

RESEARCH ARTICLE

Use of data mining techniques to classify soil CO₂ emission induced by crop management in sugarcane field

Camila Viana Vieira Farhate^{1*}, Zigomar Menezes de Souza¹, Stanley Robson de Medeiros Oliveira^{1,2}, Rose Luiza Moraes Tavares³, João Luís Nunes Carvalho⁴

1 School of Agricultural Engineering, State University of Campinas, Campinas, São Paulo, Brazil, **2** Agricultural Informatics, Computational Intelligence Laboratory, Brazilian Agricultural Research Corporation (Embrapa), Campinas, São Paulo, Brazil, **3** Department of Soil, Rio Verde University, Rio Verde, Goiás, Brazil, **4** Brazilian Bioethanol Science and Technology Laboratory, Brazilian Center for Research in Energy and Materials, Campinas, São Paulo, Brazil

* camila.vieira@feagri.unicamp.br



OPEN ACCESS

Citation: Farhate CVV, Souza ZMd, Oliveira SRdM, Tavares RLM, Carvalho JLN (2018) Use of data mining techniques to classify soil CO₂ emission induced by crop management in sugarcane field. PLoS ONE 13(3): e0193537. <https://doi.org/10.1371/journal.pone.0193537>

Editor: Asim Zia, University of Vermont, UNITED STATES

Received: July 7, 2017

Accepted: February 13, 2018

Published: March 7, 2018

Copyright: © 2018 Farhate et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors declare that the co-author of this paper, Dr. Stanley Robson de Medeiros Oliveira, is a researcher at Embrapa (Brazilian Agricultural Research Corporation) as well as a contributing professor in the graduate program at the School of Agricultural Engineering, State University of Campinas, Campinas, São Paulo, Brazil. The authors further declare that Dr. Stanley

Abstract

Soil CO₂ emissions are regarded as one of the largest flows of the global carbon cycle and small changes in their magnitude can have a large effect on the CO₂ concentration in the atmosphere. Thus, a better understanding of this attribute would enable the identification of promoters and the development of strategies to mitigate the risks of climate change. Therefore, our study aimed at using data mining techniques to predict the soil CO₂ emission induced by crop management in sugarcane areas in Brazil. To do so, we used different variable selection methods (correlation, chi-square, wrapper) and classification (Decision tree, Bayesian models, neural networks, support vector machine, bagging with logistic regression), and finally we tested the efficiency of different approaches through the Receiver Operating Characteristic (ROC) curve. The original dataset consisted of 19 variables (18 independent variables and one dependent (or response) variable). The association between cover crop and minimum tillage are effective strategies to promote the mitigation of soil CO₂ emissions, in which the average CO₂ emissions are 63 kg ha⁻¹ day⁻¹. The variables soil moisture, soil temperature (Ts), rainfall, pH, and organic carbon were most frequently selected for soil CO₂ emission classification using different methods for attribute selection. According to the results of the ROC curve, the best approaches for soil CO₂ emission classification were the following: (I)—the Multilayer Perceptron classifier with attribute selection through the wrapper method, that presented rate of false positive of 13,50%, true positive of 94,20% area under the curve (AUC) of 89,90% (II)—the Bagging classifier with logistic regression with attribute selection through the Chi-square method, that presented rate of false positive of 13,50%, true positive of 94,20% AUC of 89,90%. However, the (I) approach stands out in relation to (II) for its higher positive class accuracy (high CO₂ emission) and lower computational cost.

Oliveira has received no salary, gratification, bonus or gratuity for taking part in this research. Every effort made in this study was result of research collaboration among students, professors and researchers interested in the research line of this study. In summary, Dr. Oliveira's contribution is related to the co-supervision of this research work, as well as the use of machine learning algorithms (predictive modeling) and interpretation of the results. Embrapa did not provide any financial support for this research.

Competing interests: The authors also state that there is no competing interests with Embrapa, neither commercial interests, consultancy, patents, products in development, or marketed products. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Introduction

Carbon dioxide (CO₂) is continuously exchanged between soils and the atmosphere, mainly through the processes of photosynthesis and incorporation of organic matter derived from plants (efflux of CO₂) and the decomposition of this organic matter by soil organisms (CO₂ flux). Therefore, the amount of carbon (C) stored in the soils depends mainly on the equilibrium between the C inputs and outputs of the system [1].

In this context, increased C storage in soils worldwide could help compensate for the increase in anthropogenic CO₂ emissions, while increased soil emissions could significantly worsen atmospheric rises [2].

Thus, the potential of agriculture to mitigate greenhouse gas emissions has been the subject of intense scientific research over recent years [1]. However, after many decades of research on soil organic matter decomposition, very few robust mathematical models and experiments to predict the effect of both biotic and abiotic factors on the soil C balance have been developed [3].

Considering this issue, the alternative that we propose is the use of data mining techniques to predict CO₂ emission from the soil. Among the techniques of data mining, classification is a task that stands out in the studies of the scientific community of Knowledge Discovery in Database (KDD). The principle of this task is to predict a categorical variable, i.e., to discover a function that maps a set of records into a set of predefined variables, called classes. Such a function can be applied to new records in order to predict the class in which such records fit [4, 5].

Several algorithms can be used to perform such a task; however, some of them, such as J48, Naive Bayes, Multilayer Perceptron, SVM, Bagging, are widely used due to their usual good performance regarding classification processes. However, the models obtained through the use of different classifiers require assessment under the same conditions.

An alternative to compare predictive models is the Receiver Operating Characteristic (ROC) curve, which is a graphical method for evaluation, organization, and selection of diagnostic and/or prediction systems. This tool has been widely used in data mining processes to assess classification models and is particularly useful in areas presenting a great disparity between classes or when different costs/benefits must be taken into account for the different errors/correctness in classification [6].

Hence, our study aimed at using data mining techniques to predict the soil CO₂ emission induced by crop management in sugarcane areas in Brazil. To accomplish that, we used different variable selection methods (correlation, chi-square, wrapper) and classification (Decision tree, Bayesian models, neural networks, support vector machine, bagging with logistic regression), and finally we compared the efficiency of different approaches through the ROC curve (Receiver Operating Characteristic).

Materials and methods

Site locations and experiment descriptions

The composition of the original database for this work was carried out by collecting soil variable (physical, chemical, microbiological) and climatic data at an experimental sugarcane area located in the city of Iracemápolis, São Paulo, Brazil (22°34'50" S, 47°31'07" W; 608 m above sea level), during the sugarcane crop reform period.

The soil under investigation is classified as Rhodic Hapludox, according to the Soil Taxonomy System [7], and is described as a Latossolo Vermelho Eutroférico, according to the Brazilian Soil Classification [8], with very clayey texture. In the site characterization, the soil

presented a pH around 4.6, an average organic C content of 10.5 g C dm⁻³, base saturation of 50%, cation exchange capacity 10.25 cmol_c dm⁻³, bulk density of 1.34 kg dm⁻³, from 0–40 cm depth. The distribution of particle sizes of sand, silt, and clay were 140, 194 and 666 g kg⁻¹, respectively.

The data were obtained from a field experiment with a randomized complete block design in the split-plot design with four replicates. The plots involved treatments with and without cover crop with *Crotalaria Juncea* and subplots with two soil tillage conditions (minimum and conventional tillage). Each experimental unit (sub-plot) consisted of 15 lines of sugarcane, with a spacing of 1.5 m and 34 m in length.

Analysis of CO₂ emissions, soil temperature, and soil moisture

Soil CO₂ emissions were recorded between April 27 and August 3, 2013. Emissions were measured in the morning (i.e. between 8 and 10 am) [9], using five PVC collars (diameter = 10 cm and height = 7 cm), inserted 0.02 m into the soil at each of the 16 plots (80 collars in total). The PVC collars were distributed in sugarcane field as follow: two of them located in the in-row, one in an intermediate position considered as seedbed region (at 0.38 m distance from in-row), and two in the center of the inter-row (at 0.75 m distance from in-row).

Soil CO₂ emissions were measured using an infrared gas analyzer (IRGA) produced by LI-COR (model LI8100A, Nebraska, USA). This chamber was a closed system with an internal volume of 991 cm³ and a contact area with the soil of 71.6 cm². When recording emissions, the chamber was placed over the PVC collars to prevent any mechanical disturbance to the soil profile, which may result in overestimation of emissions.

The quantification of CO₂ emissions began 24 hours after soil tillage operations. In each plot, the CO₂ emission measurement lasted 90 seconds. The measurements were taken for 97 days after soil tillage, until CO₂ emissions had stabilized. The CO₂ emissions recorded from the five collars within each plot (two in intra-rows and three in inter-rows) were aggregated into a single measure using a weighted average (assuming an area of 27% for the rows and 73% for the inter-rows). Soil CO₂ emission over the entire study period was estimated by the integral of the area formed below the emission curves over time.

The measurements of soil temperature and soil moisture were taken at exactly the same time as soil CO₂ measurements. Soil temperature (St) was assessed at all points studied with the sensor which is part of the same LI-COR instrument (model LI8100A). This sensor consists of a 20 cm rod inserted into the soil perpendicular to the surface near the PVC collars used to measure the soil respiration. Soil moisture (Sm) was measured simultaneously with the measurement of CO₂ concentration using Time Domain Reflectometry (TDR). Probe Thetaprobe ML2 (Delta-T Devices, Cambridge, UK) is an English manufacturing tool that directly measures the water content in the soil, corresponding to the volumetric moisture content, using the principle of wave generation which releases an electromagnetic pulse to a set of rods with reflection measured in TDR. Moreover, during experimental period, rainfall events were monitored through an automatic surface weather station installed at the experiment site.

In the end of the CO₂ sampling period, disturbed and undisturbed soil samples were collected at the depths of 0.00–0.10 m, 0.10–0.20 m, 0.20–0.30 m, and 0.30–0.40 m. to evaluate the effects of the treatments on soil attributes.

Soil physical, chemical and biological variables

Total soil porosity (Sp), macroporosity (Macro), microporosity (Micro), and bulk density (Bd) were analyzed according to the Brazilian Agricultural Research Corporation methodologies [10]. The soil resistance to penetration (RP) was obtained using the Stolf's formulation [11].

The mean diameter of the aggregates (MDA) was determined according to the method described by Kemper and Chepil [12] and the calculation of the aggregate tensile strength (Ts) was performed as described by Dexter and Kroesbergen [13].

All samples were taken to the laboratory, air dried and subsequently passed through a 2.0 mm mesh. We measured soil pH (CaCl₂ 0,01 mol L⁻¹), exchangeable cations (Ca²⁺, Mg²⁺ and K⁺), phosphorus available in resin (P), organic C concentration (wet oxidation), acidity potential, CEC potential, and base saturation in accordance with the methodology proposed by Rajj [14].

For the analysis of microbial biomass C, the samples were collected and placed in a cooler and refrigerated during the transportation to the laboratory for preservation in cold storage at 4°C until analysis. The microbial biomass C (MBC) was determined through the fumigation-extraction method proposed by Vance [15].

Data mining

The original dataset was constituted of 19 variables (one dependent or response variable and 18 independent or explanatory variables) (Table 1) which were added to the dataset numbering 1,552 observations. The variable-target refers to the soil CO₂ emission and is the classification target.

The 18 independent variables were constituted of nine soil physical attributes (Sm, Ts, MDA, Macro, Micro, TP, Bd, PR, Ts), seven soil chemical attributes (H+Al, Al³⁺, Ca²⁺, Mg²⁺, K⁺, P, C), one soil microbial attribute (MBC), and a climatic variable (daily rainfall obtained at a weather station located in the sugarcane mill) (Table 1).

The data were initially evaluated by through descriptive statistics through Box-plot graphs. The box shows the data between the first and third quartiles (hinge), with median represented by a line inside the box. Vertical lines (whiskers), starting in the middle of the base (and top) of the box and ending in values (referred to as adjacent lower and upper) approximately indicate the variability of the data [16].

Table 1. Description of the 19 variables (independent and dependent) used in the database composition.

| Variable | Type | Abbreviation | Description | Unit |
|------------------------|-------------|--------------------------|-----------------------------------|---------------------------------------|
| Physical | Independent | Sm | Soil moisture | % |
| | Independent | St | Soil temperature | °C |
| | Independent | MDA | Mean diameter of the aggregate | mm |
| | Independent | Macro | Soil macroporosity | % |
| | Independent | Micro | Soil microporosity | % |
| | Independent | TP | Total porosity | % |
| | Independent | Bd | Bulk density | kg dm ⁻³ |
| | Independent | PR | Penetration resistance | MPa |
| | Independent | Ts | Tensile strength of the aggregate | kPa |
| Chemical | Independent | H+Al | Acidity potential | cmol _c dm ⁻³ |
| | Independent | Al ³⁺ | Exchangeable aluminum | cmol _c dm ⁻³ |
| | Independent | Ca ²⁺ | Exchangeable calcium | cmol _c dm ⁻³ |
| | Independent | Mg ²⁺ | Exchangeable magnesium | cmol _c dm ⁻³ |
| | Independent | K ⁺ | Exchangeable potassium | cmol _c dm ⁻³ |
| | Independent | P | Exchangeable phosphorus | mg dm ⁻³ |
| | Independent | Organic C | Soil Organic Carbon | mg dm ⁻³ |
| Microbiological | Independent | MBC | Microbial biomass carbon | Ug C g ⁻¹ |
| Climatic | Independent | P | Rainfall | mm day ⁻¹ |
| Variable-target | Dependent | CO ₂ emission | soil CO ₂ emission | kg ha ⁻¹ dia ⁻¹ |

<https://doi.org/10.1371/journal.pone.0193537.t001>

In order to identify different levels of CO₂ emissions from the soil, a goal attribute discretization was required. For this purpose, the CO₂ emission values (kg ha⁻¹ dia⁻¹) were organized in ascending order and divided equally into three emission classes: low, medium and high (Table 2).

In order to compare the performance of different classifiers through the ROC curve, a new discretization was performed considering the high class belonging to the positive class and the remaining classes (low and medium) were grouped to form the negative class (Fig 1).

To do so, an imbalance between the classes occurred since the negative class obtained 2/3 of the data and the positive class kept only 1/3, hampering the classifiers to learn the positive class. For the purpose of improving the accuracy of the model and mitigating the problem of unbalanced classes, the "under sample" method was applied to balance the number of observations per class. This method is used in the training process to slightly reduce the observations of the major (negative) class and allow the (positive) minor class to be learned by the classifier, considering that the interest class is most often the positive class. For this purpose, the Stratified Remove Folds Filter was used to select two subsets, one for training, containing 90% of the original dataset, and one for testing, with 10% of the original dataset. Subsequently, we applied the Neighborhood Cleaning Rule (NCL) filter) available in Weka software 3.4.13. The distribution of classes in the database after this operation is shown in Fig 2.

From this subset, the selection of variables was performed using different methods:

1. Absence of attribute selection upon the occurrence of use of all data.
2. Correlation-based feature selection (CFS) searching for the set of correlated variables to prevent the reuse of the same information [17].
3. The chi-square method (χ^2) is based on the concept of statistical independence.
4. Wrapper method, which functions as a "black box". In order to find the subset of variables that most satisfactorily fits the classification algorithm [18].

From the variables selected through each selection method, different classification algorithms were tested:

1. J48—Decision tree [19].
2. Naïve Bayes—Bayesian Classifier [20].
3. Multilayer Perceptron—Neural networks.
4. SMO (Sequential Minimal Optimization)—Vector Machine Support [21].
5. Bagging with Logistic Regression—Meta classifier [22].

The different approaches were validated using the supplied test set method (90% of the database for training and 10% for testing) and two metrics: (i) accuracy rate; (ii) Kappa coefficient.

Table 2. Distribution of CO₂ emission (kg ha⁻¹ dia⁻¹) according to the low, medium and high classes and their limits.

| Class | Limit |
|--------|----------------|
| Low | [3.8; 29.95] |
| Medium | [29.96; 53.63] |
| High | [53.64; 770.5] |

<https://doi.org/10.1371/journal.pone.0193537.t002>

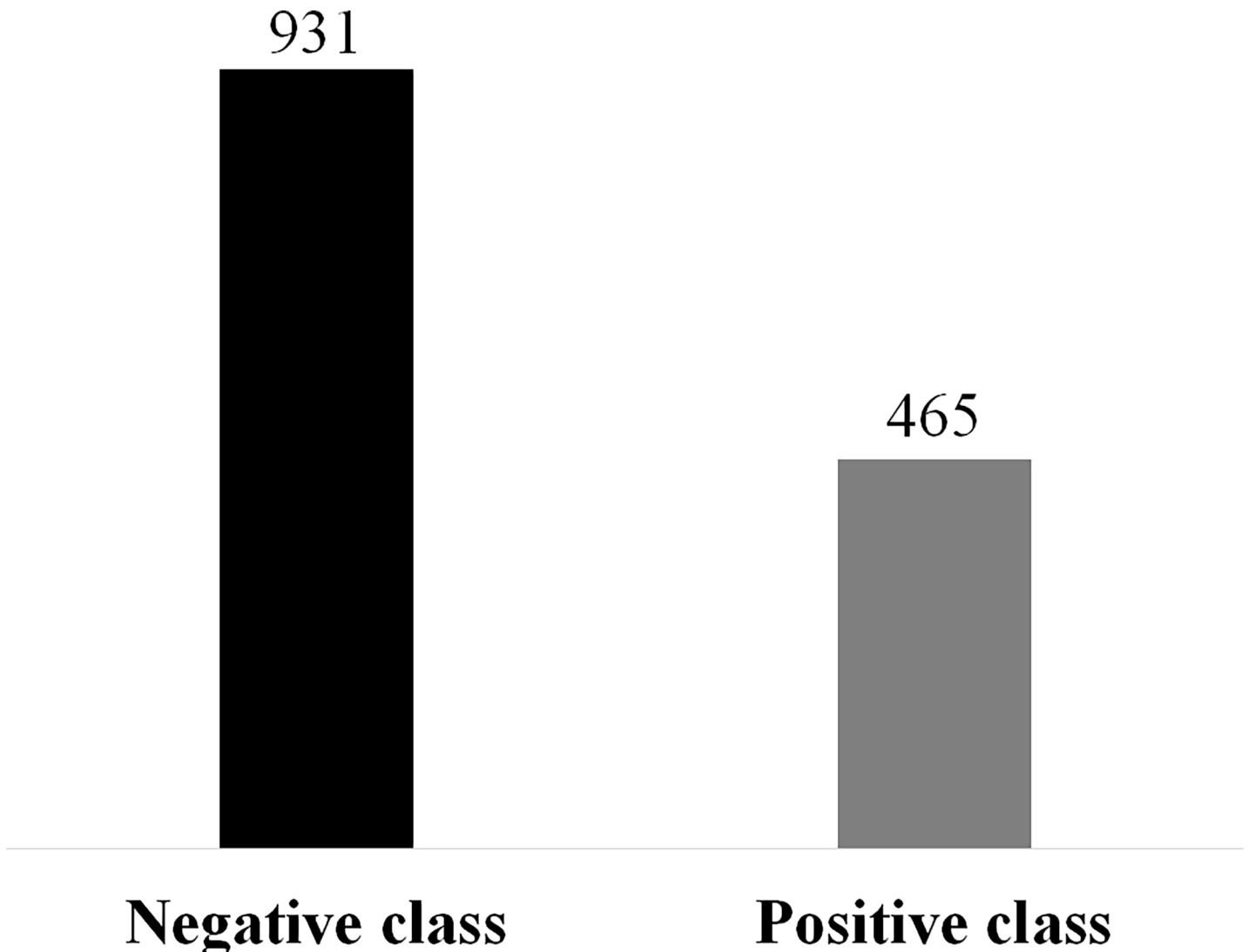


Fig 1. Discretization of CO₂ emission from the soil in the positive class (CO₂ emissions from high soil) and the negative classes (medium and low CO₂ emissions).

<https://doi.org/10.1371/journal.pone.0193537.g001>

The induction of the decision tree model resulted in the calculation of the known matrix of errors, or matrix of confusion (Table 3), widely used in statistical analysis of agreement [23].

Column 'Total' in Table 3 presents P as the total value of positive cases and N as the total of existing negative cases in the training set. In the Total, P' is the total number of cases that the model rated as positive cases and N' the total number of cases classified as negative. From the matrix of confusion, it is possible to extract the performance evaluation metrics. The rate of accuracy is the percentage of examples that were correctly classified by the classifier and can be expressed according to Eq 1.

$$\text{Accuracy} = (VP + VN)/(P + N) \quad (1)$$

To describe the measure of agreement between the predicted and observed classes, which deducts the expected number of correct answers (using a classification at random) of the actual number of the accuracy of the classifier, we used the Kappa measure (Eq 2). Kappa values vary from 0 to 1, representing bad and good classification results, respectively. The Kappa

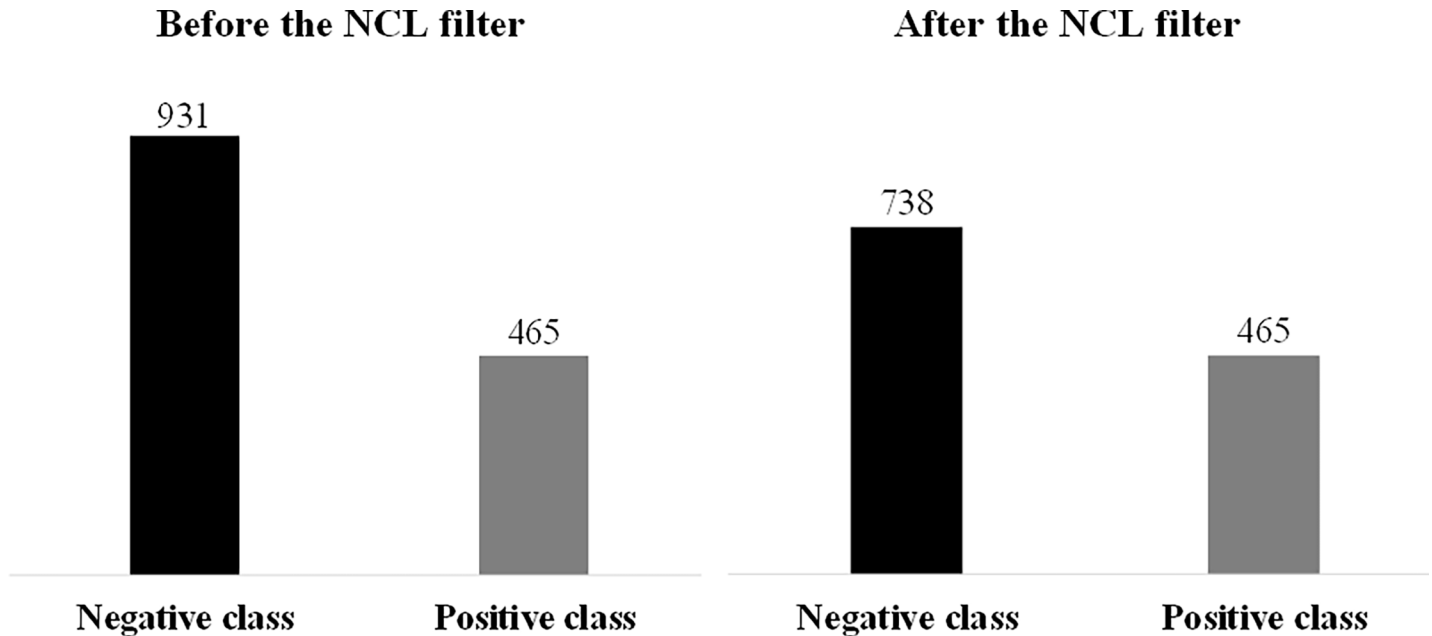


Fig 2. Demonstration of the classes distribution before and after the application of the NCL filter.

<https://doi.org/10.1371/journal.pone.0193537.g002>

coefficient can be defined by the following equation [23]:

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)} \tag{2}$$

where Pr (a) is the relative agreement observed for a given class in the matrix of confusion; Pr (e) is the probability of expected agreement for this same class.

The Kappa coefficient is calculated taking into account all classes. A possible interpretation of the performance of the models from this measure was introduced by Landis and Koch [24].

In order to compare the performance of the different classifiers associated with the attribute selection methods, the ROC curve was used, a graphical tool used to assess the algorithm using software Roc on 2.0.

The analysis of the curves generated by each classification model in the ROC space is represented by the ratio between the true positive rate (TPR) axis and the false positive rate (FPR). The classifiers positioned at the bottom of the "convex hull" have a precision rate lower than those contained in it. In addition, the larger the area under the curve, the greater the accuracy of the model. The point (0, 0) represents the strategy of never classifying an example as positive. Models that correspond to this point do not present any false positive, but are unable to classify any true positive either. The inverse strategy, of always classifying a new example as positive, is represented by the point (100%, 100%). The dot (0, 100%) represents the perfect

Table 3. A 2x2 matrix of confusion. TP = true positive; FP = false positive; FN = false negative; TN = true negative.

| | | PREDICT | | |
|-------|---------|-------------|-------------|-------|
| | | Class A | Class B | Total |
| TRUE | Class A | TP = (A, A) | FN = (A, B) | P |
| | Class B | FP = (B, A) | TN = (B, B) | N |
| Total | | P' | N' | P+N |

<https://doi.org/10.1371/journal.pone.0193537.t003>

model, i.e., all positive and negative examples are correctly classified. In contrast, the point (100%, 0) represents the model that always makes erroneous predictions [6].

Results

The behavior of soil CO₂ emission for each treatment along the experimental period is presented in Fig 3. It was possible to observe great variability of the CO₂ emission, mainly in the first days after the soil tillage practices and after rainfalls events, reaching emission values of up to 771 kg ha⁻¹ day⁻¹ (Fig 3A). Around the 40th day after tillage, there was a tendency of stabilization of soil CO₂ emissions, independent of rainfall events in the experimental area.

Analyzing the average soil CO₂ emissions, it was observed that the treatment where conservationist practices were used (cover crop and minimal tillage—MTCC) lower median values of CO₂ emission was observed compared to other treatments (Fig 3B), which presented average CO₂ emission of 63 kg ha⁻¹ day⁻¹ versus 74 kg ha⁻¹ day⁻¹ of the treatment CTCC, 71 kg ha⁻¹ day⁻¹ of the standard treatment CTWC, and finally 78 kg ha⁻¹ day⁻¹ of the MTWC.

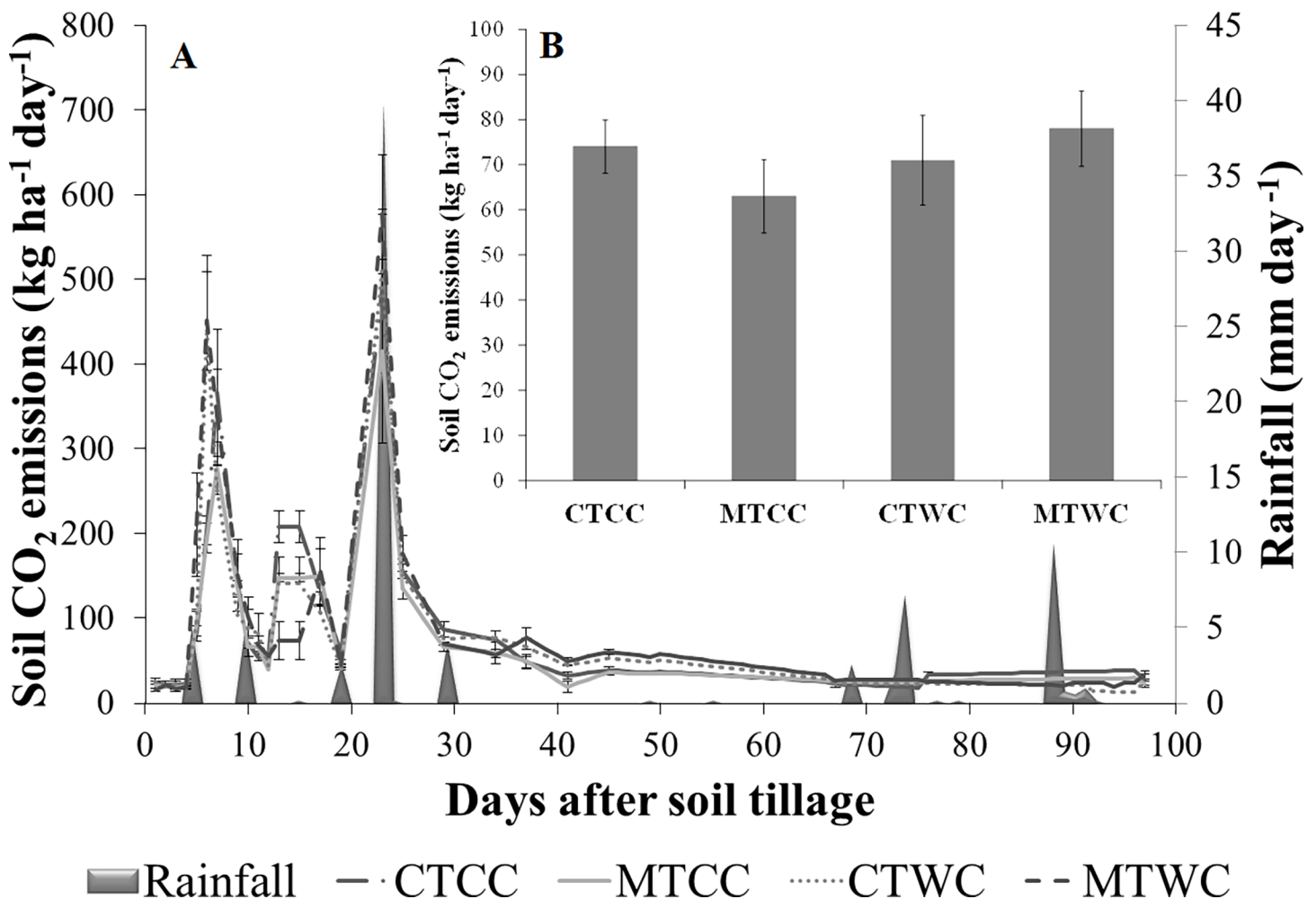


Fig 3. (A) Temporal variability of soil CO₂ emission and rainfall, and (B) average soil CO₂ emission with their respective standard deviations, in the different management systems used in the experimental area. CTCC = Conventional tillage with cover crop, MTCC = Minimal tillage with cover crop, CTWC = Conventional tillage without cover crop, MTWC = Minimal tillage without cover crop.

<https://doi.org/10.1371/journal.pone.0193537.g003>

Soil moisture was another attribute that presented high variability along the evaluated period. Soil moisture was influenced by rainfall events, where its values ranged from 4% to 36%. The soil temperature presented small variability among the treatments. In general, a tendency of small average temperature was observed in cover crop treatments (Fig 4).

In addition, it can be observed that the areas under cover crop cultivation presented higher porosity (macro, micro e TP), Al⁺³, K, P, Organic C e and lower Bd, PR, RT, Mg⁺². Specifically for the MTCC treatment, it was observed that in this treatment occurred greater Sm, Micro, TP, Al⁺³, K⁺, MBC associated with lower MDA, Bd, H + Al⁺³, Ca⁺², Mg⁺² (Figs 4 and 5).

According to the results presented in Table 4, the number of variables selected differed among the assessment methods, which has occurred since the selection of only one attribute, as in the case of the CFS method, having selected soil temperature exclusively, until the selection of ten variables, as in the case of method χ^2 (Table 4).

When analyzing the variables selected in the different approaches, we observed the recurrence of some of them, such as: soil temperature, rainfall, soil moisture, pH, and soil organic C (Table 4). Soil temperature appears in all attribute selection methods, followed by rainfall, present in six out of the seven approaches assessed, and the three variables, soil moisture, pH, and soil organic C, which appear in four of the seven methods tested.

Among the algorithms assessed for soil CO₂ emission classification, J48, Naïve Bayes and Bagging with logistic regression presented better performance when using the χ^2 method for attribute selection, with the accuracy rate of the different approaches ranging from 87.18 to 84.61 and Kappa coefficient from 0.73 to 0.66. In contrast, the SMO and Multilayer Perceptron algorithms presented a subset of variable selected through the Wrapper method with an accuracy rate varying from 89.10 to 85.90 and Kappa coefficient from 0.77 to 0.69 (Table 5).

The IV approach provided higher AUC and TP, being 89.90 and 94.2, respectively, followed by the V approach, which presented an AUC of 89.85 and a TP of 94.2. In contrast, the III approach provided the lowest AUC, 85.10 and TP, 82.7, against the remaining approaches (Table 6).

Regarding the analysis of the different approaches through the ROC curve, in general, all of them had a good performance regarding the classification, since none of them remained below the random performance line (Fig 6). In addition, we observed that the best approaches were IV and V for being the only ones to be encompassed by the convex hull. However, the IV approach has advantages over the V for being closer to the point (0.0 and 100).

Discussion

Effect of the cover plant and soil tillage to mitigate CO₂ emissions

In spite of the high variability of CO₂ emission, the extreme values were not considered as outliers, since they were associated by tillage operations and rainfall events in the experimental area.

The magnitude of CO₂ loss from soil influenced by tillage practices is highly related to the intensity of soil disturbance caused by tillage which alters soil organic matter decomposition environment due soil aeration, break the aggregates, incorporation of residue into the plow layer [25, 26].

Studies conducted under Brazilian conditions confirm this tends, i.e., Iamaguti [27], evaluated the CO₂ emissions influenced by three systems of soil tillage in sugarcane replanting area, observed that for the conventional tillage the exposed and disaggregated soil favors higher losses of CO₂. In a study on quantification the impact of sugarcane, harvest systems, straw and soil management on soil CO₂ emissions, Figueiredo [28], showed that intensive soil tillage and the incorporation of sugarcane straw into the soil increased short-term CO₂ emissions.

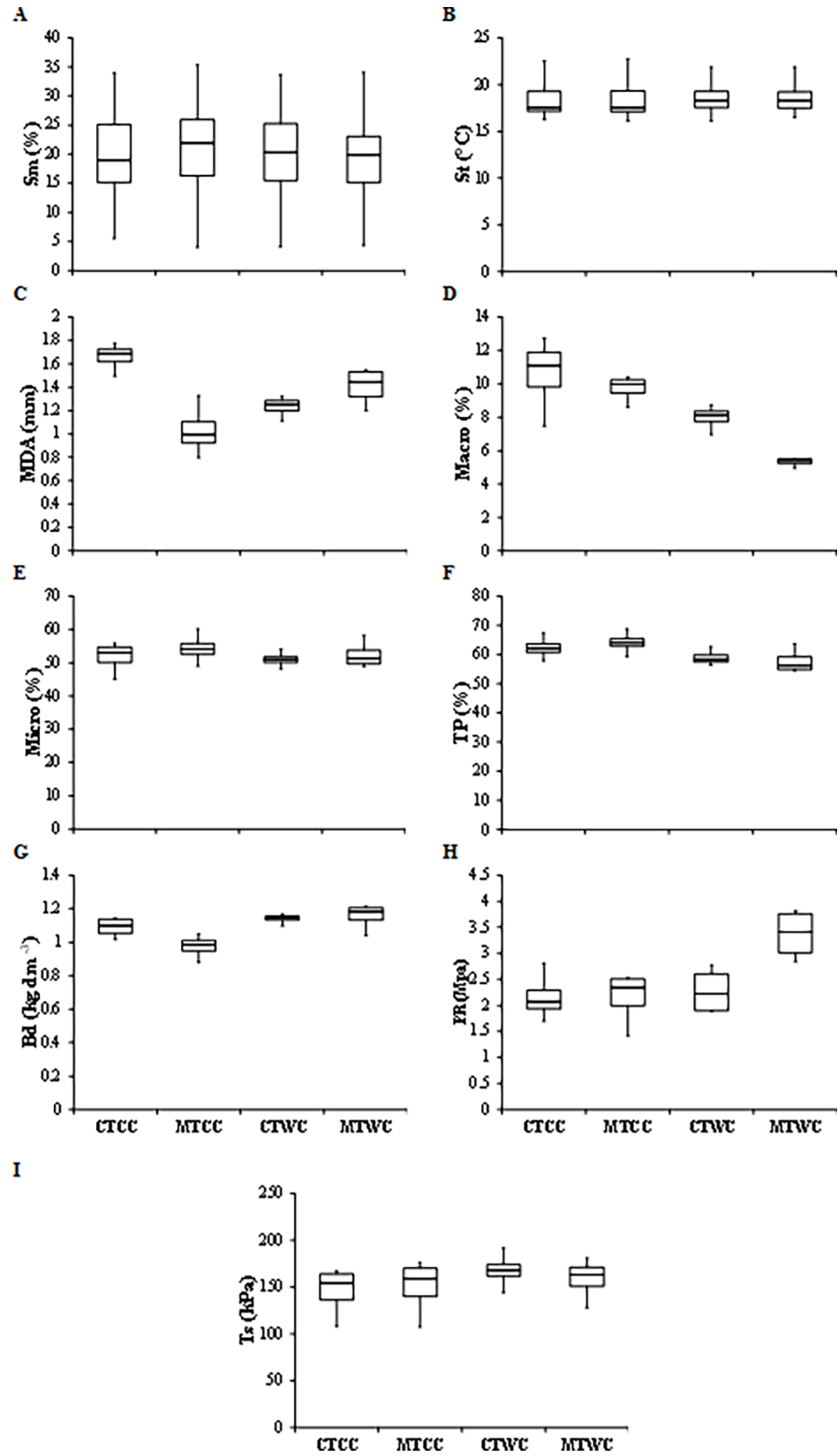


Fig 4. Box plot for the physical variables of the soil in the different management systems used in the experimental area. CTCC = Conventional tillage with cover crop, MTCC = Minimal tillage with cover crop, CTWC = Conventional tillage without cover crop, MTWC = Minimal tillage without cover crop. Sm = Soil moisture, St = Soil temperature, MDA = Mean diameter of the aggregate, Macro = Soil macroporosity, Micro = Soil microporosity, TP = Total porosity, Bd = Bulk density, PR = Penetration resistance, Ts = Tensile strength of the aggregate.

<https://doi.org/10.1371/journal.pone.0193537.g004>

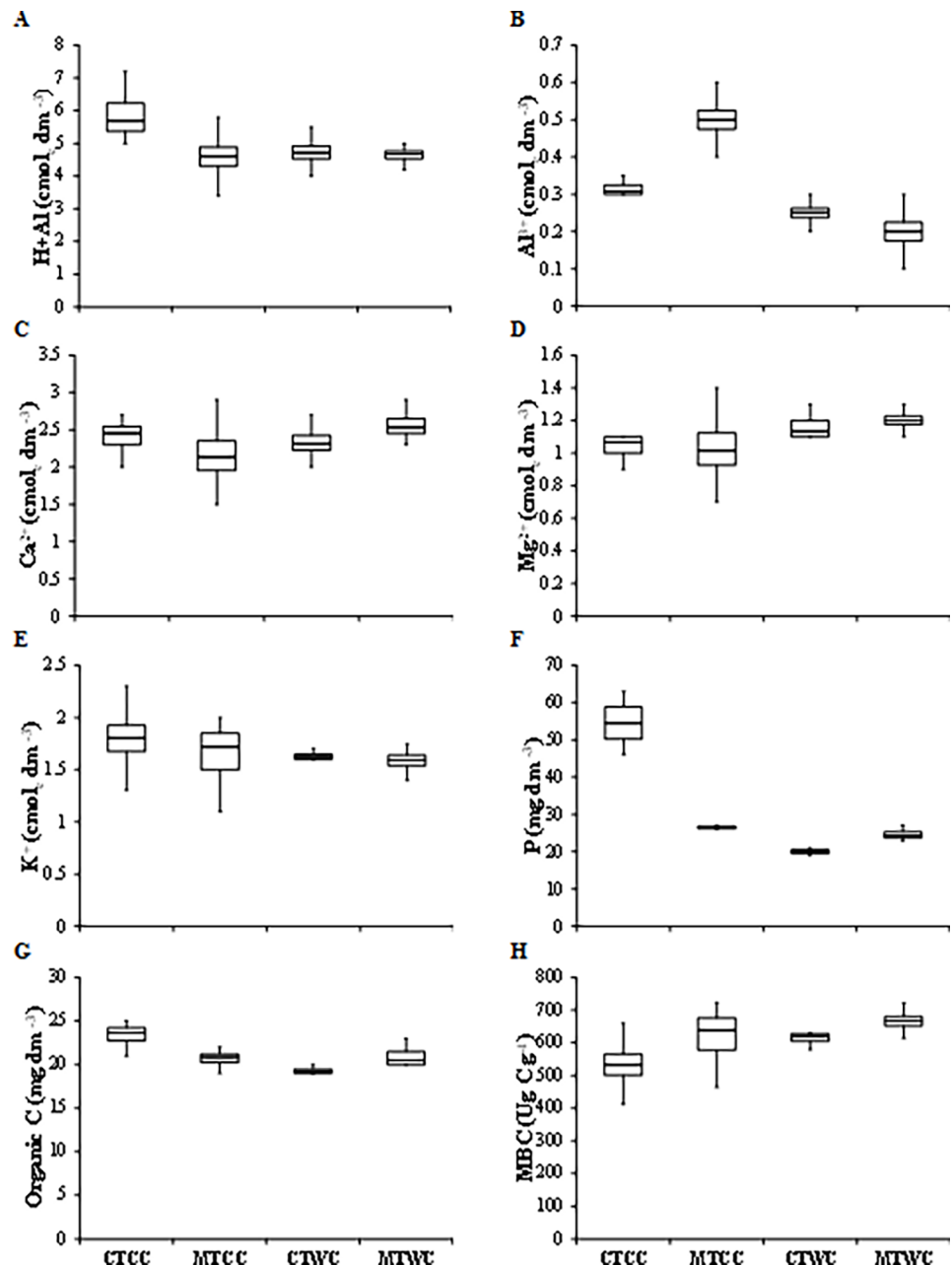


Fig 5. Box plot for soil chemical and biological variables in the different management systems used in the experimental area. CTCC = Conventional tillage with cover crop, MTCC = Minimal tillage with cover crop, CTWC = Conventional tillage without cover crop, MTWC = Minimal tillage without cover crop, H+Al = Acidity potential, Al⁺³ = Exchangeable aluminum, Ca⁺² = Exchangeable calcium, Mg⁺² = Exchangeable magnesium, K⁺ = Exchangeable potassium, P = Exchangeable phosphorus, Organic C = Organic Carbon, MBC = Microbial biomass carbon.

<https://doi.org/10.1371/journal.pone.0193537.g005>

Table 4. Variables selected through the different selection methods.

| Methods | Selected variables | Total |
|--------------------------|--|-------|
| χ^2 | Sm, St, rainfall, Macro, Tp, Bd, PR, pH, Al, P | 10 |
| CFS | St | 1 |
| Wrapper J48 | Sm, St, rainfall, Macro, Bd, Al, Ca, Organic C | 8 |
| Wrapper Naïve Bayes | St, rainfall, Macro, pH | 4 |
| Wrapper SMO | St, rainfall, MDA, Tp, Organic C | 5 |
| Wrapper Mult. Perceptron | St, Sm, pH, H+Al, P, Organic C | 6 |
| Wrapper Bagging Logistic | Sm, St, rainfall, MBC, Ts, pH, Mg, Organic C | 8 |

<https://doi.org/10.1371/journal.pone.0193537.t004>

Rainfall events, in turn, promoted the increase of soil moisture and stimulated CO₂ emissions, producing CO₂ emissions belonging to the high class. This fact is supported by several studies in sugarcane fields that demonstrate the influence of rainfalls events on soil CO₂ emissions due to changes in soil water content [28] [29] [30].

The soil moisture is an attribute essential for soil microbial metabolism, and can present both direct and indirect influence under CO₂ emissions. Adequate soil moisture tends to stimulate microbial activity in the soil rather than inhibit. According to Vicent [31] optimum soil moisture values range from 25% to 40%, above this range the CO₂ emission is limited by excess water and lack of oxygen in the soil, and below limits soil respiration by drought. In our study, the average value of soil moisture remained around 20%, whose maximum soil moisture value was 36%, not exceeded, therefore, values that could compromise the availability of O₂ in the pore space of the soil.

However, after the 40th day after soil tillage we observed a tendency to stabilize CO₂ emissions. The same occurred in other studies performed by Panosso [32] and La Scala [33]. This is probably associated to decrease of the labile C caused by tillage [34, 35].

Our results also showed that the association between the cover crop and minimum tillage were efficient methods to promote the mitigation of soil CO₂ emissions in sugarcane fields. These results are in agreement with Figueiredo [28], Tanveer [36] and Moitinho [37], that observed a lower CO₂ emission due to the reduction of the frequency of tillage and maintenance of crop residues on soil surface.

The identification of the factors influencing the emission of CO₂ from the soil in agricultural areas opens up opportunities for adopting practices that reduce the net emissions of this gas [27]. In a study on the effect of different soil tillage systems on CO₂ emissions, Lu [38], observed that the average loss of C in the form of CO₂ was significantly lower in the treatment with less soil disturbance (no-tillage) compared to conventional tillage. The authors pointed out that the lower soil temperature presented by no-tillage treatment reduces microbial activity and, consequently the emission of CO₂ in the soil. Another explanation for the lower CO₂ emissions in no-tillage treatment could be greater physical protection of the C inside the soil aggregates.

Table 5. The performance of different classifiers associated with the attribute selection methods assessed.

| Approaches | J48 | | Naive Bayes | | SMO | | Multilayer Perceptron | | Bagging Logistic | |
|------------------------------------|--------------|-------------|--------------|-------------|--------------|-------------|-----------------------|-------------|------------------|-------------|
| | Accuracy | Kappa | Accuracy | Kappa | Accuracy | Kappa | Accuracy | Kappa | Accuracy | Kappa |
| Without attribute selection | 87,18 | 0,72 | 84,62 | 0,66 | 83,97 | 0,65 | 86,54 | 0,72 | 84,62 | 0,67 |
| χ^2 | 87,18 | 0,73 | 84,61 | 0,66 | 84,63 | 0,66 | 87,18 | 0,73 | 84,62 | 0,68 |
| CFS | 84,62 | 0,68 | 82,69 | 0,64 | 83,97 | 0,67 | 83,97 | 0,67 | 82,69 | 0,65 |
| Wrapper | 85,90 | 0,70 | 82,05 | 0,60 | 85,90 | 0,69 | 89,10 | 0,77 | 83,97 | 0,67 |

<https://doi.org/10.1371/journal.pone.0193537.t005>

Table 6. Rate of false positive (FP), true positive (TP) area under the curve (AUC) for the best associations between classification and attribute selection method.

| Approaches | | FP Rate | TP Rate | AUC |
|------------|---------------------------------|---------|---------|-------|
| I | J48 x χ^2 | 15,4 | 92,3 | 89,44 |
| II | Naive Bayes x χ^2 | 12,5 | 78,8 | 89,16 |
| III | SMO x Wrapper | 12,5 | 82,7 | 85,10 |
| IV | Multilayer Perceptron x Wrapper | 13,5 | 94,2 | 89,90 |
| V | Bagging Logistic x χ^2 | 20,2 | 94,2 | 89,85 |

FP Rate = %; TP Rate = %; AUC = %.

<https://doi.org/10.1371/journal.pone.0193537.t006>

Moreover, the maintenance of crop residues on the soil surface in areas where soil disturbances are minimal, limits soil-residue contact, resulting in a reduced rate of decomposition, and consequently lower CO₂ emissions in these systems [25, 39].

The adoption of cover crop and minimum tillage (CTCC) provided desirable modifications in soil attributes, such as: greater Sm, Micro, TP, K⁺, MBC, and lower, Bd. In agreement, Torres [40], evaluated changes in the soil physical attributes with the use of different cover crops under no-tillage system, and observed positive changes in physical attributes in the soil surface

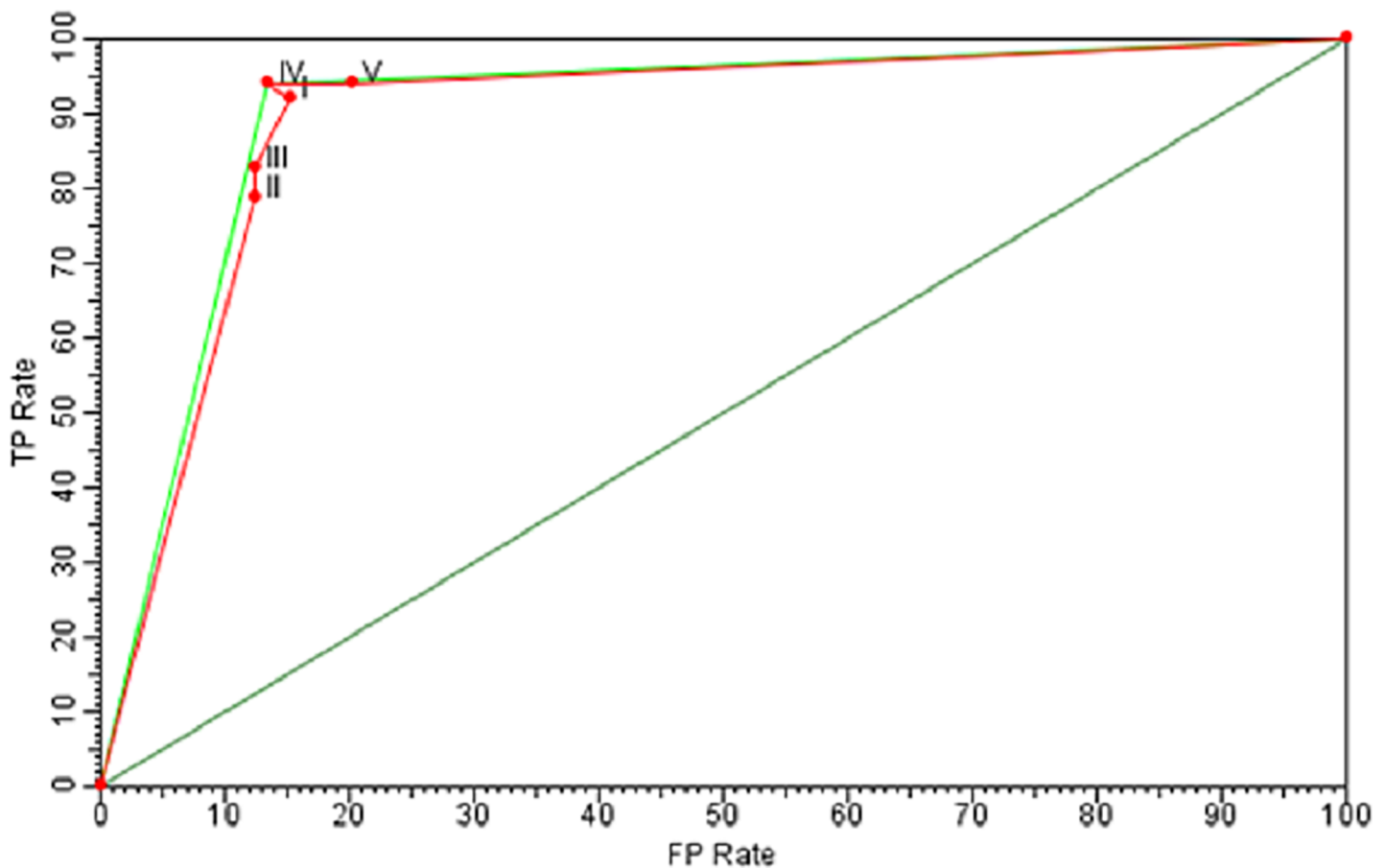


Fig 6. Performance in the ROC (Receiver Operating Characteristic) space of the different approaches used to classify soil CO₂ emission, considering the high CO₂ emission class as the "most important" class. I = J48 with χ^2 , II = Naïve Bayes with χ^2 , III = Bagging with logistic regression with χ^2 , IV = SMO using wrapper method, V = Multilayer Perceptron with wrapper method.

<https://doi.org/10.1371/journal.pone.0193537.g006>

layer. Similar results were also found by Lu [38], who observed significant changes in soil properties through the conversion of the conventional tillage system to no-tillage. However, it is important to highlight that our results represent a case study with one soil type and with specific climatic conditions, and more studies should be performed in order to analyze the results in different edaphoclimatic conditions.

Classification of CO₂ emissions

The variables of soil moisture, soil temperature, rainfall, pH, and soil organic C showed high predictive power regarding the classification of soil CO₂ emission and were frequently selected through the different methods of attribute selection assessed in this study. Several studies confirm the direct and/or indirect influence of these variables on the soil CO₂ emission.

For example, La Scala Junior [33], Corradi [41], Teixeira [42] found direct correlations between soil moisture and CO₂ emissions under different conditions. Iamaguti [27], Karhu [43] and Tavares [44] observed a significant correlation between soil temperature and soil CO₂ emission. Moitinho [29] and Silva-Olaya [30] observed higher CO₂ emissions along days with rainfall events. Fuentes [45] and Marcelo [46] reported that increasing soil pH affects the activity and microbial population in the soil and consequently their CO₂ emission. Reuss and Johnson [47] and Tossell [48], point out that carbonic acid, produced by the dissolution of CO₂ in water, is an important acidifying agent in natural systems. La Scala Junior [49] observed a linear correlation between CO₂ emission and Organic C.

According to Park [50], the closer the AUC is to the value 1, the better the overall performance of the test, which makes a test with AUC = 1 be regarded as perfectly accurate; therefore, the higher the AUC, the better the overall performance of the approach used. In this sense, the IV approach presented higher AUC, being superior to all of the remaining approaches assessed; however, the difference between IV and V, I and II approaches was very low, within the order of 0.05, 0.46 and 0.74, respectively.

In contrast, the analysis of the different approaches in the ROC space proved both the IV and the V as the most satisfactory for having been the only approaches to be encompassed by the convex hull. However, approach IV tends to have advantages over V since it is closer to the point (0.0 and 100%). Prati [6] point out that an optimal model should be as close as possible to the point (0 and 100%). In addition, Fawcett [51] refers to the ROC chart as a description of the relative compensation between the benefits (true positive) and costs (false positives) of a classification. Thus, a point located in the upper left corner demonstrates that a higher amount of positive and negative examples are classified correctly, consequently generating a lower classification cost, corroborating with the statement that the IV approach is the best one.

Through their research, several authors have found good performance both in the attribute selection Wrapper method and in the Multilayer Perceptron classifier. Liu and Yu [52] point out to a variety of available attribute selection algorithms; however, the attribute selection Wrapper method has generally performed better than other methods for selecting a subset of variables that are more suitable to the predetermined mining algorithm, but also tending to demand high computational effort and cost, especially in datasets with great dimensionality, which makes it possibly not suitable to some mining algorithms. In agreement with the results, Hall and Geoffrey [53], by comparing several methods of variables selection for the supervised classification, also concluded that the Wrapper method was the most satisfactory for the selection.

Finally, Freitas [54] investigated the use of the classifier Multilayer Perceptron for the predicting of the spatial variability of soil CO₂ emissions in sugarcane areas, also concluding that this approach satisfactorily met the expectations having proved good application potential for this type of issue.

Conclusions

The CO₂ emissions present high variability, mainly in the first days after the soil tillage, in which soil disturbance caused by tillage and rainfall events stimulate CO₂ production, while the association between cover crop and minimum tillage are effective strategies to promote the mitigation of soil CO₂ emissions.

Data mining techniques to predict soil CO₂ emission proved promising results for having allowed the development of several classifications approaches with high accuracy rate (above 80%), with variables such as soil moisture, soil temperature, rainfall, pH, and organic C showing high predictive power. Finally, among the approaches assessed in this study, the Multilayer Perceptron approach (a particular case of artificial neural networks) with the Wrapper attribute selection method tends to be more advantageous for its higher accuracy of the positive class with a consequent lower cost of classification.

Finally, we highlighted that the data presented herein represent a short-term case study conducted in specific edaphoclimatic conditions. Comprehensive studies evaluating the efficiency of the data mining techniques to predict the soil CO₂ emissions should be encouraged to test the reproducibility of these techniques under different soil and climate conditions.

Supporting information

S1 Table. Daily CO₂ emissions. Day = Days after soil tillage and replanting of sugarcane; Factor 1 = Cover crop; Factor 2 = Soil tillage; Block = Experimental design in randomized blocks; CO₂ = Soil CO₂ emissions.

(XLSX)

S2 Table. Soil physical attributes. Factor 1 = Cover crop; Factor 2 = Soil tillage; Block = Experimental design in randomized blocks; Sm = Soil moisture; St = Soil temperature; MDA = Mean diameter of the aggregate; Macro = Soil macroporosity; Micro = Soil microporosity; TP = Total porosity; Bd = Bulk density; PR = Penetration resistance; Ts = Tensile strength of the aggregate.

(XLSX)

S3 Table. Soil chemical attributes. Factor 1 = Cover crop; Factor 2 = Soil tillage; Block = Experimental design in randomized blocks; H+Al = Acidity potential; Al⁺³ = Exchangeable aluminum; Ca⁺² = Exchangeable calcium; Mg⁺² = Exchangeable magnesium; K⁺ = Exchangeable potassium; P = Exchangeable phosphorus; Organic C = Organic Carbon.

(XLSX)

S4 Table. Soil biological attributes. Factor 1 = Cover crop; Factor 2 = Soil tillage; Block = Experimental design in randomized blocks; MBC = Microbial biomass carbon.

(XLSX)

Author Contributions

Supervision: Zigomar Menezes de Souza.

Writing – original draft: Camila Viana Vieira Farhate, Stanley Robson de Medeiros Oliveira.

Writing – review & editing: Camila Viana Vieira Farhate, Zigomar Menezes de Souza, Stanley Robson de Medeiros Oliveira, Rose Luiza Moraes Tavares, João Luís Nunes Carvalho.

References

1. Paustian L, Babcock B, Hatfield JL, Lal R, Mccarl BA, Mclaughlin S, et al. Agricultural mitigation of greenhouse gases: science and policy options. In: Análisis Conference Proceedings, First National Conference on Carbon Sequestration. Washington, DC: Conference on Carbon Sequestration, 2016.
2. Rustad LE, Huntington TG, Boone RD. Controls on soil respiration: Implications for climate change. *Bio-geochemistry*, 2000; 48: 1–6.
3. Sierra CA, Trumbore SE, Davidson EA, Vicca S, Janssens I. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. *JAMES*. 2015; 7: 335–356. <https://doi.org/10.1002/2014MS000358>
4. Galvão ND, Marin HF. Data mining: a literature review. *Acta Paul Enferm*, 2009; 22: 686–90. <https://doi.org/10.1590/S0103-21002009000500014>
5. Han J, Kamber M, PEI J. Data mining: concepts and techniques. 3rd ed. Morgan Kaufmann Publishers, 2011.
6. Prati RC, Batista GEAPA, Monard MC. Curvas ROC para avaliação de classificadores. *IEEE América Latina*. 2008; 6:1–8.
7. Soil Survey Staff. Keys to Soil Taxonomy. 12st. USDA Natural Resources Conservation Service, 2014.
8. EMBRAPA. Centro Nacional de Pesquisa de Solos. Sistema brasileiro de classificação de solos. 3st ed. Rio de Janeiro. Centro Nacional de Pesquisa de Solos, 2013.
9. Zanatta JA, Alves BJR, Bayer C, Tomazi M, Marozzi AHB, Costa FFS, et al. Protocolo para medição de fluxos de gases de efeito estufa do solo. Embrapa Florestas, 2014
10. Elisabeth M., Claessen C. Manual de análise de solos. 2ed. Embrapa, Rio de Janeiro: 2011.
11. Stolf R. Teoria e teste experimental de fórmulas de transformação dos dados de penetrômetro de impacto em resistência do solo. *R. Bras. Ci. Solo*. 1991; 15: 229–235.
12. Kemper WD, Chepil WS. Size distribution of aggregates. In: Black CA. (Ed.). *Methods of soil analysis: physical and mineralogical properties, including statistics of measurement and sampling*. Part 1. Madison: American Society of Agronomy: 1965.
13. Dexter AR, Kroesbergen B. Methodology for determination of tensile strength of soil aggregates. *J Agr Eng Res*. 1985; 31: 139–147. [https://doi.org/10.1016/0021-8634\(85\)90066-6](https://doi.org/10.1016/0021-8634(85)90066-6)
14. Raij BV, Andrade JC, Cantarella H, Quaggio JA. Análise química para avaliação da fertilidade de solos tropicais. Instituto Agrônomo, Campinas: 2001.
15. Vance ED, Brookes PC, Jenkinson DS. Microbial biomass measurements in forest soils: The use chloroform fumigation-incubation method in strongly acid soils. *Soil Biol Biochem*. 1987; 19: 697–702. [https://doi.org/10.1016/0038-0717\(87\)90051-4](https://doi.org/10.1016/0038-0717(87)90051-4)
16. Mcgill R, Tukey JW, Larsen WA. Variations of box plots. *Am Stat*, 1978; 32: 12–16.
17. Hall MA. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand: 1998.
18. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell*. 1997; 97: 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)
19. Quinlan R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. 1993.
20. Duda R, Hart P. Pattern Classification and Scene Analysis. Wiley, New York. 1973.
21. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK, Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput*. 2001; 13: 637–649.
22. Breiman L. Bagging predictors. *JMLR*, 1996; 24:123–140.
23. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann; 2011.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33: 159–174. PMID: [843571](https://pubmed.ncbi.nlm.nih.gov/843571/)
25. Ussiri DAN, Lal R. Long-term tillage effects on soil carbon storage and carbon dioxide emissions in continuous corn cropping system from an alfisol in Ohio. *Soil Tillage Res*. 2009; 104: 39–47. <https://doi.org/10.1016/j.still.2008.11.008>
26. Lal R. Challenges and opportunities in soil organic matter research. *European Journal of Soil Science*, 2009; 60: 158–169. <https://doi.org/10.1111/j.13652389.2008.01114.x>
27. Iamaguti JL, Moitinho MR, Teixeira DDB, Bicalho ES, Panosso AR, La Scala N Junior. Preparo do solo e emissão de CO₂, temperatura e umidade do solo em área canavieira. *Rev. bras. eng. agríc. ambient*. 2015; 19: 497–504. <https://doi.org/10.1590/1807-1929/agriambi.v19n5p497-504>

28. Figueiredo EB, Panosso AR, Donald C, Reicosky DC, La Scala N Jr. 2015. Short-term CO₂-C emissions from soil prior to sugarcane (*Saccharum* spp.) replanting in southern Brazil. *GCB Bioenergy*. 7: 316–327.
29. Moitinho RM, Padovan MPP, Panosso AR, Teixeira DB, Ferraudo AS, La Scala N Junior. On the spatial and temporal dependence of CO₂ emission on soil properties in sugarcane (*Saccharum* spp.) production. *Soil Till Res*. 2015; 148: 127–132. <https://doi.org/10.1016/j.still.2014.12.012>
30. Silva-Olaya AM, Cerri CEP, La Scala N Junior, Dias CTS, Cerri CC. Carbon dioxide emissions under different soil tillage systems in mechanically harvested sugar cane. *Environ Res Lett*, 2012; 8: 1–8. <https://doi.org/10.1088/1748-9326/8/1/015014>
31. Vicent G, Shahriari AR, Lucot E, Badot P, Epron B D. Spatial and seasonal variations in soil respiration in a temperate deciduous forest with fluctuating water table. *Soil Biol Biochem*, 2006; 38: 2527–2535. <https://doi.org/10.1016/j.soilbio.2006.03.009>
32. Panosso AR, Camara FT, Lopes A, Pereira GT, La Scala N. JR. Emissão de CO₂ em um latossolo após preparo convencional e reduzido em períodos seco e chuvoso. *Científica*, 2006; 34: 257–262. <https://doi.org/10.15361/1984-5529.2006v34n2p257+-+262>
33. La Scala N Junior, Bolonhezi D, Pereira GT. Short-term soil CO₂ emission after conventional and reduced tillage of a no-till sugar cane area in southern Brazil. *Soil Tillage Res*. 2006; 91: 244–248. <https://doi.org/10.1016/j.still.2005.11.012>
34. Six J, Frey SD, Thies RK, Batten KM. Bacterial and fungal contributions to carbon sequestration in agroecosystems. *Soil Sci Soc Am J*, 2006; 70: 555–569. <https://doi.org/10.2136/sssaj2004.0347>
35. Teixeira LG, Lopes A, La Scala N. JR Temporal variability of soil CO₂ emission after conventional and reduced tillage described by an exponential decay in time model. *Eng Agric*, 2010; 30: 224–231. <https://doi.org/10.1590/S0100-69162010000200004>
36. Tanveer SK, Wen X, Lu XL, Zhang J, Liao Y. Tillage, Mulch and N Fertilizer Affect Emissions of CO₂ under the Rain Fed Condition. *Plos one*; 2013; 8: 1–16. <https://doi.org/10.1371/journal.pone.0072140> PMID: 24086256
37. Moitinho MR, Padovan MP, Panosso AR, La Scala N. JR 2013. Effect of soil tillage and sugarcane trash on CO₂ emission. *R Bras Ci Solo*, 2013; 37: 1720–1728, 2013. <https://doi.org/10.1590/S0100-06832013000600028>
38. Lu X, Lu X, Tanveer SK, Wen X, Liao Y. Effects of tillage management on soil CO₂ emission and wheat yield under rain-fed conditions. *Soil Research*, 2015; 54: 38–48. <https://doi.org/10.1071/SR14300>
39. Fuentes M, Hidalgo H, Etchevers J, León F, Guerrero A, Dendooven L, et al. Conservation agriculture, increased organic carbon in the top-soil macro-aggregates and reduced soil CO₂ emissions. *Plant soil*, 2012; 355: 183–197. <https://doi.org/10.1007/s11104-011-1092-4>
40. Torres JLR, Pereira MG, Assis RL, Souza ZM. Physical properties of an oxisol with cover crops in no-tillage. *R. Bras. Ci. Solo*, 2015; 39: 428–437. <https://doi.org/10.1590/01000683rbc20140597>
41. Corradi MM, Panosso AR, Martins Filho MV, La Scala N. Crop residues on short-term CO₂ emissions in sugarcane production areas. *Eng agric*. 2013; 33: 699–708, 2013. <https://doi.org/10.1590/S0100-69162013000400009>
42. Teixeira LG, Corradi MM, Fukuda A, Panosso AR, Reicosky D, Afonso Lopes A, et al. Soil and crop residue C-CO₂ emission under tillage systems in sugarcane-producing areas of southern Brazil. *Sci Agric*. 2013; 70: 327–335. <https://doi.org/10.1590/S0103-90162013000500007>
43. Karhu K, Auffret MD, Dungait JAJ, Hopkins DW, Prosser JI, Singh BK, et al. Temperature sensitivity of soil respiration rates enhanced by microbial community response. *Nature*, 2014; 513: 81–84. <https://doi.org/10.1038/nature13604> PMID: 25186902
44. Tavares RLM, Siqueira DS, Panosso AR, Castioni GAF, Souza ZM, La Scala N Junior. Soil management of sugarcane fields affecting CO₂ fluxes. *Sci Agric*. 2016; 73: 543–551. <https://doi.org/10.1590/0103-9016-2015-0174>
45. Fuentes JP, Bezdicsek DF, Flury M, Albrecht S, Smith JL Microbial activity affected by lime in a long-term no-till soil. *Soil Tillage Res*. 2006; 88: 123–131. <https://doi.org/10.1016/j.still.2005.05.001>
46. Marcelo AV, Corá JE, La Scala Júnior N. Influence of liming on residual soil respiration and chemical properties in a tropical no-tillage system. *R. Bras. Ci. Solo*. 2012; 36: 45–50. <https://doi.org/10.1590/S0100-06832012000100005>
47. Reuss JODW, Johnson DW. 1986. Acid Deposition and the Acidification of Soils and Waters. *Ecological Studies*. Springer Science and Business Media, LLC.
48. Tossell JA. H₂CO₃ (s): A New Candidate for CO₂ Capture and Sequestration. *Environ Sci Technol*, 2009; 43: 2575–2580. PMID: 19452919

49. La Scala N Junior, Marques NJ Junior, Pereira GT, Cora JE. Carbon dioxide emission related to chemical properties of a tropical bare soil. *Soil Biol Bioche*. 2000; 32: 1469–1473. [https://doi.org/10.1016/S0038-0717\(00\)00053-5](https://doi.org/10.1016/S0038-0717(00)00053-5)
50. Park SH, Goo JM, Jo CH. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J Radiol*. 2004; 5: 11–18. <https://doi.org/10.3348/kjr.2004.5.1.11> PMID: 15064554
51. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006; 27: 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
52. Liu H, Yu L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE transactions on knowledge and data engineering*. 2005; 17; 491–502. <https://doi.org/10.1109/TKDE.2005.66>
53. Hall MAE, Geoffrey Holmes G. 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE T Knowl Data En*, 15 (6), 1437–1447. <https://doi.org/10.1109/TKDE.2003.1245283>
54. Freitas LPS, Lopes MLM, Panosso AR, La Scala Júnior N, Freitas RLB, Minussi CR, et al. Previsão da variabilidade espacial da emissão de CO₂ do solo em áreas de cana-de-açúcar utilizando redes neurais artificiais. Em: *Anais do XIV International Conference on Engineering and Technology Education*, Salvador, 2016.