure, provide a valuable tool for the evaluation of soil productivity potential. In this study, we developed a regression model between the soil morphological variables of banana plantations and a crop Productivity Index (PI) previously developed for the same areas in Venezuela. For this, we implemented categorical regression, an optimal scaling procedure in which the morphological variables are transformed into a numerical scale, and can thus be entered in a multiple regression analysis. The model was developed from data from six plantations growing "Gran Nain" bananas, each with two productivity levels (high and low), in two 4-ha experimental plots, one for each productivity level. Sixty-three A horizons in thirty-six soils were described using 15 field morphological variables on a nominal scale for structure type, texture and hue, and an ordinal scale for the rest (structure grade, structure size, wet and dry consistence, stickiness, plasticity, moist value, chroma, root abundance, root size, biological activity and reaction to HCl). The optimum model selected included biological activity, texture, dry consistence, reaction to HCl and structure type variables. These variables explained the PI with an R^2 of 0.599, an expected prediction error (EPE) of 0.645 and a standard error (SE) of 0.135 using bootstrapping, and EPE of 0.662 with a SE of 0.236 using 10-fold cross validation. Our study showed how soil quality is clearly related to productivity on commercial banana plantations, and developed a way to correlate soil quality indicators to yield by using indicators based on easily measured soil morphological parameters. The methodology used in this study might be further expanded to other banana-producing areas to help identify the soils most suitable for its cultivation, thereby enhancing its environmental sustainability and profitability.

1. Introduction

The banana is one of the most important crops in the world after rice, wheat and corn, both in terms of production yields and area cultivated. This fruit constitutes the basis of the diet in tropical countries like Costa Rica, Colombia, Ecuador, Panama and Venezuela. It is also an important source of income for producers (FAO, 2020). Identification of the most suitable areas for banana cultivation is essential to increasing its productivity in tropical regions. Therefore, soil properties must be characterized to understand their relationship to crop productivity (Villarreal-Núñez et al., 2013; Delgado et al., 2010b).

Soil morphological field properties have been recognized as a valuable tool for studying a broad range of soil characteristics, including those related to soil development in agricultural areas, for the ease and speed with which they can be described (Soil Survey Staff, 2017; Calero et al., 2008; Pulido-Moncada et al., 2017). Soil morphological properties are relatively easily and economically characterized in soil pits, almost all are included in soil databases, and they can be easily determined by technicians (Delgado et al., 2010a). Unlike soil chemistry and most of its biological properties, field morphological data are generally nonnumerical (nominal or ordinal) and measured on evaluation scales with origins difficult to establish (Vaughan & Ormerod, 2005; Meulman et al.,

* Corresponding author.
E-mail addresses: barlinolivares@gmail.com, ep2olcab@uco.es (B.O. Olivares).

https://doi.org/10.1016/j.catena.2021.105718
Received 1 June 2021; Received in revised form 19 August 2021; Accepted 3 September 2021
Available online 11 September 2021

2002). Similarly, numerical relationships of different categories are still poorly developed. Such data cannot be subjected to rigorous treatment by statistical methods, such as factor analysis or multiple regression, like numerical quantitative data (Andrews et al., 2004; Linting et al., 2007; Calero et al., 2005). Therefore, there is little background for calculating soil quality indices from categorical indicators established by decision rules based on existing knowledge (Pulido-Moncada et al., 2014, Calero et al., 2018).

Soils in banana plantations in Venezuela were first characterized at the beginning of the 20th century. Other regional and farm characterization studies were carried out under Venezuelan research and development programs from 1970 to 1998 (Martínez et al., 2008). Some of the most recent soil characterization studies in banana zones are those by Hernández et al. (2007), Rey, 2009, Delgado et al. (2010a, 2010b), González-Pedraza et al. (2014), Olivares et al. (2020), González García et al. (2021a, 2021b, 2021c) and González-Pedraza and Carlos Escalante (2021). From the 1970 s onwards, the worldwide trend toward intensification of banana production systems was also observed in Venezuela, and the relationship between reduction in productivity and loss of soil quality became apparent. As a result, growers are looking for innovative sustainable management methods that can simultaneously maintain or increase productivity. Improved use of available, or newly acquired, soil information, is essential for achieving such sustainable banana production in tropical countries.

The link between soil quality and current or potential productivity is critical in the search increased productivity and sustainability in agriculture. As mentioned, most approaches are based on quantitative variables, while the potential of morphological variables has still only been moderately explored. Several methods have been studied, in which the high potential of categorical regression (CATREG), already in use in such fields as education, marketing, and agricultural economics, among others (van der Kooij, 2007; Meulman et al., 2019; Sevinç et al., 2019), has been highlighted. However, to our knowledge, it has not been extensively used in agronomic sciences and never in banana fields. This is a novel element in our study, in which soil morphological properties were explored as promising new soil indicators for assessing banana productivity in Venezuelan soils. Our study is a pioneer in the application of CATREG, an optimal scaling procedure, to the transformation of morphological variables into a numerical scale, which is completely novel in banana soils. Nevertheless, beyond its application in the case of a specific crop (banana) and geography (Venezuela), we have developed a scientific rationale that is easily transferred to other areas, not only in agriculture, but soil science in general.

According to many authors (i.e., MacEwan and Fitzpatrick, 1996; Lal, 1998; and just recently Vasu et al., 2021), it is important for soil assessment to be based on such characteristics as morphological properties, that are measured easily and inexpensively. Beyond its taxonomic utility, morphological properties can improve soil assessment. A wealth of such data is available from agricultural services (universities, research centers, etc.), just waiting to be usefully employed, and our study is a fine example of how it can be used.

This study aimed to validate the hypothesis that soil morphological properties can differentiate banana productivity levels in large areas of Venezuela using a categorical regression prediction model. In this case, qualitatively estimated soil morphological properties could be used for improving the assessment of banana productivity. Categorical regression analysis can provide an operating model for bananas in an area where there is little background of soil quality indices with categorical soil properties.

2. Materials and methods

2.1. Description of the study areas and banana plantations

Six banana plantations in the states of Aragua and Trujillo in Venezuela were selected (Table 1). The plantations in the State of Aragua (PL, SM, PZ and CH) are in the Lake Valencia Basin. They are characterized by a tropical savanna (Aw) climate with a mean annual rainfall depth of 980 mm. Rainfall in this area is seasonal for five to six rainy months, concentrated between May and October (Olivares et al., 2021). The mean annual temperature is 26.2 °C and the mean annual relative humidity is 70.0% (Olivares, 2018). The terrain relief is flat (slope 0–2%). Plantation PL is located on the fourth level of the lacustrine terrace produced by drying of Lake Valencia, while Plantations SM, PZ and CH are on alluvial soils, all with medium to silty textures. Soils in these farms are Mollisols and Inceptisols, generally with moderate to good drainage, soil pH neutral to alkaline, fertile, with medium to high organic matter content (Delgado et al., 2010a).

The second study area is located in the State of Trujillo (Plantations BA and KA), in the region southeast of Lake Maracaibo, also characterized by a tropical savanna (Aw) climate. The mean annual precipitation is 1094 mm, with two rainfall peaks, one in April-May (monthly precipitation approximately 120 mm) and the other in October (monthly precipitation approximately 145 mm). The driest months are January and February when the monthly precipitation is<50 mm (Olivares et al., 2017). The area has a mean annual temperature of 27.5 °C and a mean relative humidity of 78.0%. This area of the State of Trujillo is an alluvial plain with slopes of <1.0% with mainly Entisol soils (Rodríguez et al.,

Table 1
Geographic location and planted area of bananas (ha) of the sampled plantations in Venezuela.

Plantations code	Geographical coordinates	Height (masl)	Sites	Soil Taxonomy†	State	planted area (ha)
PL	10° 12′ N; 67° 30′ W	435	H	Mollic Ustifluvents Fluventic Haplustolls Cumulic Haplustolls	Aragua	135
			L	Oxyaquic Ustifluvents Fluventic Haplustolls		
SM	10° 12′ N; 67° 23′ W	502	H	Fluventic Haplustepts Fluventic Haplustolls	Aragua	11
			L	Fluventic Haplustolls		
PZ	10° 11′ N; 67° 31′ W	514	H	Fluventic Haplustolls	Aragua	20
			L	Fluventic Haplustepts		
CH	10° 11′ N; 67° 31′ W	498	H	Fluventic Haplustolls	Aragua	9
			L			
BA	09° 29′ N; 70° 57′ W	16	H	Oxyaquic Ustifluvents Aeric Fluvaquents	Trujillo	300
			L	Typic Ustifluvents Oxyaquic Ustifluvents		
KA	09° 28′ N; 70° 55′ W	17	H	Oxyaquic Ustifluvents Fluventic Haplustoll	Trujillo	270
			L	Typic Ustifluvents Oxyaquic ustifluvents		

† Soil Survey Staff (2014). Sites: H: High and L: Low productivity.

2006). These soils have moderate to poor drainage with neutral to alkaline pH. They are moderately fertile and average organic matter content is around 2.75% (Rey, 2009). In both areas, soil management was concentrated on fertilization, and no reclamation action was taken to improve drainage or increase organic matter.

2.2. Soil sampling

Banana productivity was estimated in sampling areas delimited by productivity level following the guidelines proposed by Rosales et al. (2008). On all the plantations, two plots or productivity levels were identified a priori as High (H) and Low (L) for estimating the Productivity Index (PI). The "Gran Nain" variety was the only variety grown, and each productivity plot had an area of 4 ha (on the large > 50 ha plantations, PL, BA and KA) with four replicated plots in each field. On large plantations (≥50 ha, PL, BA and KA), the average yield of high productivity plots was 69.8 ± 5.0 t ha^{-1} yr^{-1} and in low productivity plots, it was 59.7 ± 5.3 t ha^{-1} yr^{-1}. On the remaining small plantations (<25 ha, SM, PZ and CH) these two levels of productivity were identified in 1-ha fields, with two replicated plots in each field, for a total of 36 plots. The average yield on small plantations was 11.5 ± 0.7 t ha^{-1} yr^{-1} for high productivity plots and 1.6 t ha^{-1} yr^{-1} for those with low productivity.

2.3. Productivity index (PI)

Our study used the banana productivity index (PI) previously developed by Olivares et al. (2020) to estimate productivity in each of the evaluation plots. The PI is based on the morphometric characteristics of banana plants, such as circumference of the mother plant pseudo-stem at 1 m height (cm), number of hands per bunch (n), and height of the succession plant (cm), following the methodology proposed by Rosales et al. (2008). The PI was generated from a principal component analysis (PCA) in which the variables best represented on the first axis (Principal Component 1) were retained. In addition, the linear combination of the three biometric variables evaluated was used as a synthesis variable, where the coefficients of this linear combination were the variable loadings in PC1. Then, in the categorical regression the ordinal scaling level was considered the response variable.

2.4. Morphological characterization of soil properties

A soil profile was evaluated in each of the thirty-six replicated plots indicated in Section 2.2. The A horizon morphological properties were described and characterized in the field following the FAO (2006) and the Soil Survey Staff (2017) methodologies. Color was determined according to the Munsell soil color charts (Munsell Color Company, 1999). Texture was determined in the laboratory by sieving and sedimentation, using a Robinson pipette, (Soil Survey Staff, 2017). Fifteen morphological variables in these soil profile descriptions were studied: texture class, structure size, structure grade, structure type, moist hue, moist value, moist chroma, moist consistence, dry consistence, stickiness, plasticity, biological activity, root abundance, root size and reaction to HCl.

2.4.1. Sample analysis

Our analysis was applied to the A horizons (N = 63) in each soil. The average thickness of these horizons is 31 ± 12 cm, which roughly coincides with the area of maximum concentration of the banana tree (30 cm). We hypothesized that focusing our analysis on the A horizon would make more agronomic sense, since this is the area that most influences plant development and yield. The functional roots of the banana in our profile descriptions were concentrated in the A horizons which ranged, from about the top 30 to 40 cm deep, a root concentration range similar to the description of banana in the tropics by Gauggel et al. (2005).

2.5. Statistical modelling of field data

2.5.1. Regression with optimal scaling

First, a regression model was fitted to be able to predict the banana productivity index (PI) (*outcome*) based on the soil morphological properties (*predictors*) on the plantations sampled. However, the categorical or nonmetric nature of the predictors prevents proper application of classical regression analysis. Various techniques have been applied in an attempt to include categorical variables in multiple regression, such as the creation of dummy variables. These variables, however, introduce high multicollinearity and hinder interpretation of the model, reducing the significance of the regression coefficients (Wissmann et al., 2011) and the predictive power of the model (Hair et al., 1999; Xu et al., 2010).

The problem of multicollinearity between predictors can be dealt with by selecting a good theoretical model, which enters only those predictors with a high degree of unique variance with the dependent variable (Hair et al., 1999). However, it is hard to select categorical predictors using the usual stepwise procedure, where each variable is selected or discarded before its transformation to a numeric scale. For rigorous treatment of categorical variables, such as field morphological variables, in multivariate methods, optimal scaling methods need to be applied (Calero et al., 2018).

Categorical regression (CATREG) (van der Kooij et al., 2006; Meulman et al., 2019) is an optimal scaling technique that can transform the k_j categories ($s = 1, \ldots, j$) of the j^{th} nonmetric predictors, and the k_r categories of the response variable r, by means of nonlinear numerical functions, while minimizing the error (that is, increasing the coefficient of determination R^2 of the model). CATREG takes the classical linear regression model with nonlinear optimal scaling (Equation (1)):

$$\varphi(\mathrm{y}) = \sum_{j=1}^{p} \beta_j \varphi(\mathrm{x}) + e \tag{1}$$

where β_j are the regression coefficients of the j^{th} predictor, $\varphi(\mathrm{x})$ and $\varphi(\mathrm{y})$ the transformation functions of predictors (x) and response variable (y), respectively, and e is the model error. In this study, the only restriction applied to $\varphi(\mathrm{x})$ and $\varphi(\mathrm{y})$ was monotonicity, which makes it possible to distinguish between field soil morphological properties measured on an ordinal scale (having an intrinsic categorical order i.e., stickiness: from not sticky to very sticky) from nominal variables not having this restriction. Hue, structure type and texture class were the second type. Rewriting Equation (1) in terms of indicator matrices and category quantifications yields Equation (2):

$$\boldsymbol{G_r y_r} = \sum_{j=1}^{p} \beta_j \boldsymbol{G_j y_j} + \mathrm{e} \tag{2}$$

where β ($\beta_1, \ldots, \beta_p$), in order p (the number of predictor variables), is the vector containing regression coefficients, $\boldsymbol{y_j}$ and $\boldsymbol{y_r}$, in order k_j and k_r, the optimal scalings or numerical transformations of the categories for the predictors and response variable, respectively, and $\boldsymbol{G_j}$ and $\boldsymbol{G_r}$, in order $n \times k_j$ and n x k_r (where n is the number of cases), indicator matrices, such that 1 is when the i^{th} object is in the k_j category of variable j and 0 otherwise. CATREG estimates the regression coefficients by minimizing the least squares loss function (van der Kooij et al., 2006) in Equation (3):

$$\sigma\left(\boldsymbol{y}_r; \boldsymbol{\beta}; \boldsymbol{y}_j\right) = \left\| \boldsymbol{G_r y_r} - \sum_{j=1}^{p} \beta_j \boldsymbol{G_j y_j} \right\|^2 \tag{3}$$

The multiple correlation coefficient R^2 can be found from the ratio between the regression sum of squares and the total sum of squares (Gifi, 1990; van der Kooij, 2007) in Eq. (4):

$$R^2 = N^{-1/2} (\boldsymbol{G_r y_r})' \mathbf{v} (\mathbf{v}\mathbf{v}')^{-1/2} \tag{4}$$

where N is the number of observations and $\mathbf{v}$ is the accumulated contribution of predictor variables so that:

$$\mathbf{v} = \sum_{j=1}^{p} \beta_j \boldsymbol{G}_j \mathbf{y}_j \quad (5)$$

The statistics (*t*, *F* values) and fit and error measures, as well as the correlation matrices *R*, partial correlation and predictor tolerance, which will be used to assess the goodness of the model, are found from β_j and R^2. Tolerance is defined as one less the determination coefficient R^2 of the prediction of any predictor by the others, considering them independent variables, so it should be high to avoid multicollinearity.

2.5.2. Accuracy of categorical regression (CATREG)

The error term in the regression model ($1\text{-}R^2$) is usually not a good estimator of the prediction error, since it is found from the same data used to fit the model. To obtain a better estimate, resampling methods, such as cross validation or bootstrapping, may be used.

In CATREG, the goal is to minimize the dataset prediction error. The expected prediction error (EPE) must be known to be able to assess the reliability of predictions of future observations (van der Kooij, 2007). Therefore, ideally, a training dataset would be used to estimate a model and a test dataset. However, in this case, there is no test dataset, and therefore two resampling methods were used: K-Fold Cross-Validation (CV) and the 0.632 Bootstrap, using the mean square error (MSE) for assessing the EPE.

In K-Fold Cross-Validation, data are divided into randomized k groups of approximately the same size. K-1 groups are used to train the model and another is used for validation. This process is repeated k times using a different group as validation in each iteration. The process generates k estimates of the error which is averaged as the final estimate (Meulman et al., 2019). In bootstrapping, N cases of the same size as the sample are taken at random from the full dataset in each resampling to obtain B bootstrapped samples. Then, a model is fitted for each bootstrapped sample, estimating its prediction error from the original (not resampled) dataset. Contrary to CV, random sampling is performed with replacement, that is, an observation can be repeated in the bootstrapped sample, while another is excluded (*out-of-bag* set). The simple bootstrap estimate of the expected prediction error is found by averaging the error estimates of the B bootstrapped samples. Since the bootstrapped (training set) and the original sample (validation set) have many observations in common, the bias seems to be greater than with CV and, in general, it performs worse. Efron (1983) proposed a modification of the bootstrap prediction error called the 0.632 method, which includes the bootstrapped model error estimates in the simple bootstrap in their own *out-of-bag* observations (not the original dataset). This corrects the bias and improves model performance (van der Kooij, 2007). Here, we used 10 folds for CV and 50 bootstrap samples in 0.632 bootstrapping. In addition to the reliability of the prediction estimated by the MSE, a more flexible approach can be used to evaluate the CATREG model accuracy (Hartmann et al., 2009). For this, the dependent variable (PI, which usually ranged from −4 to + 4) was distributed into two classes or groups: N, negative with PI < 0 and P, positive with PI > 0, using zero as the cut-off point. Thus, model accuracy could be addressed by such classification accuracy measures as Efficiency and the Receiver Operating Characteristic (ROC) curve (Garosi et al., 2019). Efficiency can therefore be defined as:

$$Efficiency = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP (true positive) and TN (true negative) represent the number of correctly classified P and N plots, respectively, FP (false positive) the number of N plots that have been classified as P and FN (false negative) those P considered N.

The ROC curve plots the sensitivity (the ratio of the number of correctly classified P to total observed P) versus "1 – specificity" (specificity is the ratio between the number of correctly classified N and the total observed N). A model's predictive performance is high if high sensitivity is obtained at low values of "1 – specificity", that is, good capacity for correctly classifying P with a low number of false positives. This yields a curve closer to the upper left-hand corner (Garosi et al., 2019). The Area under the ROC curve (AUC) quantifies this relationship, so that a model is considered acceptable if AUC ≥ 0.7, excellent if AUC ≥ 0.8 and outstanding if AUC ≥ 0.9.

2.5.3. Regularization of CATREG

Regularization is used for selecting the model and avoiding overfitting in predictive techniques, since estimation of the regression coefficients by least squares may present collinearity (James et al., 2013). It is especially useful in categorical regression (Meulman et al., 2019). Of the three main regularization techniques, Ridge, Lasso and Elastic Net, Lasso is one of the most widely employed for optimal scaling regression. As first developed by Tibshirani (1996), Lasso regularization can deal with complex models with many predictors and high multicollinearity, when ordinary least squares show instability and overfitting. Both effects (instability and overfitting) are common in optimal scaling regression and can make it difficult to select a satisfactory theoretical model (Hartmann et al., 2009).

Lasso applies a λ penalty to the CATREG loss function that reduces the estimated regression coefficients, shrinking them to zero as the penalty increases (van der Kooij, 2007) in Eq. (7).

$$\sigma^{lasso}(\boldsymbol{\beta}) = \left\| \boldsymbol{G}_y \mathbf{y}_r - \sum_{j=1}^{p} \beta_j \boldsymbol{G}_j \mathbf{y}_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

Predictors with the most stable estimate of the coefficient shrink to zero more slowly, so Lasso regularization can be used to advantage for exploratory analysis instead of stepwise procedures to obtain a set of predictors with low multicollinearity. The advantage of Lasso over alternatives like Ridge regularization is that it shrinks the coefficients to zero, which the others do not. By cancelling coefficients, model interpretation (coefficient selection) becomes straightforward. Lastly, the Elastic Net regularization combines the ridge penalty rule and Lasso, but involves a complicated calculation, and its interpretation is not as immediate as Lasso. For details of the mathematics of these three regularization techniques in optimal scaling regression, see Meulman et al. (2019).

Since each λ penalty involves a regression model, the model that the regularized regression coefficients will be based on must be selected. We followed the procedure specified in Meulman et al. (2019) for this, selecting the most parsimonious model with the lowest prediction error. As an increase in the penalty usually yields lower prediction errors, the best model would be the one in which most of the coefficients are cancelled out, that is, when the sum of the standardized regression coefficients is very close to zero. However, a local minimum might be selected, such that it keeps a suitable number of coefficients, enabling an adequate theoretical model to be established. The one standard error (*1-SE*) rule can be applied for this (James et al., 2013). We chose the regularized model with the lowest number of predictors (the more parsimonious model) within one standard error of the model with the lowest EPE (the optimal model). All statistical analyses, including the regularized regression, were performed with IBM SPSS 24 (IBM Corp., 2016).

3. Results

3.1. Morphological soil properties

Tables 2 and 3 show the values of morphological soil properties: texture, colour, structure, consistence, stickiness, plasticity, biological activity, roots, and reaction to HCl of representative profiles on the high and low productivity banana sites, respectively. The characteristics of high (Mollic Ustifluvents) and low (Fluventic Haplustolls) productivity sites on Plantation PL are similar, except for texture: High productivity soils are silty loam (Table 2), while low productivity soils are generally

Table 2
Values of morphological soil properties in the A horizons of a representative profile of each high productivity level of plantations: texture, color, structure, consistence when moist and dry, stickiness and plasticity.

Plantations	Horizon†	Depth (cm)	Texture	Color	Structure	Consistence				Biological activity	Root abundance	Root size	Reaction to HCl
						When Dry	When moist	Stickiness	plasticity				
PL	Ap	0–20	sil	2,5 Y 3/2	sbk/m/gm	dsh	mfr	wss	wps	ah	raf	rsm	rhs
SM	Ap	0–17	sil	2,5 Y 4/2	sbk/f/gw	dsh	mvfr	wss	wps	ah	ra	rsm	rhm
PZ	Ap	0–22	ls	2,5 Y 3/2	sbk/f/gw	dsh	mvfr	wso	wpo	ah	ram	rsvf	rhno
PZ	A1	22–44	ls	2,5 Y 4/2	sbk/f/gw	ds	mvfr	wss	wpo	am	raf	rsvf	rhno
CH	Ap	0–18	sil	2,5 Y 3/2	abk/m/gw	dsh	mfr	wss	wps	ah	ram	rsf	rhno
CH	A1	18–38	sil	2,5 Y 4/2	abk/m/gm	dsh	mfr	wss	wps	am	raf	rsm	rhw
BA	Ap	0–18	sic	2,5 Y 4/2	abk/m/gm	dsh	mfi	ws	wp	ah	ram	rsm	rhno
BA	AC	18–40	sil	2,5 Y 4/4	m/ns/ng	ds	mfr	wss	wps	am	raf	rsm	rhno
KA	Ap	0–24	sicl	2,5 Y 3/2	sbk/m/gm	dsh	mfi	ws	wp	ah	ram	rsm	rhw
KA	A/C	24–42	sicl	2,5 Y 4/2	m/ns/ng	ds	mfr	ws	wp	am	raf	rsf	rhw

† **S**oil Survey Staff (2014). Abbreviations: **Texture:** ls = sandy loam; cl = clay loam; s = sand; L = loam; lvfs = very fine sandy loam; sic = silty clay; sicl = silty clay loam; sil = silty loam; sc = sandy clay. **Structure: Size:** ns = Structureless**;** vf = very fine; f = fine; m = medium; c = coarse. **Grade:** ng = structureless; gw = weak; gm = moderate; gs = strong. **Type:** m = massive**;** gr = granular; sbk = subangular blocky; abk = agular blocky; pr = prismatic. **Consistence: when dry:** dl = loose; ds = soft; dsh = slightly hard; dh = hard. **When moist:** mvfr = very friable; mfr = friable; mfi = firm. **Stickiness (consistence when wet):** wso = non-sticky; wss = slightly sticky; ws = stick. **Plasticity (consistency when wet):** wpo = non-plastic; wps = slightly plastic; wp = plastic. **Root abundance**: raf = few roots; ram: many roots. **Root size**: rsvf: very fine; rsf: fine; rsm: medium. **Biological activity**: ah = high activity; am = medium activity. **Reaction to HCl**: rhno = no reaction; rhw = weak reaction; rhm = moderate reaction; rhs = strong reaction.

Table 3
Values of morphological soil properties in the A horizons of a representative profile of each low productivity level of plantations: texture, color, structure, consistence when moist and dry, stickiness and plasticity.

Site	Horizon†	Depth (cm)	Texture	Color	Structure	Consistence				Biological activity	Root abundance	Root size	Reaction to HCl
						When Dry	When moist	Stickiness	plasticity				
PL	Ap	0–28	sicl	2,5 Y 3/2	sbk/m/gm	dh	mfi	wss	wps	ah	raf	rsf	rhs
SM	Ap	0–18	sicl	2,5 Y 3/2	abk/m/gm	dsh	mfi	ws	wp	ah	raf	rsm	rhm
SM	A1	18–44	sicl	2,5 Y 4/2	sbk/m/gm	dsh	mfi	ws	wp	am	raf	rsm	rhm
PZ	Ap	0–18	sicl	2,5 Y 3/2	abk/m/gm	dsh	mfi	wss	wps	am	raf	rsf	rhno
CH	Ap	0–20	sil	2,5 Y 3/2	sbk/m/gw	dsh	mfr	wss	wps	am	ram	rsf	rhno
CH	A1	20–42	sil	2,5 Y 4/2	abk/m/gw	dsh	mfr	wss	wp	al	raf	rsm	rhno
BA	Ap	0–14	sicl	2,5 Y 4/2	abk/m/gm	dsh	mfi	ws	wp	am	raf	rsvf	rhno
KA	Ap	0–22	sicl	2,5 Y 4/2	sbk/m/gs	dsh	mvfi	ws	wp	ah	ram	rsm	rhno
KA	A/C	22–42	sicl	2,5 Y 4/4	sbk/f/gm	dsh	mfi	ws	wp	al	raf	rsf	rhw

† **S**oil Survey Staff (2014). Abbreviations: **Texture:** Ls = sandy loam; cl = clay loam; s = sand; L = loam; lvfs = very fine sandy loam; sic = silty clay; sicl = silty clay loam; sil = silty loam; sc = sandy clay. **Structure: Size:** ns = Structureless**;** vf = very fine; f = fine; m = medium; c = coarse. **Grade:** ng = structureless; gw = weak; gm = moderate; gs = strong. **Type:** m = massive**;** gr = granular; sbk = subangular blocky; abk = angular blocky; pr = prismatic. **Consistence: when dry:** dsh = slightly hard; dh = hard. **When moist:** mvfr = very friable; mfr = friable; mfi = firm; mvfi = very firm. **Stickiness (consistence when wet):** wss = slightly sticky; ws = sticky. **Plasticity (consistency when wet):** wps = slightly plastic; wp = plastic. **Root abundance**: nra = no root; raf = few; ram: many. **Root size**: nra = no root; rsvf: very fine; rsf: fine; rsm: medium. **Biological activity**: ah = high activity; am = medium activity; al = low activity. **Reaction to HCl**: rhno = no reaction; rhw = weak reaction; rhm = moderate reaction; rhs = strong reaction.

silty clay loam (Table 3), with the same very dark greyish brown colour (2.5 Y 3/2), and a blocky subangular structure, with moderate-to-medium particle size. At both sites vegetation growth is limited by the high calcium carbonate content from its parent material, which originated in Lake Valencia. At Plantation SM, high productivity soils (Fluventic Haplustepts) usually have high biological activity, and a moderate reaction to HCl, with weak structural stability and lithological discontinuity (Tables 2 and 3). However, low productivity soils (Fluventic Haplustolls) were limited by their susceptibility to sealing and compaction due to the high proportion of silt and fine sand (100 and

200 μm) and very fine sand (50 and 100 μm) (Table 3).

The high productivity soils (Fluventic Haplustolls) at Plantation PZ have geomorphological characteristics favouring high biological activity (Table 2). The low productivity (Fluventic Haplustepts) sites show an abrupt textural change coupled with heavier soil texture, which is a significant limitation for banana productivity (Table 3). At Plantation CH, friable consistence along with weakly adhesive and plastic characteristics favour high biological activity and the common presence of fine roots at both productivity sites (Table 2). In the low productivity soils, characteristics are related to the sharp textural change in depth and poor structure. Soils at both sites are classified as Fluventic Haplustolls (Table 3).

At Plantation BA in the State of Trujillo, the high productivity sites have dark greyish brown (2.5 Y 4/2) silty clay soils (Oxyaquic Ustifluvents). The consistence is firm when moist, adhesive and plastic with high biological activity, no reaction to HCl and abundant roots (Table 2). On the other hand, low productivity soils (Typic Ustifluvents) on this plantation also have certain limitations associated with poor drainage and the presence of a water table. They are generally unstructured with abrupt textural changes, and few very fine roots (Table 3). At Plantation KA, the high and low productivity soils are silty clay loam, predominantly Entisols (Oxyaquic Ustifluvents) (Table 2), while in lower productivity soils the limitations are associated with poor soil structure (Table 3).

3.2. Model selection

From the original set of 15 predictors, we obtained a significant model with a high coefficient of determination, $R^2 = 0.746$ ($p < 0.007$). However, as most of the regression coefficients in the that first model was not significant (Table 4), it had to be refined. The coefficients of texture, structure type, dry consistence, biological activity and reaction to HCl were significant ($p < 0.05$). These variables should be considered in a stepwise procedure since they respond to soil processes related to plant physiology and eventual yield. Therefore, to fit a better model with significant coefficients while keeping R^2 as high as possible, we checked the MSE as a measure of the expected prediction error (EPE) given by alternative models in which predictors were progressively removed (Fig. 1). Thus, the optimum model included five predictors (Fig. 1). Table 5 shows the expected prediction errors in the nonregularized and regularized models. Considering the average EPEs of both bootstrap and cross-validation estimations, the most accurate models seem to be Elastic Net (average EPE of 0.676) and Lasso (average EPE of 0.677), because they had the lowest EPE. We used the Lasso model to select the optimal set of predictors since it is simpler and easier to interpret, and in addition, there are no appreciable differences in accuracy from the Elastic Net model.

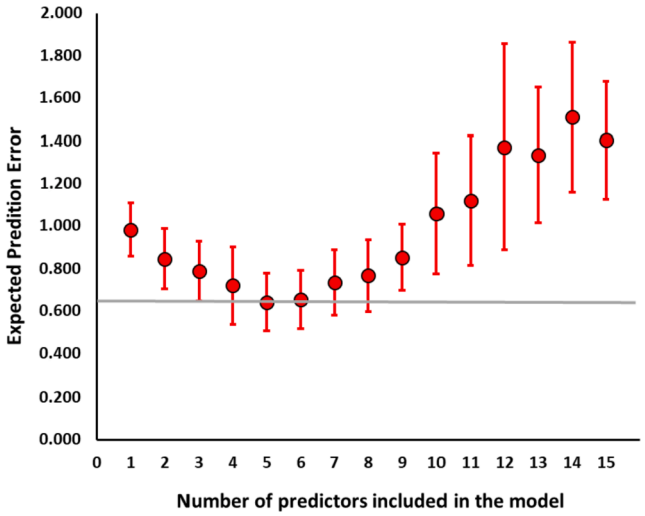

Fig. 1. Expected Prediction Error (EPE) of 15 different no-regularized models, employing the 0.630 Bootstrapping procedure (standard errors are showed by the vertical bars. The grey line shows the lowest EPE).

Table 5
Expected Prediction Error (EPE) and Standard Error (SE) of regularized CATREG models for the intial set of fifteen predictors.

Regularization	0.632 Bootstrapping		10-fold Cross Validation	
	EPE	SE	EPE	SE
No regularized	1.405	0.277	1.398	0.334
Ridge	0.792	0.125	0.836	0.116
Lasso	0.861	0.104	0.493	0.087
Elastic Net	0.859	0.104	0.493	0.087

The Lasso paths (Fig. 2) drawn represent the 42 regularized models performed by increasing λ (Equation (7)) by 0.02 per step, from a standardized sum of coefficients of 1 (unshrunk coefficients, right side of the graph) to 0 (left side of the graph). The variables with the regression coefficients that are shrunken earliest, lowering their penalized coefficients to zero, should be taken as those with the lowest predictive power, regardless of their starting value (right of the plot), so they can be removed from the model. On the contrary, coefficients that keep non-zero values at 1-SE of the regularized optimal model may be left in the theoretical model. Our optimum model included eight predictors with a λ of 0.260, an EPE of 0.789 and a SE of 0.122, while the most

Table 4
Regression coefficients, correlation with the outcome and tolerance of the modelfrom the initial set of fifteen predictors.

Predictors	Coefficients					Correlation with the outcome		Tolerance	
	Beta	B*	gl	*F*	*p-value*	*r*	Partial *r*	After	Before
Texture	0.397	0.224	5	3.127	**0.023**	0.118	0.556	0.723	0.704
Moist hue	0.193	0.195	2	0.980	0.388	−0.017	0.275	0.557	0.617
Moist value	0.175	0.203	1	0.742	0.396	0.049	0.272	0.663	0.641
Moist chroma	0.245	0.277	2	0.780	0.468	−0.092	0.298	0.414	0.411
Structure type	0.409	0.216	3	3.598	**0.026**	0.275	0.575	0.748	0.263
Structure size	−0.005	0.397	1	0.000	0.989	−0.251	−0.007	0.397	0.148
Structure grade	0.281	0.547	1	0.264	0.612	−0.298	0.299	0.316	0.151
Dry consistence	−0.612	0.234	2	6.829	**0.004**	−0.297	−0.645	0.483	0.628
Moist consistence	−0.515	0.349	3	2.178	0.113	−0.253	−0.508	0.334	0.312
Stickiness	−0.609	0.462	2	1.735	0.195	−0.193	−0.469	0.193	0.189
Plasticity	0.430	0.383	2	1.256	0.300	−0.186	0.334	0.172	0.189
Biological activity	0.792	0.311	2	6.480	**0.005**	0.273	0.688	0.364	0.221
Root abundance	−0.112	0.452	1	0.061	0.807	0.172	−0.137	0.392	0.192
Root size	0.052	0.409	3	0.016	0.997	0.119	0.072	0.493	0.370
Reaction to HCl	−0.532	0.240	2	4.898	**0.015**	−0.252	−0.645	0.638	0.398

* Bootstrap SE estimates of Beta coefficients (1,000 boostrap resamplings)

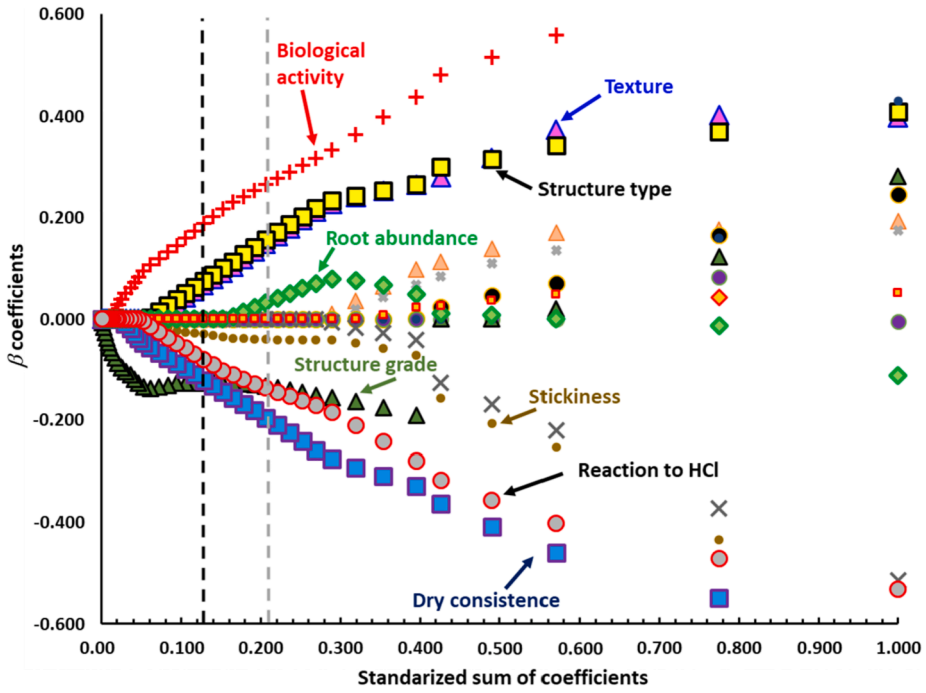

Fig. 2. Categorical regression (CATREG)-Lasso coefficients path estimated with the 0.632 bootstrap (dashed bar shows the most parsimonious –black vertical line – and the optimal –grey vertical line– models within one standard error). The coefficients in the right side of the plot (standardized sum of coefficients of 1.000) are the same of that in Table 4.

parsimonious, with an EPE of 0.861 (λ of 0.400) was fitted using seven predictors. In this case, there was no important difference in simplicity (parsimony) between the two regularized models. Only the standardized coefficient of root abundance was reduced to zero in the most parsimonious model with respect to the optimum. The eight predictors for the optimum model were: 1) biological activity, 2) dry consistency, 3) texture, 4) structure grade, 5) structure type, 6) HCl reaction, 7) stickiness, and 8) root abundance. From this starting set of predictors (Model 4 in Table 6), alternative nonregularized models of seven (Model 3), six (Model 2) and five (Model 1) predictors were tested in a backward stepwise regression to find the statistically significant model with the highest R^2 and minimum prediction error (Table 6). Model 1 was finally selected, because it met all these conditions: statistical significance of R^2 ($p < 0.0001$) and of all its coefficients ($p < 0.050$), while yielding the minimum bootstrapped EPE (0.645, see also Fig. 1) and a good enough R^2 (0.599) close to the model with the highest determination coefficient (Model 4, $R^2 = 0.645$).

The statistical significance of the regression coefficients was the main issue with these alternative models, because they all yielded quite similar prediction errors (within one SE) and significant overall fit ($p < 0.0001$). In Model 4, with eight predictors, elimination of root abundance was supported by the *1-SE* rule, as discussed above, while stickiness and structure grade presented coefficients far from statistical significance ($p = 0.195$ and 0.612, respectively) in the starting model (Table 4), as well as lower tolerances of all predictors (except plasticity). Therefore, Model 1 with five predictors seemed to be the best model for predicting the PI, and it was chosen as the definitive CATREG optimal scaling model (Table 7).

According to the sign of the coefficients, the ordinal variable, biological activity, is positively correlated with the dependent variable, PI, which implies that higher biological activity is correlated with higher productivity, while harder (drier consistence) soils with carbonates (reaction to HCl) are associated with lower productivity. The transformation functions (optimal scaling) indicate the linearity of this relationship, as well as the direction of the variation of the nominal variables, texture and structure (Fig. 3).

3.3. Optimal scaling of the field soil morphological variables

Fig. 3 shows optimal scaling of the selected categorical regression model. Of the transformation functions, dry consistence and biological activity were practically linear, which implies a proportional increase

Table 6
Selection of the optimal CATREG model.

Model	Number of predictors	R^2	*p*-value	EPE (SE) [1]		Predictors (*p*-value for regression coefficients)[2]
				Bootstrapped	10-fold crossvalidation	
#1	5	0.599	<0.0001	0.645 (0.135)	0.662 (0.236)	Bio. Act. (<0.0001), Texture (<0.0001), Dry cons. (0.001), Str. Type (0.000), HCl (0.017)
#2	6	0.633	<0.0001	0.656 (0.145)	0.651 (0.134)	Bio. Act. (<0.0001), Texture (0.001), Dry cons. (0.001), Str. Type (0.001), HCl (0.013), Str. Grade (0.065)
#3	7	0.636	<0.0001	0.738 (0.115)	0.685 (0.139)	Bio. Act. (<0.0001), Texture (0.008), Dry cons. (0.001), Str. Type (0.003), HCl (0.013), Str. Grade (0.449), Stickiness (0.838)
#4	8	0.645	<0.0001	0.770 (0.169)	0.778 (0.153)	Bio. Act. (0.044), Texture (0.006), Dry cons. (0.001), Str. Type (0.002), HCl (0.182), Stickiness (0.818), Str. Grade (0.478), Root abundance (0.657)

[1] EPE: expected prediction error; SE: standard error (in bracket).
[2] Bio. Act.: biological activity; Str. Grade: structure grade; Str. Type: structure type; Dry cons.: dry consistence; HCl: reaction to HCl.

Table 7
Regression coefficients, correlation with the outcome and Tolerance of the optimal CATREG model.

5 predictors (Model #1)	Beta	B*	*p*-value	Correlation with outcome		Tolerance	
				r	Partial *r*	After	Before
Texture	0.371	0.157	<0.0001	0.221	0.495	0.945	0.963
Structure type	0.463	0.154	<0.0001	0.309	0.571	0.904	0.856
Dry consistence	−0.440	0.151	0.001	−0.232	−0.551	0.902	0.793
Biological activity	0.581	0.108	<0.0001	0.292	0.645	0.849	0.907
Reaction to HCl	−0.336	0.160	0.017	−0.302	−0.463	0.968	0.976

* Bootstrap SE estimate of Beta coefficients (1,000 samples)

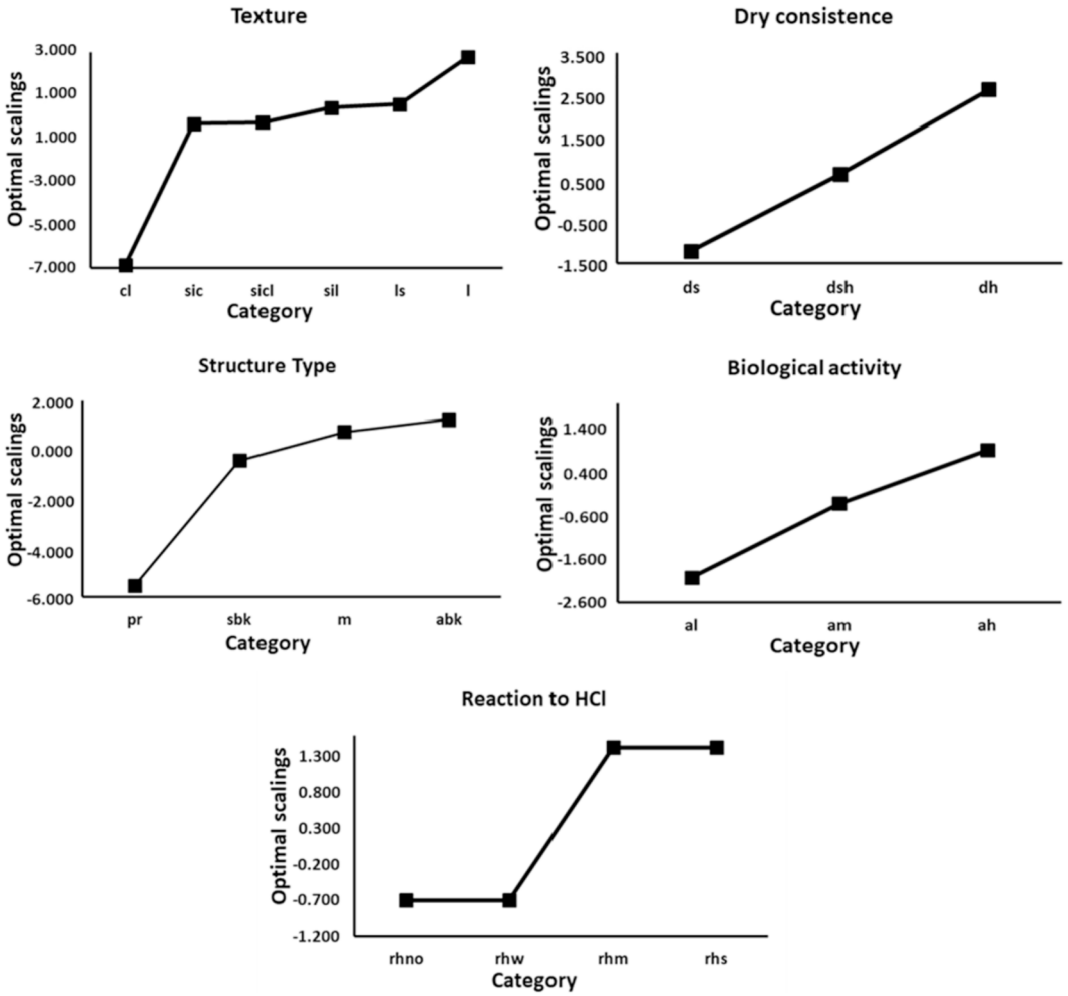

Fig. 3. Optimal scaling's of texture, dry consistence, structure type, biological activity, and reaction to HCl. Abbreviations: Texture: ls = sandy loam; cl = clay loam; l = loam; sic = silty clay; sicl = silty clay loam; sil = silty loam. Dry consistence: ds = soft; dsh = slightly hard; dh = hard. Structure type: m = massive; pr = prismatic; sbk = subangular blocky; abk = angular blocky. Biological activity: al = low activity; am = medium activity; ah = high activity. Reaction to HCl: rhno = no reaction; rhw = weak reaction; rhm = moderate reaction; rhs = strong reaction.

(biological activity) or decrease (dry consistency) in productivity (PI) with these variables. Two soil productivity groups were formed according to the acid reaction to their carbonate content: 1) no reaction (rhno) and weak reaction (rhwo), and 2) moderate (rhm) or strong (rhs) reaction. As this predictor has a negative correlation (beta coefficient) with PI, moderate or strong reactions must be associated with a decrease in productivity.

For structure type, categories were ordered (entered in the model as nominal, and therefore, with unrestricted order) as prismatic (pr) < subangular blocks (sbk) < massive (m) < angular blocks (abk). As the beta coefficient was positive, a positive correlation was defined between more developed (angular) structures and productivity, while prismatic structure, "extreme" development of the angular structure, was as unfavorable as massive structure, which could be explained by its characteristic loss of porosity (macro porosity). Texture categories, also entered as nominal, were ordered on a rough gradient from finer clay (cl) to loamy textures (l), and as the beta coefficient was positive, this was the most closely correlated with productivity (PI). (Fig. 3).

3.4. Prediction accuracy

Fig. 4 shows the prediction plot of observations versus predictions. As mentioned above, the CATREG model with the lowest bootstrap estimate of the prediction error was selected. However, it yielded a

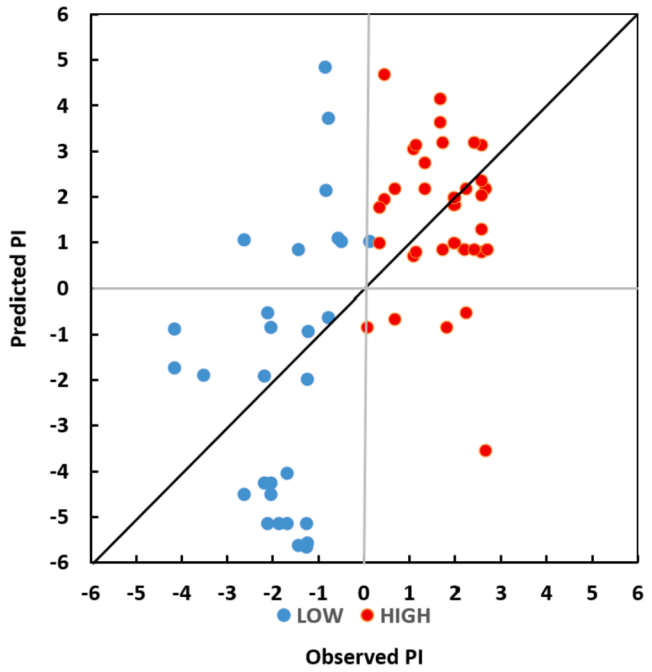

Fig. 4. Observed vs. Predicted values of the productivity index (PI) for the five predictors optimal categorical regression model. The colour of each point shows whether the plot belongs to a High (red) or Low (blue) productivity plantation. Grey lines: cut-off value for dependent variable classification in P (PI > 0) and N (PI < 0) groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

relatively high MSE (EPE) of 0.645 (RMSE of 0.803), accounting for 11.7% of the outcome ranking, which varies from −4.178 to 2.680 (X-axis in Fig. 4).

For a more general assessment of model performance, the observed and predicted dependent variables were classified into two groups or classes, with a PI of zero as the cut-off point: one level, N, with PI < 0 (low PI) and another with PI > 0 (high PI). It should be mentioned that most of the horizons of soils in high productivity sites (Table 1) also had positive predicted PI values. Table 8 shows the confusion matrix of the resulting classification, and various measures of their discrimination accuracy.

Under these conditions, the model showed high sensitivity of 86%, and misclassified only five cases recognized as positive or high PI (bottom right-hand quadrant in Fig. 4). Specificity, although slightly lower (75%), was also satisfactory, as was classification efficiency (81%) and the area under the ROC curve (AUC), which was 0.834, considered excellent. Therefore, although the prediction based on continuous PI values is not as reliable as might be desired, prediction of the level of productivity of the plot from the morphological indicators would be quite suitable. Finally, as deduced from Fig. 4, the statistical relationship between predicted PI and *a priori* plot type, High or Low productivity, was also significant: The mean predicted PI for High productivity was 1.579 (SD = 3.018), and mean predicted PI for Low productivity was −1.872 (SD = 1.642), t = -5.530 (p = 0.000).

Table 8
Confusion matrix and discrimination accuracy of classified values of the outcome variable: Sensitivity, Specificity, Efficiency and Area under the curve ROC (AUC).

		Predicted PI level		Total observed
		P	N	
Observed PI level	**P**	30	5	35
	N	7	21	28
Total predicted		37	26	63

P: positive PI (>0), N: negative PI (<0)
Sensitivity = 0.86, Specificity = 0.75, Efficiency = 0.81, AUC = 0.834

4. Discussion

The objective of this study was to validate the hypothesis that a quantitative relationship between banana productivity and key soil morphological properties is feasible. This was demonstrated using categorical regression analysis with transformation (Table 7 and Fig. 3). Categorical regression with optimal scaling was implemented to find nonlinear transformations (Lasso) to select a sparse model with stable predictors and bootstrap 0.632 to evaluate prediction accuracy. With this approach we identified a subset of five variables that best predicts banana productivity levels in two areas of Venezuela. These variables included texture, soil structure type, dry consistence, biological activity indicators and HCl reaction. The model developed enables biophysical interpretation, clearly related to banana productivity. The categorical regression developed was able to correctly discriminate between areas of high and low productivity on the same plantation, and captured the trend in variation in productivity among plantations as their soil morphological variables changed (Fig. 4). The accuracy of this model is in line with the prediction accuracy found by van der Kooij (2007).

An interesting feature of the regression model developed in our study is that, combined with the optimal scaling developed, it can be easily interpreted by banana production technicians in relation to some key soil properties. In addition, the key soil variables used can usually be found in soil profile descriptions, and with little explanation can also be interpreted by growers or other stakeholders. Texture, structure (grade and type) and biological activity are closely related to productivity. Soil texture influences the availability of water and nutrients, as well as aeration, drainage and accessibility in the use of agricultural implements. The model calibrated identified this, with a higher score in productivity for the loamy textures. This is in line with observations at Plantations PL, BA and KA, where the high productivity plots had loamy textures (L/L-SL) and low productivity plots had a moderate to fine texture in the A horizon. In the largest plantations, the high productivity plots were in the alluvial plains (BA, KA, SM, PZ and CH) which had a higher clay content than the lower productivity plots on the same plantations. Overall, this agrees with Vaquero (2005) who concluded that the areas with low banana productivity had soil horizons with coarse textures (very sandy soils) and very fine textures (clayey, with content of clay>60%) due to the direct influence on the water retention capacity and permeability. In this regard, medium-to-fine textured soils (loam to silty loam) with good structure and porosity, developed a deeper, more extensive root system (Vaquero, 2005; Rey, 2009; Delgado et al., 2010b). Higher biological activity, an indicator of healthy soil, was also positively correlated to higher productivity as indicated by the regression coefficient and the scaling ranking. This is not only an important technical result, but also for spreading good agricultural practices among banana stakeholders, since it shows a clear link between soils with good biological activity and higher productivity.

The fourth variable which contributed to our predictive model of banana productivity was the type of soil structure (Table 7 and Fig. 3). This is not surprising, since it reflects the process of soil formation and anthropogenesis (Hernández et al., 2010). Voorhees et al. (1971) stated that soil structure is a fundamental function in pedogenesis and for plant nutrition, because of its enormous significance in improving fertility and regulating the microbial activity of soils. In our analysis, soils with massive or highly developed prismatic structure had a lower score in productivity. This agrees with the results of Gauggel et al. (2005), who found rapid deterioration of the banana root system in coarse and very coarse blocks and prismatic structures, as is the case of the low productivity soils on the plantations located in the State of Trujillo (BA and KA). It also agrees with the results of Gauggel et al. (2005) and Villarreal-Núñez et al. (2013), who noted deterioration of the root system where soil with massive clay structures at shallow depths forms barriers.

Dry consistency was the most significant variable associated with the PI (Table 7), where consistence was strong in lower productivity soil (Table 7 and Fig. 3). In alluvial soils, firm and very firm consistence with weak or no structure can cause compaction, which according to Dorel (1993), lowers banana productivity. The results of Vaquero (2005), who found that banana root density was higher in soils with friable to very friable consistence, low bulk density and low resistance to penetration, also coincide with ours. In our study, root density was drastically reduced in areas with high penetration resistance associated with a firm or very firm soil consistence and bulk density over 1.2 g cm^{-3} (Data not shown).

High carbonate content, indicated by the reaction to HCl, the fifth parameter in our model (Table 7 and Fig. 3), can also decrease banana productivity (Cigales & Pérez, 2011). This is associated with the limited response to fertilization of very calcareous soils, which can even prevent bunch development and reduce the size of the pseudo stem and plant height, and sometimes facilitate the appearance of diseases (Olivares et al., 2020). Phosphorus, iron, zinc, and nitrogen deficiencies can be explained by excessive presence of carbonates. When carbonate accumulates at a certain depth in the soil profile, the apical bud can die, even after normal initial development (Vanilarasu and Balakrishnamurthy, 2014; El-Khawaga, 2013).

Overall, our study has shown how the use of categorical regression analysis with optimal scaling can deliver an operating model able to incorporate the effect of qualitative soil information into banana productivity. When properly scaled to other soil types and farms, it has the potential of being a useful tool for farmers, technicians or investors for identifying the best areas for banana plantations. It can also contribute to independent management in different areas within the same plantation based on relatively easily acquired soil information. One of the major advantages of this model is that it is based on relatively simple field evaluations at a moderate cost, and soil information from field surveys carried out previously for other purposes can be used.

5. Conclusions

The five morphological properties of the soil (soil texture, soil structure type, dry consistence, biological activity and HCl reaction) in our empirical categorical regression model have a clear agronomic relationship with banana productivity. The proposed model could be used in the field for reliable identification of areas of high and low potential banana productivity in other banana growing areas such as, the states of Barinas, Sucre and Zulia in Venezuela after local assessment. Identification of the main soil morphological properties associated with banana productivity by applying categorical regression can contribute to the long-term sustainability of banana soils in Venezuela, and other tropical areas.

Our results suggest the potential for further studies of quantitative transformation of soil morphological properties and application of categorical regression, as carried out in this study, with different levels of banana productivity This methodology can be easily applied to other crops, requiring little or no expert knowledge.

This study can serve as an example of a relatively straight forward way to quantitatively assess the effect of soil properties on banana productivity using information generated from soil surveys which are relatively inexpensive and often already available for other reasons. Calibrating similar correlations between banana productivity indicators, or even actual yield records, with soil morphological properties in specific areas, using the methodology proposed in this manuscript could be done in a few months at relatively moderate cost, which would be compensated by the savings in planting the banana plots in the most suitable areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors recognize the financial support for international mobility of the Ibero-American Secretary General with Fundación Carolina and Action KA107 of Erasmus + Program from Agrifood Campus of International Excellence (ceiA3) (2020). Also, by project "Technological innovations for the management and improvement of the quality and health of banana soils in Latin America and the Caribbean" financed by FONTAGRO and coordinated by Bioversity International (before INIBAP) and project SHui (European Commission Grant Agreement number: 773903). Funding for open access charge: Universidad de Córdoba / CBUA.

References

Andrews, S.S., Karlen, D.L., Cambardella, C.A., 2004. The soil management assessment framework: A quantitative soil quality evaluation method. Soil. Sci. Soc. Am. J. 68 (6), 1945–1962. https://doi.org/10.2136/sssaj2004.1945.

Calero, J., Serrano J.M., Aranda V., Sánchez D., Vila M.A., Delgado G. 2005. Analysis and characterization of olive tree cultivation system in Granada province (South of Spain) with optimal scaling and multivariate techniques. Agrochimica, 49, 118-131. Available online at https://n9.cl/k3rdk. Last accessed 07 July 2020.

Calero, J., Aranda, V., Montejo-Raez, A., Martin-Garcia, J.M., 2018. A new soil quality index based on morpho-pedological indicators as a site-specific web service applied to olive groves in the Province of Jaen (South Spain). Comput. Electron Agric. 146, 66–76. https://doi.org/10.1016/j.compag.2018.01.016.

Calero, J., Delgado, R., Delgado, G., Martín-García, J.M., 2008. Transformation of categorical field soil morphological properties into numerical properties for the study of chronosequences. Geoderma 145 (3-4), 278–287. https://doi.org/10.1016/j.geoderma.2008.03.022.

Cigales, M., Pérez, O., 2011. Soil variability and water requirement of banana cultivation in a locality in the Pacific of Mexico. Av. Investig. Agropecu, 15, 21-31. Available online at https://n9.cl/hxjtg. Last accessed 07 July 2020.

Delgado, E. Trejos, J, Villalobos, M, Martinez, G., Lobo, D., Rey, J., Rodríguez, G., Rosales, F. and Pocasangre, L. 2010b. Determination of a soil quality and health index for banana plantations in Venezuela. Interciencia, 35, 927-933. Available online at https://n9.cl/5ly9u. Last accessed 07 February 2021.

Delgado, E., Rosales, F., Trejos, J., Villalobos, M., Pocasangre, L., 2010a. Soil quality and health index for banana plantations in four countries of Latin America and the Caribbean. Bioagro, 22, 53-60. Available online at https://n9.cl/qm3e. Last accessed 07 February 2021.

Dorel, M., 1993. Banana development in an andosol in Guadeloupe: effect of soil compaction. Fruits 48, 83–88. Last accessed 07 February 2021.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Amer. Statist. Assoc 78, 316–331. https://doi.org/10.2307/2288636.

El-Khawaga, A.S., 2013. Response of Grand Naine banana plants grown under different soil moisture levels to antitranspirants application. Asian J. Crop Sci 5 (3), 238–250. https://doi.org/10.3923/ajcs.2013.238.250.

FAO, 2006. Guidelines for soil profile description. Fourth edition. FAO, Roma.

FAO, 2020. Banana Market Review and Banana Statistics 2018-20. Rome. Available online at. Last accessed 07 February 2021. http://www.fao.org/economic/est/est-commodities/bananas/en/.

Garosi, Y., Sheklabadi, M., Conoscenti, C., Pourghasemi, H.R., Van Oost, K., 2019. Assessing the performance of GIS- based machine learning models with different accuracy measures for determining susceptibility to gully erosion. Sci. Total Environ. 664, 1117–1132. https://doi.org/10.1016/j.scitotenv.2019.02.093.

Gauggel, C.A., Sierra, F., Arévalo, G., 2005. The problem of banana root deterioration and its impact on production: Latin America's experience. In Turner, D.W. and Rosales, F.E. (Eds) Banana root system: towards a better understanding for its productive management. Proceedings of an International Symposium held in San José, Costa Rica on 3-5 November 2003. International Network for the Improvement of Banana and Plantain (INIBAP), Montpellier, pp. 13-22. Available online at https://n9.cl/b3wi8. Last accessed 07 February 2021.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chichester.

González García, H., González Pedraza, A. F., Atencio, J., & Soto, A. 2021a. Evaluation of quality of banana soils through microbial activity in the south the lake of Maracaibo, Zulia state, Venezuela. Rev. Fac. Agron. (LUZ), 38(2), 216-240. Available online at https://n9.cl/wdnr5. Last accessed 19 February 2021.

González García, H.G., Pedraza, A.F.G., Yzquierdo, G.R., Pacheco, R.L., Vásquez, M.B., 2021b. Vigor of plantain plants (*Musa* AAB cv. Harton) and its relationship with physical, chemical and biological characteristics of the soil. Agron. Costarricense 45 (2), 115–134. https://doi.org/10.15517/rac.v45i2.47772.

González García, H.G., González, A.F., Pineda, M., Escalante, H., Yzquierdo, G.A.R., Bracho, A.S., 2021c. Edaphic microbiota in plantain lots of contrasting vigor and relationships with soil properties. Bioagro 33 (2), 143–148. https://doi.org/10.51372/bioagro332.8.

González-Pedraza, A.F., Atencio, J., Cubillán, K., Almendrales, R., Ramírez, L., Barrios, O., 2014. Microbial activity in soils cultivated with plantain (Musa AAB

plantain subgroup cv. Harton) with different vigor of plants. Rev. Fac. Agron. (LUZ) 31, 526–538. https://doi.org/10.51372/bioagro332.8.
González-Pedraza, A. F., Carlos Escalante, J. 2021. Nitrogen mineralization in soils cultivated with plantain (Musa AAB Subgroup plátano cv. Hartón), Zulia state, Venezuela. Rev. Fac. Agron. (LUZ), 38(3), 525-547. Available online at https://n9.cl/so6u. Last accessed 19 July 2021.
Hair, J.F., Anderson, R.E., Tatham, R.L., BlLack, W.C., 1999. Análisis Multivariante, fifth ed. Prentice, Hall, Madrid. España.
Hartmann, A., Zeeck, A., van der Kooij, A.J., 2009. Severity of Bulimia Nervosa. Measurement and classification into health or pathology. Psychopathology 42, 22–31. https://doi.org/10.1159/000173700.
Hernández, A., Bojorquez, J., Morell, F., Cabrera, A., Ascanio, M., Garcia, J., Madueño, A., Najera, O., 2010. Fundamentos de la estructura de suelos tropicales. Universidad Autónoma de Nayarit, Mexico.
Hernández, Y., Marín, M., García, J., 2007. Response to the plantain (Musa AAB cv. Horn) yield as a function of the mineral nutrients and its phenologycal cycle. Part I. Growth and Production. Rev. Fac. Agron. (LUZ), 24, 607-626. Available online at https://n9.cl/uwuxm. Last accessed 19 February 2021.
IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer 112. https://doi.org/10.1007/978-1-4614-7138-7.
Lal, R., 1998. Soil quality and agricultural sustainability. In: Lal, R. (Ed.), Soil quality and agricultural sustainability. Sleeping Bear Press Inc, Chelsea, MI, pp. 3–12.
Linting, M., Meulman, J.J., Groenen, P.J.F., van der Kooij, A.J., 2007. Nonlinear principal components analysis: Introduction and application. Psychological Methods 12, 336–358. https://doi.org/10.1037/1082-989X.12.3.336.
MacEwan, R. J., Fitzpatrick, R. W., 1996. The Pedological Context for Assessment of Soil Quality. In: MacEwan, R. J., Carter, M. R. (Eds.), Soil quality is in the hands of the land manager. Proceedings of an international symposium. Advances in soil quality for land management: Science, Practice and Policy, 17-19 April 1996, University of Ballarat, Ballarat, Victoria, Australia, pp. 10–16.
Martínez, G., Delgado, E., Rodríguez, D., Hernández, J., Del Valle, R., 2008. Brief Analysis of Musaceae Production in Venezuela. Prod. Agrop, 1, 24-29. Available online at https://n9.cl/d48mt. Last accessed 19 February 2021.
Meulman, J.J., van der Kooij, A.J., Babinec, A., 2002. New features of categorical principal components analysis for complicated data sets, including data mining. In: Gaul, W., Ritter, G. (Eds.), Classification, Automation. and New Media, pp. 207–217. https://doi.org/10.1007/978-3-642-55991-4_22.
Meulman, J.J., van der Kooij, A.J., Duisters, K.L.W., 2019. ROS Regression: Integrating Regularization with Optimal Scaling Regression. Statistical Sci. 34, 361–390. https://doi.org/10.1214/19-STS697.
Munsell Color Company, 1999. Munsell Soil Color Charts. Munsell Color, Macbeth Division of Kollmorgen, Maryland, USA.
Olivares, B., 2018. Tropical conditions of seasonal rain in the dry-land agriculture of Carabobo, Venezuela. Lgr 27, 86–102. https://doi.org/10.17163/lgr.n27.2018.07.
Olivares, B., Cortez, A., Parra, R., Lobo, D., Rodríguez, M.F., Rey, J.C., 2017. Evaluation of agricultural vulnerability to drought weather in different locations of Venezuela. Rev. Fac. Agron. (LUZ), 34,103-129. Available online at https://n9.cl/hc5xs. Last accessed 19 February 2021.
Olivares, B., Paredes, F., Rey, J., Lobo, D., Galvis-Causil, S., 2021. The relationship between the normalized difference vegetation index, rainfall, and potential evapotranspiration in a banana plantation of Venezuela. STJSSA 18, 58–64. https://doi.org/10.20961/stjssa.v18i1.50379.
Olivares, B.O., Araya-Alman, M., Acevedo-Opazo, C., Rey, J.C., Cañete, P., Giannini, F., Balzarini, M., Lobo, D., Navas-Cortés, J.A., Landa, B.B., Gómez, J.A., 2020. Relationship between soil properties and banana productivity in the two main cultivation areas in Venezuela. J. Soil Sci. Plant Nutr. 20 (4), 2512–2524. https://doi.org/10.1007/s42729-020-00317-8.
Pulido-Moncada, M., Gabriels, D., Lobo, D., Rey, J.C., Cornelis, W.M., 2014. Visual field assessment of soil structural quality in tropical soils. Soil Tillage Res. 139, 8–18. https://doi.org/10.1016/j.still.2014.01.002.
Pulido-Moncada, M., Penning, L.H., Timm, L.C., Gabriels, D., Cornelis, W.M., 2017. Visual examination of changes in soil structural quality due to land use. Soil Tillage Res. 173, 83–91. https://doi.org/10.1016/j.still.2016.08.011.
Rey, J.C. Martínez G., Rodríguez G., Lobo D., Delgado E. Trejos J., Pocasangre L. Rosales F., 2009. Aspects on soil quality and welfare in Venezuela bananas. Prod. Agrop., 2, 52-55. Available online at https://n9.cl/5qmzy. Last accessed 19 February 2021.
Rodríguez, G., Núñez, MC., Lobo, D., Martínez. G., Rey, JC., Espinoza, J., Muñoz, N., González D., Rosales, F., Pocasangre, L., Delgado, E., 2006. Banana root health in lots with different productivity levels in a soil at the oriental coast of Maracaibo Lake, Venezuela. XVII Reunión Internacional ACORBAT: Banano un negocio sustentable. Joinville, Santa Catarina, Brasil. Nov 15–20. p. 355.
Rosales, F.E., Pocasangre, L.E., Trejos, J., Serrano, E., Peña, W., 2008. Guía de Diagnóstico de la Calidad y Salud de Suelos bananeros. France Bioversity Int. Last accessed 19 February 2021.
Sevinç, G., Aydogdu, M.H., Cancelik, M., Sevinc, M.R., 2019. Farmers' Attitudes toward Public Support Policy for Sustainable Agriculture in GAP-Sanliurfa, Turkey. Sustainability 11, 6617. https://doi.org/10.3390/su11236617.
Soil Survey Staff., 2017. Soil Survey Manual Hadbook 18. United States Department of Agriculture. Washington D.C.
Soil Survey Staff., 2014. Keys to Soil Taxonomy. Keys to Soil Taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B, 58, 267–288. Available online at https://n9.cl/pqv2k. Last accessed 19 February 2021.
Van der Kooij, A. J., 2007. Prediction accuracy and stability of regression with optimal scaling transformations. Unpublished doctoral dissertation, Leiden University. Netherlands. Available online at https://n9.cl/ew8pz. Last accessed 19 February 2021.
van der Kooij, A.J., Meulman, J.J., Heiser, W.J., 2006. Local minima in categorical multiple regression. Comput. Statistics Data Anal. 50 (2), 446–462. https://doi.org/10.1016/j.csda.2004.08.009.
Vanilarasu, K., Balakrishnamurthy, G., 2014. Influences of organic manures and amendments in soil physiochemical properties and their impact on growth, yield and nutrient uptake of banana. The Bioscan, 9, 525-529. Available online at https://n9.cl/awlm. Last accessed 19 February 2021.
Vaquero, MR., 2005. Soil physical properties and banana root growth. In: Turner, D.W and Rosales, F.E (Eds) Banana root system: towards a better understanding for its productive management: proceedings of an International Symposium held in San José, Costa Rica on 3-5 November 2003. International Network for the Improvement of Banana and Plantain, Montpellier, pp 125–131. Available online at https://n9.cl/szt4m. Last accessed 19 February 2021.
Vasu, Duraisamy, Tiwari, Gopal, Sahoo, Sonalika, Dash, Benukantha, Jangir, Abhishek, Sharma, Ram Prasad, Naitam, Ravindra, Tiwary, Pramod, Karthikeyan, Karunakaran, Chandran, Padikkal, 2021. A minimum data set of soil morphological properties for quantifying soil quality in coastal agroecosystems. Catena 198, 105042. https://doi.org/10.1016/j.catena.2020.105042.
Vaughan, I.P., Ormerod, S.J., 2005. Increasing the value of principal components analysis for simplifying ecological data: a case study with rivers and river birds. J. Appl. Ecol, 42, 487–497. Available online at https://n9.cl/2wnp0. Last accessed 19 February 2021.
Villarreal-Núñez, J., et al., 2013. Soil quality indexes in areas cultivated with banana in Panama. Agron. Mesoam 24, 301–315. Last accessed 19 February 2021.
Voorhees, W.B., Amemiya, M., Allmaras, R.R., y Larson, W.E., 1971. Some Effects of Aggregate Structure Heterogeneity on Root Growth. Soil Sci. Soc. Am. J 35 (4), 638–643. https://doi.org/10.2136/sssaj1971.03615995003500040043x.
Wissmann, M., Shalabh, D., Toutenburg, H., 2011. Role of categorical variables in multicollinearity in linear regression model. J. Appl. Stat 19, 99–115. Last accessed 19 February 2021.
Xu, J., Capretz, L.F., Ho, D., 2010. Building an OSS Quality Estimation Model with CATREG. International J Comput Sci Eng 2, 1952–1958. Last accessed 19 February 2021.