







Determining the geographical origin of lettuce with data mining applied to micronutrients and soil properties

Camila Maione¹, Eloá Moura Araujo², Sabrina Novaes dos Santos-Araujo², Alexys Giorgia Friol Boim², Rommel Melgaço Barbosa¹ *, Luís Reynaldo Ferraciu Alleoni²

¹Universidade Federal de Goiás/Instituto de Informática, Alameda Palmeiras, Quadra D, Campus Samambaia – 74690-900 – Goiânia, GO – Brasil.

²Universidade de São Paulo/ESALQ – Depto. de Ciência do Solo, C.P. 09 – 13418-900 – Piracicaba, SP – Brasil.

*Corresponding author <rmbweb@gmail.com>

Edited by: Thomas Kumke

Received March 19, 2020
Accepted August 31, 2020

ABSTRACT: Lettuce (*Lactuca sativa*) is the main leafy vegetable produced in Brazil. Since its production is widespread all over the country, lettuce traceability and quality assurance is hampered. In this study, we propose a new method to identify the geographical origin of Brazilian lettuce. The method uses a powerful data mining technique called support vector machines (SVM) applied to elemental composition and soil properties of samples analyzed. We investigated lettuce produced in São Paulo and Pernambuco, two states in the southeastern and northeastern regions in Brazil, respectively. We investigated efficiency of the SVM model by comparing its results with those achieved by traditional linear discriminant analysis (LDA). The SVM models outperformed the LDA models in the two scenarios investigated, achieving an average of 98 % prediction accuracy to discriminate lettuce from both states. A feature evaluation formula, called F-score, was used to measure the discriminative power of the variables analyzed. The soil exchangeable cation capacity, soil contents of low crystallized Al and Zn content in lettuce samples were the most relevant components for differentiation. Our results reinforce the potential of data mining and machine learning techniques to support traceability strategies and authentication of leafy vegetables.

Keywords: ICP-OES, traceability, tropical soils, heavy metals, feature selection

Introduction

Lettuce is among the most consumed vegetables worldwide and is considered the most produced and consumed leafy vegetable in Brazil. Lettuce is low in calories, fat, and Na, while being a good source of fibers, Fe, folate, vitamins and several other bioactive compounds that are beneficial to human health (Kim et al., 2016). Since the consumption *per capita* of fruits and vegetables is about 40 kg per yr⁻¹ in Brazil, much less than 143 kg per yr⁻¹ consumed in a developed country, such as the United States (Mainville and Peterson, 2005), lettuce is an important source of vegetable-based nutrients for the Brazilian population.

According to the most recent research conducted by IBGE (Brazilian Institute of Geography and Statistics) in 2006, the states of São Paulo (SP) and Pernambuco (PE) are the main lettuce producers in the southeastern and northeastern regions of Brazil, respectively. Growers can sell to a diversity of buyers, including intermediaries (purchase at farm gate), small supermarkets, large supermarket chains, wholesale markets, processors, and directly to consumers. This fragmented production chain makes the traceability and quality assurance of lettuce a difficult task. Moreover, most farmers neglect to use methods and techniques that add value to the product, such as food safety, traceability of inputs, improvements of handling and planting, among others (Carvalho et al., 2014).

This study has the following main objectives:

- i. We propose the use of support vector machines (SVM)

and feature selection to determine the geographical origin of lettuce samples based on their elemental composition and soil properties. We discriminate lettuce samples from São Paulo and Pernambuco, major lettuce producers in Brazil (Figure 1).

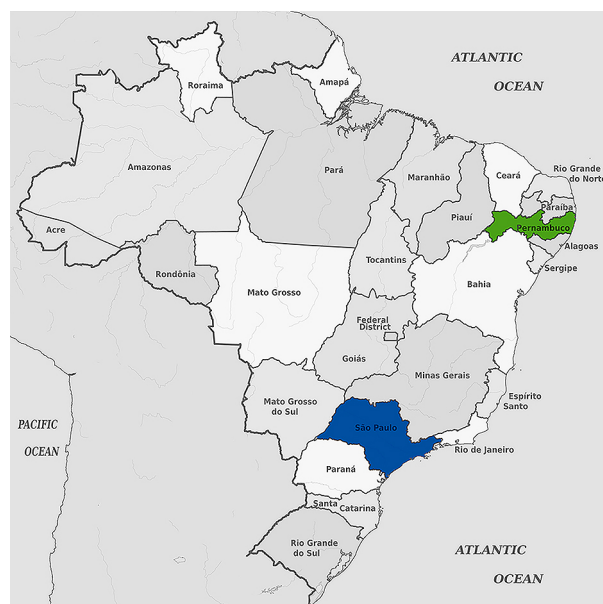


Figure 1 – Map of administrative divisions of Brazil. Pernambuco and São Paulo States are highlighted in green and blue colors, respectively.

ii. In order to ascertain the efficiency of the SVM model, we developed simple linear discriminant models for the same data and compared the results and performance measurements.

iii. We also investigated the discriminative power of each variable and identified a subset of variables that mostly impact differentiation through the use of a feature selection technique, called F-score, a novel approach to lettuce discrimination.

iv. We expect to show the potential of data mining and machine learning techniques to support traceability strategies and authentication of leafy vegetables.

Materials and Methods

Lettuce and soil samples analyzed

We collected 194 lettuce samples and soil samples from farms in São Paulo (n = 72) and Pernambuco (n = 122), Brazil. Coordinates of the sampling sites are shown in Table 1. Soil samples were dried in the shade and then sieved (2 mm mesh). Lettuce leaves were washed in running water to remove impurities, dried (45 – 60 °C) and ground in a stainless-steel mill (< 1 mm).

Soil texture was obtained by the densimeter method (Gee and Or, 2002) and the pH was obtained by potentiometry using a combined electrode immersed in the soil: water suspension (1:2.5) and soil: 1 mol L⁻¹ KCl solution (1:2.5). Potential acidity (H + Al) was obtained by extraction with 1 mol L⁻¹ calcium acetate (pH 7.0) and titration with NaOH using phenolphthalein as indicator. The organic carbon (OC) content of soils was obtained by dry combustion in an elemental analyzer. A 1 mol L⁻¹ KCl solution extracted the levels of exchangeable Ca, Mg and Al. Levels of available K and P were extracted with a double acid solution (Mehlich-1), following the protocol of Anderson and Ingram (1992). Based on these extractions, we obtained the following values: $\Delta\text{pH} = \text{pH}_{\text{KCl}} - \text{pH}_{\text{H}_2\text{O}}$; $\text{CEC}_T (\text{Ca}^{2+} + \text{Mg}^{2+} + \text{K}^+ + \text{H} + \text{Al})$; CEC_e

($\text{Ca}^{2+} + \text{Mg}^{2+} + \text{K}^+ + \text{Al}^{3+}$); SB ($\text{Ca}^{2+} + \text{Mg}^{2+} + \text{K}^+$); V % ($[\text{SB} \times 100]/\text{CEC}_T$); and m % ($[\text{Al}^{3+} \times 100]/\text{CEC}_e$). Levels of well-crystallized Fe and Al ($\text{Fe}_2\text{O}_3\text{DCB}$ and $\text{Al}_2\text{O}_3\text{DCB}$) were extracted with Na dithionite-citrate-bicarbonate (DCB) (Inda Junior and Kämpf, 2003; Mehra and Jackson, 1960), while amorphous Fe and Al were extracted with acidic ammonium oxalate ($\text{Fe}_2\text{O}_{3\text{OXA}}$ and $\text{Al}_2\text{O}_{3\text{OXA}}$) (McKeague and Day, 1966).

The pseudo-total concentrations of Cu, Ni, and Zn in the soil were extracted by acid extraction in a microwave oven using the EPA 3051A method (1:3 HCl:HNO₃, v/v). Plant material digestion followed Araújo et al. (2002), using HNO₃ and H₂O₂ in microwave assisted digestion. Contents of Cu, Ni and Zn were determined by inductively coupled plasma / optical emission spectroscopy (ICP OES) using the conventional sample introduction system. Data quality control was measured using standard reference material (SRM 2709a – San Joaquin Soil) from the National Institute of Standards and Technology (NIST, USA) and an analytical blank in triplicate. The concentrations of analytical blanks were below the quantification limit (0.01 mg L⁻¹ for Cu and Ni; 0.05 mg L⁻¹ for Zn). Precision (n = 3), expressed as relative standard deviation (RSD), was < 10 % for all elements. More details can be found at Santos-Araujo and Alleoni (2016).

Data mining for prediction of food origin

In the past decade, authenticity and traceability of foodstuffs became a desirable feature for consumers and producers worldwide (Baroni et al., 2015) and the search for methods that ensure authenticity of food has received great attention from researchers. A strategy that emerged in recent literature was the use of data mining and multivariate data analysis to discriminate the geographical origin of foodstuffs and vegetables based on their chemical components. Successful applications of this methodology and other similar were reported for rice (Maione et al., 2018), honey (Maione et al., 2019), Italian and Turkish lemon (Potorti et al., 2018),

Table 1 – Approximated geographical coordinates of cities where the analyzed lettuce samples were collected. Column # is the number of samples collected from the location.

City	State	#	Latitude	Longitude	Altitude (m)	City	State	#	Latitude	Longitude	Altitude (m)
Piracicaba	SP	4	-22.73	-47.62	1.276	Ibiúna	SP	1	-23.76	-47.46	1.683
Limeira	SP	3	-22.56	-47.34	1.311	São Roque	SP	1	-23.68	-47.23	1.590
Nova Odessa	SP	3	-22.79	-47.30	1.270	Sorocaba	SP	3	-23.62	-47.18	1.542
Jundiá	SP	2	-22.92	-47.00	1.388	Embú Guaçu	SP	2	-23.39	-47.56	1.232
Leme	SP	3	-23.15	-46.83	1.412	São José do Rio Pardo	SP	1	-23.86	-46.84	1.430
Mogi Mirim	SP	1	-22.41	-46.96	1.296	Pombos	PE	20	-21.68	-46.92	1.461
Itatiba	SP	2	-22.43	-46.95	1.303	Amaraji	PE	10	-8.38	-35.45	1.019
Salesópolis	SP	2	-23.06	-46.83	1.451	Vitória de St Antão	PE	50	-8.13	-35.30	885
Biritiba Mirim	SP	1	-23.55	-45.95	1.452	Recife	PE	10	-8.07	-34.94	745
Mogi das Cruzes	SP	3	-23.57	-46.08	1.430	Glória de Goitá	PE	49	-8.03	-35.28	896
Suzano	SP	3	-23.56	-46.15	1.463	Chã Grande	PE	19	-8.23	-35.46	1.195
Piedade	SP	1	-23.75	-47.46	1.658						

tea (Moreda-Piñeiro et al., 2003), chocolate (Cambrai et al., 2010), alcoholic beverages (Alcázar et al., 2012; Ceballos-Magaña et al., 2012; Coetzee et al., 2005), coffee (Oliveira et al., 2015; Serra et al., 2005), tomato (Mahne Opatić et al., 2018), and others. Therefore, we proposed the use of data mining techniques, namely SVM and feature selection algorithms, to determine the geographical origin of lettuce samples based on their chemical composition.

Data mining is an efficient process to find hidden patterns and information in large and complex data sets where simple multivariate data analysis techniques and statistical methods are often unable to model efficiently, such as the principal component analysis and the discriminant analysis. Data mining techniques combine concepts and methods from artificial intelligence, mathematical optimization, linear algebra, and statistical analysis in order to perform either predictive or exploratory analysis on labeled or unlabeled data (Kotsiantis et al., 2006). Although data mining processes emerged from the multivariate data analysis and statistical techniques to handle larger and more complex data sets (Izenman, 2008), these processes can be applied to smaller data sets to extract meaningful information, often preferred due to their sophisticated algorithms that are capable of performing probabilistic reasoning. Furthermore, these algorithms are constantly evolving.

Classification models can be described as data mining tools that can predict information, represented by a categorical variable, in data samples. These models observe similar and previously labeled samples and use the information learned from this observation to build a function or model that is capable of generalizing the learned information in new and unknown data samples, as long as they are described by the same set of variables as the observed samples. This learning process is known as supervised learning in the field of artificial intelligence. Support vector machines, created by Cortes and Vapnik (1995), are an example of a popular classification model in the recent data mining literature and are successfully employed to discriminate and classify data from different fields for various purposes.

Our previous literature search revealed that this is the first attempt to discriminate the geographical origin of Brazilian lettuce samples based on the machine learning technique for data mining, such as support vector machines, also applied to chemical composition and soil parameters. In order to ascertain the efficiency of the SVM model, we developed simple linear discriminant models for the same data and compared the results and performance measurements. We also investigated the discriminative power of each variable and identified a subset of variables that mostly impact differentiation through the use of a feature selection technique called F-score, a novel approach to the discrimination of lettuce.

Support vector machines (SVM)

SVM is described as an optimization function to find the decision boundary with the largest margin possible to separate the data, minimizing the risk of overfitting and improving the generalization performance. The decision boundary is computed by the following Eq. 1:

$$w \cdot x + b = 0 \quad (1)$$

where: x is the values obtained from the variables of the training samples, w refers to weights whose linear combination computes the class label, and b is a model parameter, the decision boundary with the largest margin possible is achieved by the minimization Eq. 2:

$$\min_w \frac{\|W\|^2}{2} \quad (2)$$

Classification models based on SVM are widely used in the literature to perform the predictive analyses on data from several problems and fields. Only in the last two years, SVM has been successfully employed to solve problems in domains, such as geology (García-Nieto et al., 2019; Huang et al., 2017; Jung et al., 2018; Kumar et al., 2017; Mahvash and Hezarkhani, 2018; Pu et al., 2019), hydrological sciences (Choubin et al., 2019b, 2018; Kisi et al., 2019; Rahmati et al., 2019; Sajedi-Hosseini et al., 2018), climate and weather (Fan et al., 2018; Kundu et al., 2017; Yu et al., 2018, 2017), fault detection in various systems and processes (Ali et al., 2018; Fazai et al., 2019; Ghalyani and Mazinan, 2019; Han et al., 2019; Liu and Zio, 2018; Manjural Islam and Kim, 2019; Saari et al., 2019; Xi et al., 2019), health and medicine (Battineni et al., 2019; Di et al., 2019; Liu et al., 2019; Lukmanto et al., 2019; Vougas et al., 2019), agriculture (Akbarzadeh et al., 2018; Feng et al., 2019; Fernandes et al., 2019; Griffel et al., 2018; Leena and Saju, 2019; Radhakrishnan and Ramanathan, 2018; Zhou et al., 2019) power and energy systems (Ma et al., 2018; Wang et al., 2018; Zendeheboudi et al., 2018), urban wastes and infrastructure (Karimi et al., 2019; Solano Meza et al., 2019; Tang et al., 2019; Xiao et al., 2019; Zhu et al., 2019), speech recognition (Bhavan et al., 2019; Braga et al., 2019; Rahmeni et al., 2019; Wang et al., 2019), and many others. In addition, we chose to work with the SVM models in this project due to two main advantages. First, the models are capable of performing kernel trick and project the data into higher dimensions to better classify non-linearly separable data such as ours. Second, because the SVM models are known to perform relatively better on small data sets in comparison to other machine learning algorithms, such as neural networks, which are heavily dependent on the amount of data available for training.

In this study, we employed SVM with the radial basis (RBF) kernel function. The use of kernel functions allows SVM to project the original data into a new dimensional space and find a linear decision boundary

to separate the transformed samples when they cannot be linearly separated in the original dimensional space. In addition to the required parameter C , which can be described as the cost imposed by the SVM model on a misclassification, the RBF kernel also requires a γ parameter, namely the value used by the kernel to perform the kernel trick and handle non-linear classification. Both parameters must be chosen carefully, since increasing their value indiscriminately potentially results in overfitting, high variance, and low biases, while very restrictive values lead to an under-fitted model that cannot capture patterns in the data. We determine these values through a grid search on values $C = \{0.25, 0.5, 1, 1.5, 2, 3\}$ and $\gamma = \{3, 2, 1, 0.5, 0.1, 0.01, 0.02, 0.03, 0.05, 0.06\}$ for each SVM model developed. The model with the best performance was selected.

Linear discriminant analysis (LDA)

The linear discriminant analysis (LDA) is a classification technique to maximize the ratio of between-class variance to within-class variance to achieve maximal separability. The LDA creates a decision boundary called discriminant function (DF), which is a linear combination of the variables that describe the data and that best separates the classes. Considering a problem for classes y_1 and y_2 , the linear DF is defined as Eq. 3 (Duda et al., 2001):

$$g(x) = w'V + w_0 \quad (3)$$

where: x is an arbitrary sample, V is the variable set values for sample x , w is the weight vector, and w_0 is a bias value. We aimed to find w and w_0 values for $g(x) > 0$, otherwise, the class label associated to x is y_1 , and y_2 .

The LDA has been widely used recently in several classification problems and, despite traditional, it is still a well-known and popular method to discriminate food data, largely reported in literature reviews (Abbas et al., 2018; Berrueta et al., 2007; Callao and Ruisánchez, 2018; Cavanna et al., 2018; Esteki et al., 2019, 2018a, 2018b; Granato et al., 2018; Jiménez-Carvelo et al., 2019; Kemsley et al., 2019; Medina et al., 2019a, 2019b; Oliveri, 2017; Peris and Escuder-Gilabert, 2016; Ropodi et al., 2016; Valdés et al., 2018; Wadood et al., 2020). Therefore, we expect this model to perform well in our data set and that its use certify the efficiency of the SVM model by comparing the results obtained by both methods.

Performance measures

The data available for analysis must be divided into a training set to build the classification model and as a test set to verify the model prediction performance, also called the holdout method.

The default holdout method has two main disadvantages. First, it requires the data set to be divided

into two subsets for training and testing the model, respectively. When the available data set is relatively small, similar to that analyzed in this study, dividing the data set can be unfeasible, as the resulting subsets can be too small to effectively train a reliable classification model (Varma and Simon, 2006). Moreover, since only a single subset is used for training the model and this subset is commonly generated by random selection, meaningful information possibly contained in data samples assigned to the test set is not considered and is thus wasted. In order to tackle these issues, we used a training and validation method called the k -fold cross validation, a solution to the lack of sufficiently large training and testing sets (Duda et al., 2001). This method divides the data set into k mutually exclusive subsets (folds) of similar size. The classification model is trained and tested for k iterations. In each iteration, one subset, different from the subsets previously used, is selected to test the model while the others are used for training. Therefore, all the data samples available for analysis are eventually considered in the construction of the classification model. The final accuracy of the model is computed as the average of the accuracies obtained in each iteration.

After the test phase, the tested samples can be categorized as true positives, true negatives, false positives or false negatives. True positives (TP) and true negatives (TN) are the number of positive and negative samples correctly classified, respectively. False positives (FP) refer to the number of negative samples incorrectly classified as positives and false negative (FN) is the number of positive samples incorrectly classified as negative. Performance measurements of accuracy, sensitivity and specificity (Tan et al., 2005) are computed based on these values. Accuracy refers to the overall probability of the model to correctly classify an arbitrary sample (Choubin et al., 2019a). Sensitivity refers to the overall probability of the model to correctly classify an arbitrary sample, which originally belongs to the positive class. Specificity is the overall probability of the model to correctly classify an arbitrary sample, which originally belongs to the negative class. Therefore, the three performance measurements are computed with Eq. 4–6:

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4)$$

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100 \quad (5)$$

$$Specificity(\%) = \frac{TN}{FP + TN} \times 100 \quad (6)$$

Estimating the relevance of the parameters analyzed

One of our objectives was to evaluate the discriminative power of the descriptive variables and try to build

classification models capable of discriminating lettuce from two distinct locations with high performance using only variables considered relevant for the decision-making of the classifier. Disregarding variables with low or null influence on the information mapped by the class label could also provide advantages, such as improvement of prediction accuracy, dimensionality reduction, reduction of time to build and run classification models, among others.

Filter methods are variable selection methods applied to the training data prior to the learning phase of the models, allowing irrelevant variables to be identified and discarded before training occurs. Filter methods evaluate variables by computing the intercorrelation between each other and the correlation between the variables and the class label. The best rated variables are present little dependence from other variables while presenting the highest dependence as possible from the class label. Since filter methods are algorithmically simple and operate with low computational cost, a common strategy is to use them to evaluate all the variables individually, to set up subsets with combinations of the best ranked variables, to apply them to a classification model and to check the final prediction accuracy obtained to attest their discriminative power. There are several examples of popular filter methods for the multivariate data analysis, such as information gain, chi-square, random forest importance (Izenman, 2008), mutual information, Correlation-based Feature Selection (CFS), and others (Bommert et al., 2020).

In this study, we used a variable selection algorithm called F-score. This function presented by Chen and Lin (2006) measures the discrimination of two sets of real numbers. For a single descriptive variable from our data set, we can divide its measurements into two distinct sets called positive and negative sets, which hold the variable measurements for lettuce samples from SP and PE, respectively. The value produced from this function, when applied to a variable to measure the discrimination between its positive and negative sets, can be used as a score for measuring the variable contribution to the class label. Given the training samples x_k , $k = \{1, \dots, m\}$, if the number of samples belonging to the SP and PE classes are n_{SP} and n_{PE} , respectively, the F-score value of the i -th variable, which reflects the discrimination between positive and negative samples, is calculated by the Eq. 7:

$$F(i) = \frac{(\bar{X}_i^{(SP)} - \bar{X}_i)^2 + (\bar{X}_i^{(PE)} - \bar{X}_i)^2}{\frac{1}{n_{SP} - 1} \sum_{k=1}^{n_{SP}} (X_{k,i}^{(SP)} - \bar{X}_i^{(SP)})^2 + \frac{1}{n_{PE} - 1} \sum_{k=1}^{n_{PE}} (X_{k,i}^{(PE)} - \bar{X}_i^{(PE)})^2} \quad (7)$$

where: \bar{X}_i , $\bar{X}_i^{(SP)}$, $\bar{X}_i^{(PE)}$ are the average of the i -th variable of the whole, positive, and negative data sets, respectively; $\bar{X}_{k,i}^{(SP)}$ is the i -th variable of the k -th positive sample, and $\bar{X}_{k,i}^{(PE)}$ is the i -th variable of the k -th negative sample. The numerator indicates the discrimination between the positive and negative sets, and the

denominator indicates the discrimination within each set. The larger the F-score, the more discriminative this variable.

Balancing the data set with the K-means clustering algorithm

Imbalanced data sets are inconvenient and present various challenges for data mining and the multivariate data analysis (Chawla, 2005; Haixiang et al., 2017; He and Garcia, 2009; Jo and Japkowicz, 2004; López et al., 2013; Prati et al., 2004). Overall, classification models trained on imbalanced data tend to express a good prediction performance for samples of the majority class and a lower performance for samples of the minority class. This decrease occurs not necessarily due to the difference in the class proportion, but due to other natural factors of imbalanced data, such as the presence of small disjuncts, low density of data, data overlapping and others (He and Garcia, 2009; López et al., 2013). In this study, we tackled the imbalanced data issue with the aid of a clustering algorithm called K-means.

Clustering algorithms are considered a branch of unsupervised learning and are basically employed in exploratory data analyses, where no hypothesis about the data nor previously known class labels existed. These techniques are useful to aid the identification of natural groupings existing within the data based on a similarity (or dissimilarity) pattern. Partitional clustering algorithms, such as K-means, divide the data set into mutually exclusive clusters in a way that samples assigned to a same cluster must be as similar as possible and as different as possible from samples associated to other clusters.

The K-means algorithm could be summarized in the following steps (Jain, 2010): (i) randomly selects k data samples and names their centroids. Each centroid is associated to a different cluster label; (ii) for each non-centroid sample in the data set, it find its nearest centroid and associates this sample to the same cluster as the centroid found; (iii) for each cluster formed, it updates the centroid to be the center of cluster mass; and (iv) repeats the previous steps until no new changes are made to the clusters, or a stopping criterion is reached.

The K-means algorithm is not new and is still highly reported in the literature due to its simplicity, low computational cost, and good performance (Jain, 2010). In this study, we used the K-means algorithm to aid data balancing due to under sampling. Considering that we want to discard m samples of a certain class from the data set, we divide the data labeled as this class into $(n - m)$ clusters with the K-means algorithm and keep only the determined centroids as data samples. Because the centroids found for each cluster could be considered the most representative samples in a data partition, this strategy reduces the information loss that naturally occurs under sampling.

Analysis strategy

The entire statistical and predictive analysis was conducted on the RStudio software, version 1.1.463. Our analysis methodology is presented in Figure 2 and summarized in the following steps:

1. Data samples obtained from Pernambuco (PE) state are under sampled in order to match the number of available samples from São Paulo (SP) state. The goal of this step is to create a balanced data set that could be reliably used to develop classification models without biases. The K-means clustering algorithm is applied only to the samples obtained from PE and 36 clusters were determined. The centroids computed for each cluster were retained in the data set, while the other samples from PE were discarded from analysis. Therefore, the final data set comprised all 36 originally samples collected in SP and 36 samples from PE retained as centroids from the clustering algorithm.
2. In order to perform five-fold cross validation, the balanced data set is randomly divided into five mutually exclusive subsets (folds), properly keeping the original proportion of the two states (50 % - 50 %) in each set.
3. For each validation:
 - a. The selected validation is used as a test set while the remaining folds are merged and used as a training set;
 - b. The training and test sets are standardized to avoid potential biases caused by the different unit

measurements and ranges of values of the variables. The variables are centered by subtracting their means from their values, and then the centered variables are divided by their standard deviations;

- c. F-score values are computed for each variable of the training set. The selection threshold is set as the maximum F-score value obtained by the variables divided by 3;
- d. The SVM and LDA models are developed using the entire training data and the training data with only the variables that received F-score values higher than the threshold, resulting in a total of four models developed;
- e. Accuracy, sensitivity, and specificity values are computed for the four models.

4. The average accuracy, sensitivity, and specificity obtained by the models are determined and presented as final performance measurements.

Results and Discussion

Properties and micronutrients in lettuce and soil samples

Metal uptake by plants is influenced by several soil properties (Kumpiene et al., 2017). Therefore, we evaluated 25 soil variables (Table 2). We designated letters to represent the variables analyzed in lettuce and soil samples to simplify visualization in our graphics and

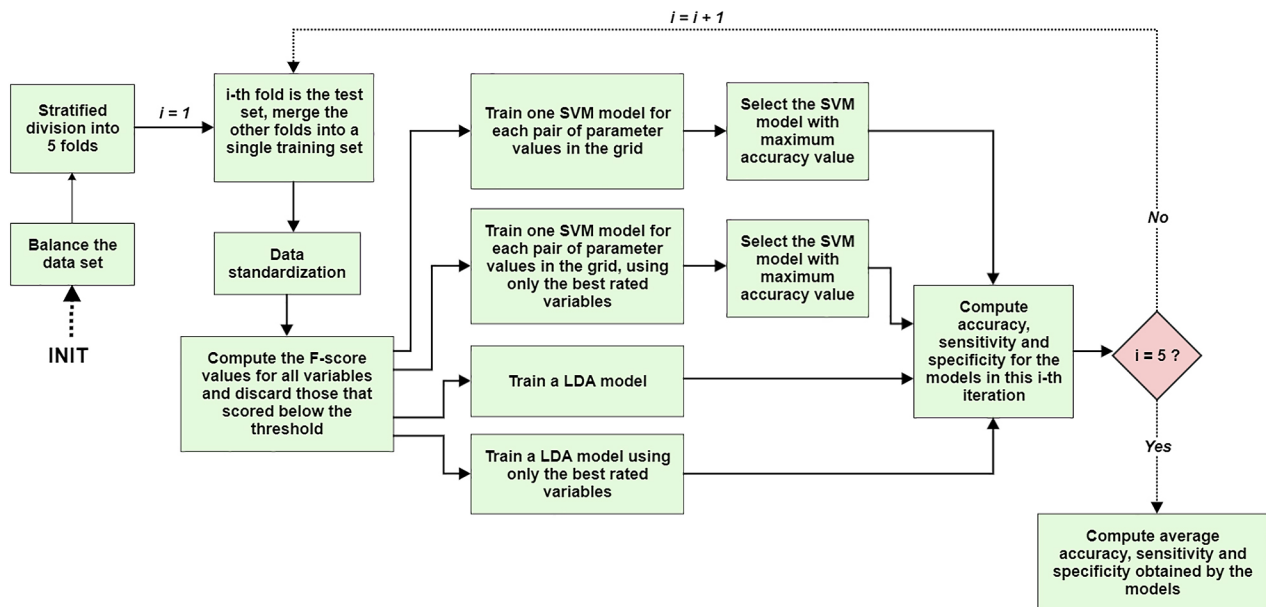


Figure 2 – Methodology for construction of the SVM (support vector machines) and LDA (linear discriminant analysis) models and performance measure estimation.

Table 2 – Analyzed variables in the determination of the geographical origin of lettuce samples and their respective mean and standard deviation values for each state, São Paulo (SP) and Pernambuco (PE).

ID	Description	Mean ± SD		ID	Description	Mean ± SD	
		SP	PE			SP	PE
Var1	pH determined in water suspension	5.84 ± 0.74	6.52 ± 0.75	Var15	Pseudototal concentration of Ni in soil (mg kg ⁻¹)	9.45 ± 6.52	2.95 ± 3.02
Var2	pH determined in KCl solution	5.51 ± 0.72	5.98 ± 0.7	Var16	Pseudototal concentration of Cu in soil (mg kg ⁻¹)	41.34 ± 41.61	14.43 ± 11.34
Var3	Potential acidity (mmol _c dm ⁻³)	24.47 ± 16.1	6670.18 ± 10955.68	Var17	Pseudototal concentration of Zn in soil (mg kg ⁻¹)	72.81 ± 87.01	38.43 ± 50.33
Var4	Levels of sand (g kg ⁻¹)	461.17 ± 155.72	658.06 ± 193.23	Var18	Amorphous Al extracted with acid ammonium oxalate (g kg ⁻¹)	14.66 ± 11.56	1.21 ± 0.84
Var5	Levels of clay (g kg ⁻¹)	366.9 ± 142	204.64 ± 164.03	Var19	Amorphous Fe extracted with acid ammonium oxalate (g kg ⁻¹)	4.82 ± 4.93	1.28 ± 1.04
Var6	Levels of silt (g kg ⁻¹)	171.93 ± 87.4	137.29 ± 84.47	Var20	Levels of well-crystallized iron extracted with sodium dithionite–citrate–bicarbonate (g kg ⁻¹)	22.13 ± 20.2	7.94 ± 6.7
Var7	Levels of exchangeable calcium (mmol _c dm ⁻³)	57.65 ± 52.6	630.19 ± 910.64	Var21	Levels of well-crystallized aluminum extracted with sodium dithionite–citrate–bicarbonate (g kg ⁻¹)	10.77 ± 5.3	5.04 ± 3.85
Var8	Levels of exchangeable magnesium (mmol _c dm ⁻³)	17.1 ± 8.99	623.59 ± 1113.78	Var22	ΔpH obtained by pH _{KCl} – pH _{H2O}	-0.33 ± 0.26	-0.54 ± 0.18
Var9	Levels of exchangeable aluminum (mmol _c dm ⁻³)	1.13 ± 1.97	0.61 ± 0.4	Var23	CEC _T obtained by Ca ²⁺ + Mg ²⁺ + K ⁺ + (H+Al) (mmol _c dm ⁻³)	104 ± 55.79	29.17 ± 16.57
Var10	Levels of available phosphorus extracted with a double acid solution (mg kg ⁻¹)	530.82 ± 348.03	311.67 ± 239.87	Var24	CEC _e obtained by (Ca ²⁺ + Mg ²⁺ + K ⁺ + Al ³⁺) (mmol _c dm ⁻³)	80.67 ± 57.32	17.68 ± 6.12
Var11	Levels of available potassium extracted with a double acid solution (mmol _c dm ⁻³)	4.78 ± 2.76	4.5 ± 2.1	Var25	Base saturation obtained by ((SB x 100)/CEC _T) (%)	73.15 ± 18.83	68.39 ± 26.38
Var12	Concentrations of Zn in plant (mg kg ⁻¹)	170.18 ± 110.78	49.63 ± 23.36	Var26	Aluminum saturation obtained by ((Al ³⁺ x 100)/CEC _T) (%)	2.31 ± 5.53	3.71 ± 2.84
Var13	Concentrations of Cu in plant (mg kg ⁻¹)	6.79 ± 1.82	8.81 ± 3.17	Var27	Sum of bases obtained by (Ca ²⁺ + Mg ²⁺ + K ⁺) (mmol _c dm ⁻³)	79.53 ± 57.73	17.07 ± 6.1
Var14	Concentrations of Ni in plant (mg kg ⁻¹)	0.32 ± 0.19	0.29 ± 0.4	Var28	Organic carbon content in soil (%)	27.04 ± 32.71	17.62 ± 12.58

specific mentions throughout the rest of the paper. Soil samples collected from lettuce farms in SP had fairly higher amount of nutrients and other beneficial traits than soil samples obtained from PE.

Soil samples from SP farms presented mean values for pseudo-total concentrations of Ni, Cu and Zn of 9.45, 41.34 and 72.81 mg kg⁻¹, respectively, while samples from PE presented 2.95, 14.43 and 38.43 mg kg⁻¹, respectively. Ni, Cu, and Zn are essential metals for plants. Biondi et al. (2011) demonstrated that soils from PE have low capacity to release Cu and Ni to plants. Their study also indicated a significant association of most metals to clayey soils. Considering that the soil samples from PE are mostly composed of sand (approximately 70 %), the low clay content of these soils may help explain the relatively low metal concentration in the pseudototal fraction of soils from PE.

Amorphous and crystalline Al and Fe oxide minerals play a major role in stabilizing soil structure,

and their presence in soils has a favorable effect on soil physical properties (Goldberg, 1989). Furthermore, kaolinite, Fe and Al oxides compose the dominant mineralogy in the clay fraction of most Brazilian soils (Fink et al., 2014), responsible for chemical reactions that control the availability of essential and non-essential elements.

Phytoavailable metal forms are sorbed to amorphous metal oxides (Rodrigues et al., 2010). Levels of well-crystallized Fe and Al were higher in soil samples from SP than for those from PE, presenting mean values of 22.13 g kg⁻¹ and 10.77 g kg⁻¹ respectively, against 7.94 g kg⁻¹ and 5.04 g kg⁻¹ respectively for soils from PE. Soil samples from SP also showed fairly higher numbers of amorphous Al than soil samples from PE, with mean values of 14.66 mg kg⁻¹ and 1.21 mg kg⁻¹ for soils of both states, respectively. As for amorphous Fe, mean values were 4.82 mg kg⁻¹ and 1.28 mg kg⁻¹ for soils from SP and PE, respectively.

Mean levels of available P in soil samples from SP and PE states were 530.82 mg kg⁻¹ and 311.67 mg kg⁻¹, respectively. Although samples from PE presented higher levels of exchangeable Ca and Mg, the soil samples from SP state showed higher CEC (for both CEC_T and CEC_e), which depends on the levels of Ca, Mg, K, Al and potential acidity in the soil samples. Mean values of CEC_T were 104 mmol_c dm⁻³ and 29.17 mmol_c dm⁻³ for soils from SP and PE states, respectively, while mean values of CEC_e were 80.67 mmol_c dm⁻³ and 17.68 mmol_c dm⁻³ for soils from SP and PE states, respectively. Finally, values for the sum of bases were also considerably higher in soils from SP (mean 79.53 mmol_c dm⁻³) than in soils from PE (mean 17.07 mmol_c dm⁻³). Maybe soils in SP farms are better fertilized than in PE.

Statistical and predictive analysis

Using the standardized measurements of all variables shown in Table 2 as input values, the SVM models reached an average value of 98.67 % for accuracy, 97.14 % for sensitivity, and 100 % for specificity. The LDA models trained with the same input values performed fairly lower than SVM for all performance measurements, presenting 66 % average accuracy, 71.43 % average sensitivity, and 60.71 % average specificity. Although more complex to build than linear discriminant models and designed to handle large and complex data bases, SVM models are excellent tools to determine the geographical origin of lettuce, even when trained on a relatively small amount of data.

In order to determine the individual importance of each component for the discrimination of the lettuce samples from both regions, we applied the F-score equation to each training set during the cross-validation process. The F-score values achieved by each variable

in each iteration are presented in descending order in Figure 3. The variables referring to the sum of bases obtained by (Ca²⁺ + Mg²⁺ + K⁺) and soil cation exchangeable capability (CEC_T and CEC_e) were retained in all training sets with very high F-score values. Levels of exchangeable Ca, well-crystallized Al, amorphous Al, sand and pseudo-total concentration of Ni measured from the soil plus the Zn levels in the plant were retained by four of five considered training sets. Overall, we conclude that these eight factors were the most relevant variables for the discrimination of lettuce samples from both states according to the F-score metric.

To ascertain the real relevance of the best rated components, we also built SVM and LDA models using only values for variables that achieved F-score values higher than the determined threshold (Table 3), while the others were discarded. The average performance measurements, namely accuracy, sensitivity and specificity, achieved by the models with and without variable selection are summarized in Table 4. The SVM model trained with only the best rated variables achieved 95.81 % average accuracy, 94.29 % average sensitivity and 97.14 % average specificity, a very small decrease in performance in comparison to the values achieved when all 28 variables were used for training. On the other hand, the LDA model experienced a slight decrease in performance, presenting 86.29 % average accuracy, 94.29 % average sensitivity, and 78.57 % average specificity when only the five best rated variables were used for training. Although the SVM and LDA models achieved the same average sensitivity values when combined with variable selection, the SVM models still clearly outperformed the LDA models in both scenarios for all other performance measures.

The best classification model achieved was the SVM model trained on all variables available from the data set. The accuracy value achieved is approximately

Table 3 – Variables retained by the F-score in the training set in each iteration of the cross-validation process. Variables discarded achieved F-score value below the computed threshold *max*/3.

CV iteration	Variables retained by the F-score in the training set
# 1	Var1, Var22, Var7, Var27, Var23, Var24, Var20, Var21, Var18, Var4, Var15, Var13, Var12
# 2	Var7, Var27, Var23, Var24, Var21, Var18, Var5, Var4, Var15, Var12
# 3	Var27, Var23, Var24
# 4	Var1, Var7, Var27, Var23, Var24, Var20, Var21, Var19, Var18, Var5, Var4, Var15, Var13, Var12
# 5	Var1, Var22, Var7, Var27, Var23, Var24, Var21, Var18, Var5, Var4, Var15, Var12

Table 4 – Averaged performance measures computed for the SVM (support vector machines) and LDA (linear discriminant analysis) models trained with the whole variable set and with only the best rated variables according to the F-score.

Performance measure	SVM model		LDA model	
	All variables	Best rated variables	All variables	Best rated variables
	%			
Accuracy	98.67	95.81	66	86.29
Sensitivity	97.14	94.29	71.43	94.29
Specificity	100	97.14	60.71	78.57

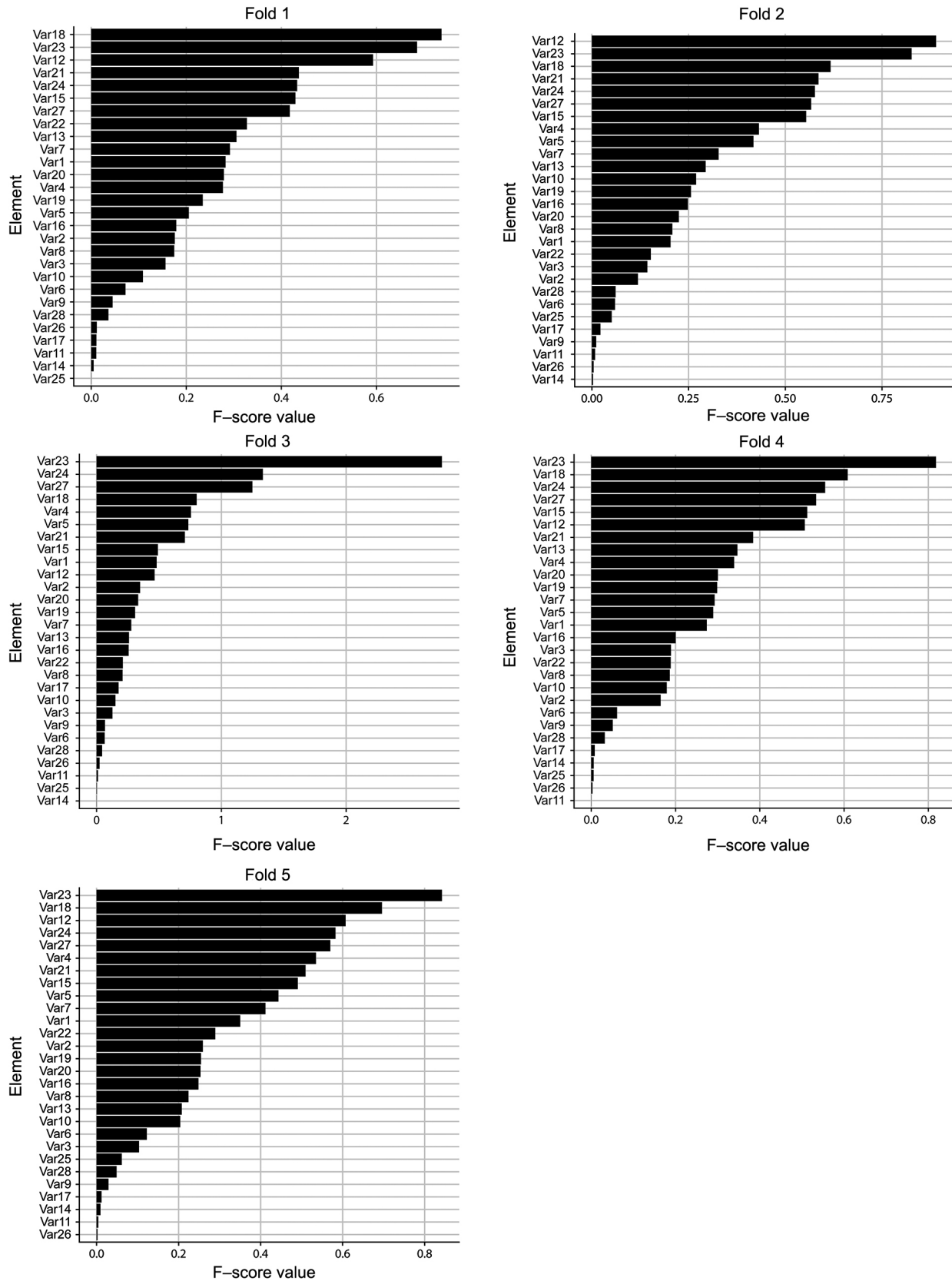


Figure 3 – Relative importance of each variable to determine the lettuce geographical origin, computed according to the F-score equation.

3 %, 32 % and 12 % higher than the accuracy values obtained from the SVM model with feature selection, the LDA model without feature selection and the LDA model with feature selection, respectively. This slight increase in error for the SVM model is expected since several variables were discarded, the model was possibly deprived from meaningful information contained in them; however, the SVM model with variable selection still presented a high average accuracy value for predicting the geographical origin of lettuce when only approximately eight out of 28 variables were used. This is a substantial decrease in the dimensionality and size of the data set, and consequently reduction of the required effort from researchers to gather and prepare the necessary data.

Mean values of the sum of bases, CEC_T , CEC_e , exchangeable Ca, well-crystallized Al, sand, amorphous Al, nickel in soil and Zn in plant for the lettuce samples produced in SP and PE are shown in Figure 4. Samples

collected in SP presented relatively higher values for almost all of these components than the samples obtained from PE. França et al. (2017) studied lettuce production in PE, and although the Zn content in their soil samples was higher than those observed by Santos-Araujo and Alleoni (2016), the Zn content in the plant was much lower. Because different varieties of lettuce present different Zn uptakes even when cultivated under the same soils conditions (França et al., 2017), lettuce varieties grown in SP may differ from varieties cultivated in PE, resulting in different levels of soil-plant transference.

The strong effect of soil variables on the plant classification could be explained by relevance of soil properties in plant uptake. A previous study carried out by Santos-Araujo and Alleoni (2016) showed that the most important covariates for predicting the Zn content in vegetables sampled in SP were CEC_e , pH, organic carbon, and the pseudo-total content of Zn and Cu.

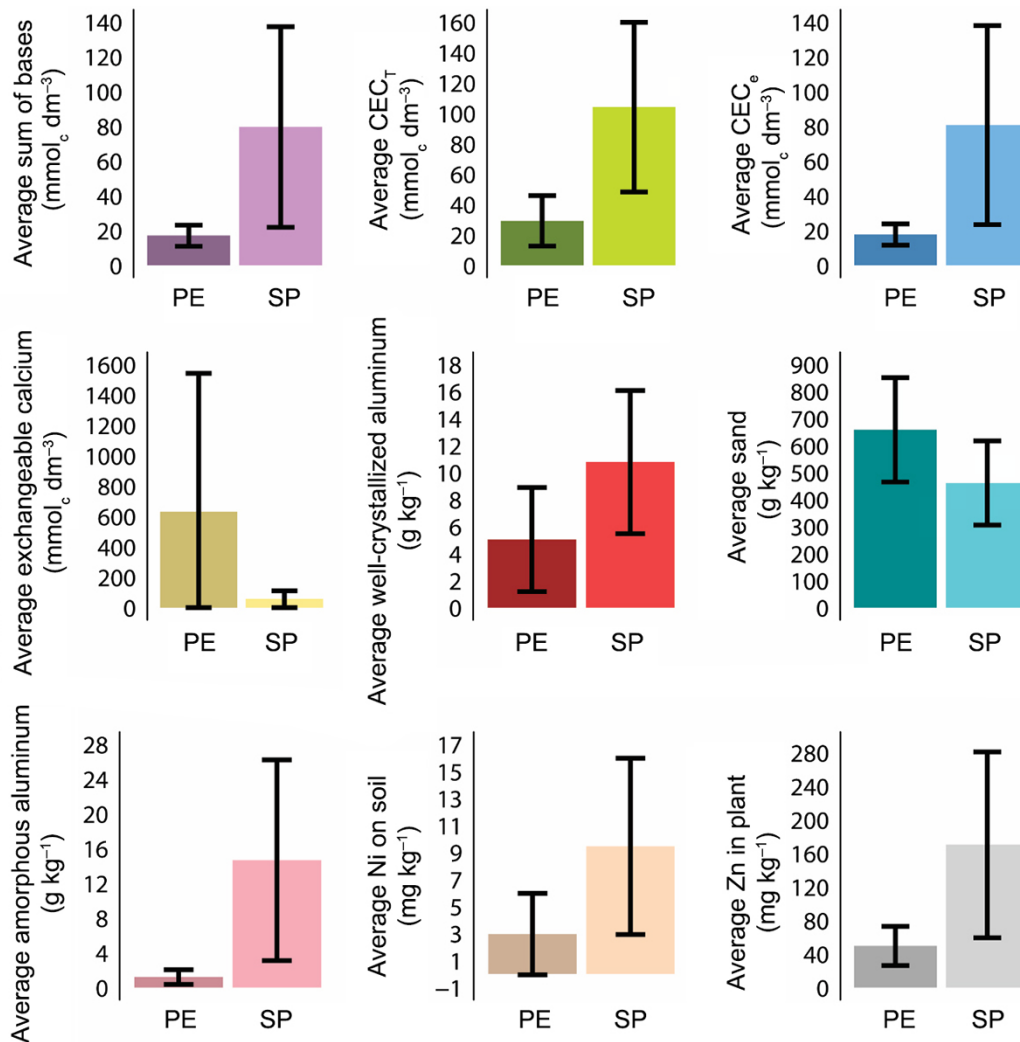


Figure 4 – Mean values of the sum of bases, CEC_T , CEC_e , exchangeable calcium in soil, well-crystallized aluminum in soil, sand, amorphous aluminum in soil, nickel in soil and zinc in plant for the lettuce samples produced in São Paulo (SP) and Pernambuco (PE) state.

As production is a result of cultivated area multiplied by yield, it is possible to use soil productivity to infer the incorporation of agricultural technology (Camargo Filho and Camargo, 2017). Therefore, the inclusion of soil parameters in the model for plant classification complements the assortment and may give insights into the geographical origin of lettuce.

Conclusion

The Sustainable Capitol Hill (SCH) and Michigan State University list several reasons for customers to buy and consume locally produced food (Klavinski, 2013; SCH, 2019). Because local food involves a shorter time and less transportation effort from harvest to customer table, it is likely to be safer to consume, fresher, less contaminated, more flavorful, and higher in nutritional value. It is also easier for customers to monitor the food origin and investigate practices and substances used to grow and harvest the crops. Purchasing local food also benefits the local economy as the money is retained within the community and reinvested in local businesses and services, also supporting local farmers, considerable importance in economic and food supply crises.

Verifying the geographic origin of food is a substantial matter to ascertain that this important kind of food was produced by a trusted source that takes quality and safety into account. In this study, we proposed a novel methodology to determine the geographical origin of Brazilian lettuce based on their elemental composition and soil properties through the use of SVM, LDA, and feature selection. We analyzed the contents of several chemical variables and soil properties determined for 72 lettuce samples obtained from São Paulo and Pernambuco States in Brazil. Through the use of a filter method for feature selection, we estimated that soil cation exchangeable capacity, exchangeable Ca, well-crystallized Al, sand, amorphous Al and Ni in soil, Zn levels in the plant and the sum of bases obtained by ($\text{Ca}^{2+} + \text{Mg}^{2+} + \text{K}^{+}$) were generally the most important variables for differentiating lettuce samples produced in both regions. We developed classification models based on SVM, which were capable of discriminating lettuce samples from both regions with a high accuracy level, presenting approximately 98.67 % correct predictions when all 28 chemical variables were used for training, and 95.81 % correct predictions when only the most important variables were used for training. These values surpass those obtained by the LDA model, a well-known, reliable and widely employed model for classification of food samples, which scored 66 % and 86.29 % prediction accuracy when all variables and only the best rated variables were used for training, respectively. The values achieved proved that, when combined with the chemical composition of lettuce samples determined by ICP OES and certain soil properties, classification models based on SVM could successfully determine the geographical origin of lettuce samples with excellent accuracy, at the

same time attesting that data mining techniques could powerfully support traceability strategies and ensure vegetable authenticity. Our previous literature search reveals that this is the first attempt to discriminate the geographical origin of Brazilian lettuce samples based on a powerful machine learning technique for data mining, such as SVM, also applied to chemical composition and soil parameters.

Acknowledgments

The authors would like to thank the São Paulo Research Foundation (FAPESP), Processes number 12/03682-2, 15/19332-9 and 15/25416-0; the National Council for Scientific and Technological Development (CNPq); and the Coordination for the Improvement of Higher Level Personnel (CAPES) for their financial support.

Authors' Contributions

Conceptualization, and Data acquisition: Araujo, E.M.; Santos-Araujo, S.N.; Boim, A.G.F.; Alleoni, L.R.F. **Design of methodology and Data analysis:** Maione, C. **Writing and editing:** Maione, C.; Araujo, E.M.; Santos-Araujo, S.N.; Boim, A.G.F.; Barbosa, R.M.; Alleoni, L.R.F.; Barbosa, R.M.

References

- Abbas, O.; Zadavec, M.; Baeten, V.; Mikuš, T.; Lešić, T.; Vulić, A.; Prpić, J.; Jemeršić, L.; Pleadin, J. 2018. Analytical methods used for the authentication of food of animal origin. *Food Chemistry* 246: 6-17.
- Akbarzadeh, S.; Paap, A.; Ahderom, S.; Apopei, B.; Alameh, K. 2018. Plant discrimination by Support Vector Machine classifier based on spectral reflectance. *Computers and Electronics in Agriculture* 148: 250-258.
- Alcázar, Á.; Jurado, J.M.; Palacios-Morillo, A.; Pablos, F.; Martín, M.J. 2012. Recognition of the geographical origin of beer based on support vector machines applied to chemical descriptors. *Food Control* 23: 258-262.
- Ali, S.M.; Hui, K.H.; Hee, L.M.; Leong, M.S. 2018. Automated valve fault detection based on acoustic emission parameters and support vector machine. *Alexandria Engineering Journal* 57: 491-498.
- Anderson, J.M.; Ingram, J.S.I. 1992. *Tropical Soil Biology and Fertility: A Handbook of Methods*. CAB International, Wallingford, UK.
- Araújo, C.L.; Nogueira, A.R.A.; Nobrega, J.A. 2002. Effect of acid concentration on closed-vessel microwave-assisted digestion of plant materials. *Spectrochimica Acta Part B: Atomic Spectroscopy* 57: 2121-2132.
- Baroni, M.V.; Podio, N.S.; Badini, R.G.; Inga, M.; Ostera, H.A.; Cagnoni, M.; Gautier, E.A.; García, P.P.; Hoogewerff, J.; Wunderlin, D.A. 2015. Linking soil, water, and honey composition to assess the geographical origin of Argentinean honey by multielemental and isotopic analyses. *Journal of Agricultural and Food Chemistry* 63: 4638-4645.

- Battineni, G.; Chintalapudi, N.; Amenta, F. 2019. Machine learning in medicine: performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked* 16: 100200.
- Berrueta, L.A.; Alonso-Salces, R.M.; Héberger, K. 2007. Supervised pattern recognition in food analysis. *Journal of Chromatography A* 1158: 196-214.
- Bhavan, A.; Chauhan, P.; Hitkul, Shah, R.R. 2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems* 184: 104886.
- Biondi, C.M.; Nascimento, C.W.A.; Fabricio Neta, A.B.; Ribeiro, M.R. 2011. Concentrations of Fe, Mn, Zn, Cu, Ni and Co in benchmark soils of Pernambuco, Brazil. *Revista Brasileira de Ciência do Solo* 35: 1057-1066.
- Bommert, A.; Sun, X.; Bischl, B.; Rahnenführer, J.; Lang, M. 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 143: 106839.
- Braga, D.; Madureira, A.M.; Coelho, L.; Ajith, R. 2019. Automatic detection of Parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence* 77: 148-158. <https://doi.org/10.1016/j.engappai.2018.09.018>
- Callao, M.P.; Ruisánchez, I. 2018. An overview of multivariate qualitative methods for food fraud detection. *Food Control* 86: 283-293.
- Camargo Filho, W.P.; Camargo, F.P. 2017. A quick review of the production and commercialization of the main vegetables in Brazil and the world from 1970 to 2015. *Horticultura Brasileira* 35: 160-166.
- Cambrai, A.; Marcic, C.; Morville, S.; Sae Houer, P.; Bindler, F.; Marchioni, E. 2010. Differentiation of chocolates according to the cocoa's geographical origin using chemometrics. *Journal of Agricultural and Food Chemistry* 58: 1478-1483.
- Carvalho, K.L.; Costa, R.P.; Souza, R.C. 2014. Strategic management of relationships in lettuce supply chain. *Production* 24: 271-282.
- Cavanna, D.; Righetti, L.; Elliott, C.; Suman, M. 2018. The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends Food Science and Technology* 80: 223-241.
- Ceballos-Magaña, S.G.; Jurado, J.M.; Muñoz-Valencia, R.; Alcázar, A.; Pablos, F.; Martín, M.J. 2012. Geographical authentication of tequila according to its mineral content by means of support vector machines. *Food Analytical Methods* 5: 260-265.
- Chawla, N.V. 2005. Data mining for imbalanced datasets: an overview. p. 853-867. In: Maimon O.; Rokach L., eds. *Data mining and knowledge discovery handbook*. Springer, New York, NY, USA.
- Chen, Y.-W.; Lin, C.-J. 2006. Combining SVMs with various feature selection strategies, p. 315-324. In: Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L.A. eds. *Feature extraction*. Springer, Berlin, Germany.
- Choubin, B.; Borji, M.; Mosavi, A.; Sajedi-Hosseini, F.; Singh, V.P.; Shamsirband, S. 2019a. Snow avalanche hazard prediction using machine learning methods. *Journal of Hydrology* 577: 123929. <https://doi.org/10.1016/j.jhydrol.2019.123929>
- Choubin, B.; Darabi, H.; Rahmati, O.; Sajedi-Hosseini, F.; Kløve, B. 2018. River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. *Science of the Total Environment* 615: 272-281.
- Choubin, B.; Moradi, E.; Golshan, M.; Adamowski, J.; Sajedi-Hosseini, F.; Mosavi, A. 2019b. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment* 651: 2087-2096.
- Coetzee, P.P.; Steffens, F.E.; Eiselen, R.J.; Augustyn, O.P.; Balcaen, L.; Vanhaecke, F. 2005. Multi-element analysis of South African wines by icp-ms and their classification according to geographical origin. *Journal of Agricultural and Food Chemistry* 53: 5060-5066.
- Cortes, C.; Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20: 273-297.
- Di, Z.; Gong, X.; Shi, J.; Ahmed, H.O.A.; Nandi, A.K. 2019. Internet addiction disorder detection of Chinese college students using several personality questionnaire data and support vector machine. *Addictive Behaviors Reports* 10: 100200.
- Duda, R.O.; Hart, P.E.; Stork, D.G. 2001. *Pattern Classification*, 2ed. Wiley-Interscience, Hoboken, NJ, USA.
- Esteki, M.; Shahsavari, Z.; Simal-Gandara, J. 2019. Food identification by high performance liquid chromatography fingerprinting and mathematical processing. *Food Research International* 122: 303-317.
- Esteki, M.; Shahsavari, Z.; Simal-Gandara, J. 2018a. Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products. *Food Control* 91: 100-112.
- Esteki, M.; Simal-Gandara, J.; Shahsavari, Z.; Zandbaaf, S.; Dashtaki, E.; Vander Heyden, Y. 2018b. A review on the application of chromatographic methods, coupled to chemometrics, for food authentication. *Food Control* 93: 165-182.
- Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. 2018. Comparison of Support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Conservation and Management* 164: 102-111. <https://doi.org/10.1016/j.enconman.2018.02.087>
- Fazai, R.; Abodayeh, K.; Mansouri, M.; Trabelsi, M.; Nounou, H.; Nounou, M.; Georghiou, G.E. 2019. Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems. *Solar Energy* 190: 405-413.
- Feng, P.; Wang, B.; Liu, D.L.; Yu, Q. 2019. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern, Australia. *Agricultural Systems* 173: 303-316. <https://doi.org/10.1016/j.agsy.2019.03.015>
- Fernandes, A.M.; Utkin, A.B.; Eiras-Dias, J.; Cunha, J.; Silvestre, J.; Melo-Pinto, P. 2019. Grapevine variety identification using "Big Data" collected with miniaturized spectrometer combined with support vector machines and convolutional neural networks. *Computers and Electronics in Agriculture* 163: 104855. <https://doi.org/10.1016/j.compag.2019.104855>

- Fink, J.R.; Inda, A.V.; Bayer, C.; Torrent, J.; Barrón, V. 2014. Mineralogy and phosphorus adsorption in soils of south and central-west Brazil under conventional and no-tillage systems. *Acta Scientiarum. Agronomy* 36: 379–387.
- França, F.C.S.S.; Albuquerq, A.M.A.; Almeida, A.C.; Silveira, P.B.; Silva Filho, C.A.; Hazin, C.A.; Honorato, E.V. 2017. Heavy metals deposited in the culture of lettuce (*Lactuca sativa* L.) by the influence of vehicular traffic in Pernambuco. *Brazilian Food Chemistry* 215: 171–176.
- García-Nieto, P.J.; García-Gonzalo, E.; Fernández, J.R.A.; Muñiz, C.D. 2019. Modeling of the algal atypical increase in La Barca reservoir using the DE optimized least square support vector machine approach with feature selection. *Mathematics and Computers in Simulation* 166: 461–480.
- Gee, G.W.; Or, D. 2002. Particle-size analysis. p. 241–254. In: Dane, J.H.; Topp, G.C., eds. *Methods of soil analysis. Part 4. Physical methods*. Soil Science Society of America, Madison, WI, USA.
- Ghalyani, P.; Mazinan, A.H. 2019. Performance-based fault detection approach for the dew point process through a fuzzy multi-label support vector machine. *Measurement: Journal of the International Measurement Confederation* 144: 214–224.
- Goldberg, S. 1989. Interaction of aluminum and iron oxides and clay minerals and their effect on soil physical properties: a review. *Communications in Soil Science and Plant Analysis* 20: 1181–1207.
- Granato, D.; Santos, J.S.; Escher, G.B.; Ferreira, B.L.; Maggio, R.M. 2018. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology* 72: 83–90.
- Griffel, L.M.; Delparte, D.; Edwards, J. 2018. Using support vector machines classification to differentiate spectral signatures of potato plants infected with potato virus Y. *Computers and Electronics in Agriculture* 153: 318–324. <https://doi.org/10.1016/j.compag.2018.08.027>
- Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. 2017. Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications* 73: 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- Han, H.; Cui, X.; Fan, Y.; Qing, H. 2019. Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features. *Applied Thermal Engineering* 154: 540–547.
- He, H.; Garcia, E.A. 2009. Learning from Imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21: 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Huang, G.; Qiu, W.; Zhang, J. 2017. Modelling seismic fragility of a rock mountain tunnel based on support vector machine. *Soil Dynamics and Earthquake Engineering* 102: 160–171.
- Inda Junior, A.V.; Kämpf, N. 2003. Evaluation of pedogenic iron oxide extraction procedures with sodium dithionite-citrate-bicarbonate. *Revista Brasileira de Ciência do Solo* 27: 1139–1147.
- Izenman, A.J. 2008. *Modern Multivariate Statistical Techniques*. Springer Science, New York, NY, USA.
- Jain, A.K. 2010. Data clustering: 50 years beyond K-means, pattern recognition letters. *Pattern Recognition Letters* 31: 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jiménez-Carvelo, A.M.; González-Casado, A.; Bagur-González, M.G.; Cuadros-Rodríguez, L. 2019. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity: a review. *Food Research International* 122: 25–39.
- Jo, T.; Japkowicz, N. 2004. Class imbalances versus small disjuncts. *ACM SIGKDD Exploration Newsletters* 6: 40. <https://doi.org/10.1145/1007730.1007737>
- Jung, H.; Jo, H.; Kim, S.; Lee, K.; Choe, J. 2018. Geological model sampling using PCA-assisted support vector machine for reliable channel reservoir characterization. *Journal of Petroleum Science and Engineering* 167: 396–405.
- Karimi, F.; Sultana, S.; Shirzadi Babakan, A.; Suthaharan, S. 2019. An enhanced support vector machine model for urban expansion prediction. *Comput. Environment and Urban Systems* 75: 61–75.
- Kemsley, E.K.; Defernez, M.; Marini, F. 2019. Multivariate statistics: considerations and confidences in food authenticity problems. *Food Control* 105: 102–112.
- Kim, M.J.; Moon, Y.; Tou, J.C.; Mou, B.; Waterland, N.L. 2016. Nutritional value, bioactive compounds and health benefits of lettuce (*Lactuca sativa* L.). *Journal of Food Composition and Analysis* 49: 19–34.
- Kisi, O.; Choubin, B.; Deo, R.C.; Yaseen, Z.M. 2019. Incorporating synoptic-scale climate signals for streamflow modelling over the Mediterranean region using machine learning models. *Hydrological Sciences Journal* 64: 1240–1252.
- Klavinski, R. 2013. 7 benefits of eating local foods. Available at: https://www.canr.msu.edu/news/7_benefits_of_eating_local_foods [Accessed Nov 4, 2019]
- Kotsiantis, S.B.; Zaharakis, I.D.; Pintelas, P.E. 2006. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* 26: 159–190.
- Kumar, D.; Thakur, M.; Dubey, C.S.; Shukla, D.P. 2017. Landslide susceptibility mapping and prediction using support vector machine for Mandakini River Basin, Garhwal Himalaya, India. *Geomorphology* 295: 115–125.
- Kumpiene, J.; Giagnoni, L.; Marschner, B.; Denys, S.; Mench, M.; Adriaensen, K.; Vangronsveld, J.; Puschenreiter, M.; Renella, G. 2017. Assessment of methods for determining bioavailability of trace elements in soils: a review. *Pedosphere* 27: 389–406.
- Kundu, S.; Khare, D.; Mondal, A. 2017. Future changes in rainfall, temperature and reference evapotranspiration in the central India by least square support vector machine. *Geoscience Frontiers* 8: 583–596. <https://doi.org/10.1016/j.gsf.2016.06.002>
- Leena, N.; Saju, K.K. 2019. Classification of macronutrient deficiencies in maize plants using optimized multi class support vector machines. *Engineering in Agriculture, Environment and Food* 12: 126–139. <https://doi.org/10.1016/j.eaef.2018.11.002>
- Liu, J.; Xu, H.; Chen, Q.; Zhang, T.; Sheng, W.; Huang, Q.; Song, J.; Huang, D.; Lan, L.; Li, Y.; Chen, W.; Yang, Y. 2019. Prediction of hematoma expansion in spontaneous intracerebral hemorrhage using support vector machine. *EBioMedicine* 43: 454–459.
- Liu, J.; Zio, E. 2018. A scalable fuzzy support vector machine for fault detection in transportation systems. *Expert Systems with Applications* 102: 36–43.

- López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 250: 113–141. <https://doi.org/10.1016/j.ins.2013.07.007>
- Lukmanto, R.B.; Suharjo, Nugroho, A.; Akbar, H. 2019. Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. *Procedia Computer Science* 157: 46–54.
- Ma, Z.; Ye, C.; Li, H.; Ma, W. 2018. Applying support vector machines to predict building energy consumption in China. *Energy Procedia* 152: 780–786.
- Mahne Opatić, A.; Nečemer, M.; Lojen, S.; Masten, J.; Zlatić, E.; Šircelj, H.; Stopar, D.; Vidrih, R. 2018. Determination of geographical origin of commercial tomato through analysis of stable isotopes, elemental composition and chemical markers. *Food Control* 89: 133–141.
- Mahvash, N.M.; Hezarkhani, A. 2018. Application of support vector machine for the separation of mineralised zones in the Takht-e-Gonbad porphyry deposit, SE Iran. *Journal of African Earth Sciences* 143: 301–308.
- Mainville, D.Y.; Peterson, H.C. 2005. Fresh Produce Procurement Strategies in a Constrained Supply Environment: Case Study of Companhia Brasileira de Distribuicao. *Applied Economic Perspectives and Policy* 27: 130–138.
- Maione, C.; Barbosa, F.; Barbosa, R.M. 2019. Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: a review. *Computers and Electronics in Agriculture* 157: 436–446. <https://doi.org/10.1016/j.compag.2019.01.020>
- Maione, C.; Barbosa, R.M. 2018. Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: a review. *Critical Reviews in Food Science and Nutrition* 12: 1868–1879. <https://doi.org/10.1080/10408398.2018.1431763>
- Manjurul Islam, M.M.; Kim, J.M. 2019. Reliable multiple combined fault diagnosis of bearings using heterogeneous feature models and multiclass support vector machines. *Reliability Engineering & System Safety* 184: 55–66.
- McKeague, J.A.; Day, J.H. 1966. Dithionite and oxalate extractable Fe and Al as aids in differentiating various classes of soils. *Canadian Journal of Soil Science* 46: 13–22.
- Medina, S.; Pereira, J.A.; Silva, P.; Perestrelo, R.; Câmara, J.S. 2019a. Food fingerprints: a valuable tool to monitor food authenticity and safety. *Food Chemistry* 278: 144–162.
- Medina, S.; Perestrelo, R.; Silva, P.; Pereira, J.A.M.; Câmara, J.S. 2019b. Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science and Technology* 85: 163–176.
- Mehra, J.P.; Jackson, M.L. 1960. Iron oxides removal from soils and clays by a dithionite-citrate-bicarbonate system buffered with bicarbonate sodium. *Clays and Clay Minerals* 7: 317–327.
- Moreda-Piñeiro, A.; Fisher, A.; Hill, S.J. 2003. The classification of tea according to region of origin using pattern recognition techniques and trace metal data. *Journal of Food Composition and Analysis* 16: 195–211.
- Oliveira, M.; Ramos, S.; Delerue-Matos, C.; Morais, S. 2015. Espresso beverages of pure origin coffee: mineral characterization, contribution for mineral intake and geographical discrimination. *Food Chemistry* 177: 330–338.
- Oliveri, P. 2017. Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues: a tutorial. *Analytica Chimica Acta* 982: 9–19.
- Peris, M.; Escuder-Gilbert, L. 2016. Electronic noses and tongues to assess food authenticity and adulteration. *Trends in Food Science and Technology* 58: 40–54.
- Potorti, A.G.; Bella, G. Di; Mottese, A.F.; Bua, G.D.; Fedè, M.R.; Sabatino, G.; Salvo, A.; Somma, R.; Dugo, G.; Turco, V.L. 2018. Traceability of protected geographical indication (PGI) Interdonato lemon pulps by chemometric analysis of the mineral composition. *Journal of Food Composition and Analysis* 69: 122–128.
- Prati, R.C.; Batista, G.E.A.P.A.; Monard, M.C. 2004. Class imbalances versus class overlapping: an analysis of a learning system behavior. p. 312–321. In: Monroy R.; Arroyo-Figueroa, G.; Sucar, L.E.; Sossa, H., eds. *MICAI 2004: advances in artificial intelligence*. Springer, Berlin, Germany.
- Pu, Y.; Apel, D.B.; Liu, V.; Mitri, H. 2019. Machine learning methods for rockburst prediction-state-of-the-art review. *International Journal of Mining Science and Technology* 29: 565–570. <https://doi.org/10.1016/j.ijmst.2019.06.009>
- Radhakrishnan, S.; Ramanathan, R., 2018. A support vector machine with Gabor features for animal intrusion detection in agriculture fields. *Procedia Computer Science* 143: 493–501.
- Rahmati, O.; Choubin, B.; Fathabadi, A.; Coulon, F.; Soltani, E.; Shahabi, H.; Mollaefar, E.; Tiefenbacher, J.; Cipullo, S.; Ahmad, B.B.; Bui, D.T. 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. *Science of the Total Environment* 688: 855–866. <https://doi.org/10.1016/j.scitotenv.2019.06.320>
- Rahmeni, R.; Aicha, A.B.; Ayed, Y.B. 2019. Speech spoofing countermeasures based on source voice analysis and machine learning techniques. *Procedia Computer Science* 159: 668–675. <https://doi.org/10.1016/j.procs.2019.09.222>
- Rodrigues, S.M.; Henriques, B.; Silva, E.F.; Pereira, M.E.; Duarte, A.C.; Römkens, P.F.A.M. 2010. Evaluation of an approach for the characterization of reactive and available pools of twenty potentially toxic elements in soils. Part I. The role of key soil properties in the variation of contaminants' reactivity. *Chemosphere* 81: 1549–1559.
- Ropodi, A.I.; Panagou, E.Z.; Nychas, G.J.E. 2016. Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines. *Trends in Food Science and Technology* 50: 11–25.
- Saari, J.; Strömbergsson, D.; Lundberg, J.; Thomson, A. 2019. Detection and identification of windmill bearing faults using a one-class support vector machine (SVM). *Measurement: Journal of the International Measurement Confederation* 137: 287–301.
- Sajedi-Hosseini, F.; Malekian, A.; Choubin, B.; Rahmati, O.; Cipullo, S.; Coulon, F.; Pradhan, B. 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment* 644: 954–962.
- Santos-Araujo, S.N.; Alleoni, L.R.F. 2016. Concentrations of potentially toxic elements in soils and vegetables from the macroregion of São Paulo, Brazil: availability for plant uptake. *Environmental Monitoring and Assessment* 188: 1–17.

- Serra, F.; Guillou, C.G.; Reniero, F.; Ballarin, L.; Cantagallo, M.I.; Wieser, M.; Iyer, S.S.; Héberger, K.; Vanhaecke, F. 2005. Determination of the geographical origin of green coffee by principal component analysis of carbon, nitrogen and boron stable isotope ratios. *Rapid Communications in Mass Spectrometry* 19: 2111-2115.
- Solano Meza, J.K.; Orjuela Yepes, D.; Rodrigo-Illari, J.; Cassiraga, E. 2019. Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. *Heliyon* 5: e02810.
- Sustainable Capitol Hill [SCH]. 2019. Top 10 Benefits of eating local, seasonal, organic food. Available at: <https://sustainablecapitolhill.org/eating-locally-in-capitol-hill>. [Accessed Nov 4, 2019]
- Tan, P.-N.; Steinbach, M.; Kumar, V. 2005. Introduction to data mining. Addison Wesley, Boston, MA, USA.
- Tang, J.; Chen, X.; Hu, Z.; Zong, F.; Han, C.; Li, L. 2019. Traffic flow prediction based on combination of support vector machine and data denoising schemes. *Physica A: Statistical Mechanics and its Applications* 534: 120642.
- Valdés, A.; Beltrán, A.; Mellinas, C.; Jiménez, A.; Garrigós, M.C. 2018. Analytical methods combined with multivariate analysis for authentication of animal and vegetable food products with high fat content. *Trends in Food Science and Technology* 77: 120-130.
- Varma, S.; Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7: 91.
- Vougas, K.; Sakellaropoulos, T.; Kotsinas, A.; Foukas, G.R.P.; Ntargaras, A.; Koinis, F.; Polyzos, A.; Myrianthopoulos, V.; Zhou, H.; Narang, S.; Georgoulis, V.; Alexopoulos, L.; Aifantis, I.; Townsend, P.A.; Sfrikakis, P.; Fitzgerald, R.; Thanos, D.; Bartek, J.; Petty, R.; Tsirigos, A.; Gorgoulis, V.G. 2019. Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining. *Pharmacology and Therapeutics* 203: 107395.
- Wadood, S.A.; Boli, G.; Xiaowen, Z.; Hussain, I.; Yimin, W. 2020. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchemical Journal* 152: 104295.
- Wang, J.; Shan, Y.; Xie, X.; Kuang, J. 2019. Output-based speech quality assessment using autoencoder and support vector regression. *Speech Communication* 110: 13-20. <https://doi.org/10.1016/j.specom.2019.04.002>
- Wang, X.; Luo, D.; Zhao, X.; Sun, Z. 2018. Estimates of energy consumption in China using a self-adaptive multi-verse optimizer-based support vector machine with rolling cross-validation. *Energy* 152: 539-548.
- Xi, P.P.; Zhao, Y.P.; Wang, P.X.; Li, Z.Q.; Pan, Y.T.; Song, F.Q. 2019. Least squares support vector machine for class imbalance learning and their applications to fault detection of aircraft engine. *Aerospace Science and Technology* 84: 56-74.
- Xiao, R.; Hu, Q.; Li, J. 2019. Leak detection of gas pipelines using acoustic signals based on wavelet transform and Support Vector Machine. *Measurement: Journal of the International Measurement Confederation* 146: 479-489.
- Yu, C.; Li, Y.; Bao, Y.; Tang, H.; Zhai, G. 2018. A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. *Energy Conversion and Management* 178: 137-145.
- Yu, P.S.; Yang, T.C.; Chen, S.Y.; Kuo, C.M.; Tseng, H.W. 2017. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of Hydrology* 552: 92-104. <https://doi.org/10.1016/j.jhydrol.2017.06.020>
- Zendehboudi, A.; Baseer, M.A.; Saidur, R. 2018. Application of support vector machine models for forecasting solar and wind energy resources: a review. *Journal of Cleaner Production* 199: 272-285.
- Zhou, Z.; Morel, J.; Parsons, D.; Kucheryavskiy, S.V.; Gustavsson, A.M. 2019. Estimation of yield and quality of legume and grass mixtures using partial least squares and support vector machine analysis of spectral data. *Computers and Electronics in Agriculture* 162: 246-253.
- Zhu, S.; Chen, H.; Wang, M.; Guo, X.; Lei, Y.; Jin, G. 2019. Plastic solid waste identification system based on near infrared spectroscopy in combination with support vector machine. *Advanced Industrial and Engineering Polymer Research* 2: 77-81.