Soils and Plant Nutrition | Research Article

SCIENTIA AGRICOLA

DOI: http://dx.doi.org/10.1590/1678-992X-2016-0473

ISSN 1678-992X

Classification of soil respiration in areas of sugarcane renewal using decision tree

Camila Viana Vieira Farhate[1]*, Zigomar Menezes de Souza[1], Stanley Robson de Medeiros Oliveira[1,2], João Luís Nunes Carvalho[3], Newton La Scala Júnior[4], Ana Paula Guimarães Santos[1]

[1]University of Campinas/FEAGRI, Av. Cândido Rondon, 501 – 13083-875 – Campinas, SP – Brazil.
[2]Embrapa Agricultural Informatics – Computational Intelligence Lab., Av. André Tosello, 209 – 13083-886 – Campinas, SP – Brazil.
[3]Brazilian Center for Research in Energy and Materials – Brazilian Bioethanol Science and Technology Lab., R. Giuseppe Máximo Scolfaro, 10000 – 13083-100 – Campinas, SP – Brazil.
[4]São Paulo State University – Dept. of Exact Sciences, Via de Acesso Prof. Paulo Donato Castellane, s/n – 14884-900 – Jaboticabal, SP – Brazil.
*Corresponding author <camila.vieira@feagri.unicamp.br>

Edited by: Thomas Kumke

Received November 28, 2016
Accepted February 08, 2017

ABSTRACT: The use of data mining is a promising alternative to predict soil respiration from correlated variables. Our objective was to build a model using variable selection and decision tree induction to predict different levels of soil respiration, taking into account physical, chemical and microbiological variables of soil as well as precipitation in renewal of sugarcane areas. The original dataset was composed of 19 variables (18 independent variables and one dependent (or response) variable). The variable-target refers to soil respiration as the target classification. Due to a large number of variables, a procedure for variable selection was conducted to remove those with low correlation with the variable-target. For that purpose, four approaches of variable selection were evaluated: no variable selection, correlation-based feature selection (CFS), chi-square method (χ^2) and Wrapper. To classify soil respiration, we used the decision tree induction technique available in the Weka software package. Our results showed that data mining techniques allow the development of a model for soil respiration classification with accuracy of 81 %, resulting in a knowledge base composed of 27 rules for prediction of soil respiration. In particular, the wrapper method for variable selection identified a subset of only five variables out of 18 available in the original dataset, and they had the following order of influence in determining soil respiration: soil temperature > precipitation > macroporosity > soil moisture > potential acidity.

Keywords: soil CO^2 emission, data mining, variable selection, soil temperature, soil organic matter

Introduction

Soil respiration is defined as carbon dioxide (CO_2) released from the soil surface into the atmosphere through the combined activity of the roots and decomposing organisms of soil organic matter (Stockmann et al., 2013). It is a significant component in the global carbon balance (C), since it is the main contributor in the transmission of C from the pedosphere into the atmosphere (Vicca et al., 2014).

Moisture and temperature have a strong influence on soil respiration (Fang and Moncrieff, 2001), as well as precipitation (Moitinho et al., 2015), soil drainage (Lal, 2004), availability and quality of organic matter (Campos et al., 2011), soil texture (Dilustro et al., 2005), topography (Brito et al., 2009), soil preparation system (Lal, 2004), application of limestone (Marcelo et al., 2012), porosity, pH (Orchard and Cook, 1983) and erosion processes (Lal, 2001).

Currently, an ultimate objective in carbon cycle science is to understand the spatial and temporal controls of CO_2 dynamics in terrestrial ecosystems (Leon et al., 2014). Thus, a more detailed understanding of the components of the global C balance allows to identify sources and drains of C and develop strategies to mitigate the risks of climate change (Lal, 2004). However, estimating sequestration or soil respiration into the atmosphere in production systems is difficult and complex due to the diversity of agricultural practices in large areas and significant variations of both soil and climate (Smith et al., 2010).

In this sense, data mining, defined by Berry and Linoff (1997) as the use and analysis of large amounts of data to discover significant patterns and rules, becomes an interesting tool to estimate different levels of soil respiration from related variables. Among the data mining techniques, we highlight the algorithms for decision trees (Monard and Baranauskas et al., 2002), mainly due to the interpretative potential of the symbolic model generated (Han et al., 2011).

Decision tree is a supervised learning method that provides a model represented graphically by nodes and branches, similar to a tree, but in the inverted sense. The decision nodes are: root node, located at the structure top, and the internal nodes, which contain a value test on one of the most relevant attributes where the results of these tests form the branches. The leaf-knots refer to the classes of the response variable and represent the result of the prediction obtained by the model (Witten et al., 2011). This study is based on the hypothesis that data mining techniques are efficient to generate models to predict soil respiration through related variables. Therefore, our objective was to build a model using variable selection and decision tree induction to predict different levels of soil respiration, taking into account physical, chemical and microbiological variables of soil and precipitation in renewal of sugarcane areas.

Materials and Methods

The experiment was conducted in a sugarcane crop belonging to the Iracema sugarcane Mill, located in the municipality of Iracemápolis, São Paulo, Brazil (22°34′50″ S, 47°31′07″ W; 608 m above sea level). The climate is CWa, humid temperate climate, with dry winter and hot summer according to the Köppen and Geinger classification (Köppen and Geinger, 1928). In the region, the average annual precipitation is 1294 mm and relative air temperature is 20.4 °C.

The soil under investigation is classified as Rhodic Hapludox, according to the Soil Taxonomy System (Soil Survey Staff, 2014), and is described as a Latossolo Vermelho Eutroférrico, according to the Brazilian Soil Classification, with very clayey texture. In the site characterization, the soil presented a pH around 4.6, an average organic carbon content of 10.5 g C dm^{-3}, base saturation of 50 %, cation exchange capacity 10.25 $cmol_c$ dm^{-3}, N-NH_4 content of 3.21 mg kg^{-1}, N-NO_3 content of 15.10 mg kg^{-1}, from 0-40 cm depth. The distribution of particle sizes of sand, silt, and clay were 140 g kg^{-1}, 194 g kg^{-1} and 666 g kg^{-1} respectively.

The experimental design was randomized blocks with a layout of plots subdivided into four replications. Each plot was split into an area with and without crop rotation, and each subplot received one of two types of soil tillage (minimum tillage and conventional tillage). Each experimental unit comprised 15 rows of sugarcane, at 1.5 m spacing and 34 m in length.

The treatments in this study are detailed as follows: Management system: i) area without crop rotation (WR); ii) area with crop rotation (CR) that used sunn hemp during the renewal of the sugarcane crop.

Soil Tillage systems: i) conventional tillage (CT) in the form of subsoiling, two harrowing procedures, and furrowing; ii) minimum tillage (MT) which consisted only of subsoiling and furrowing. For both treatments, we adopted conventional traffic held at the mill, even during harvesting (Figure 1).

The soil location of the experimental area has been used for sugarcane production in the last 70 years; however, the experiment began at the time of sugarcane crop renewal. The experiment was set up initially by mechanical elimination of ratoons followed by subsoiling into a 0.40 m depth across the entire area. Subsequently, sunn hemp was line-planted by distributing 25 kg ha^{-1} of seeds at a 0.5 m spacing during the recommended period of leguminous planting.

During this period, the area without crop rotation remained fallow, with only the presence of spontaneous sugarcane plants and weeds. At the end of the crop cycle, four metal frames covering 1 m^2 were randomly released to evaluate dry matter production of sunn hemp; 8 t ha^{-1} of sunn hemp dry matter were produced. In Apr 2013, the sunn hemp was desiccated.

Subsequently, the conventional tillage plots were prepared with two light harrowing procedures at a 0.40 m depth. For plots with minimum cultivation, only one furrowing was applied at a 0.40 m depth using the same equipment as applied to the conventional tillage. Planting was done manually, where the stalks were distributed in rows, cut into smaller pieces and later covered mechanically. At that time, it was applied 0.5 L ha^{-1} of fungicide and 250 g ha^{-1} of insecticide and planting fertilization performed with 300 kg ha^{-1} formulated as 12-23-23. Sugarcane grown in the experimental area was the variety RB96 6928 recommended for environments of high potential.

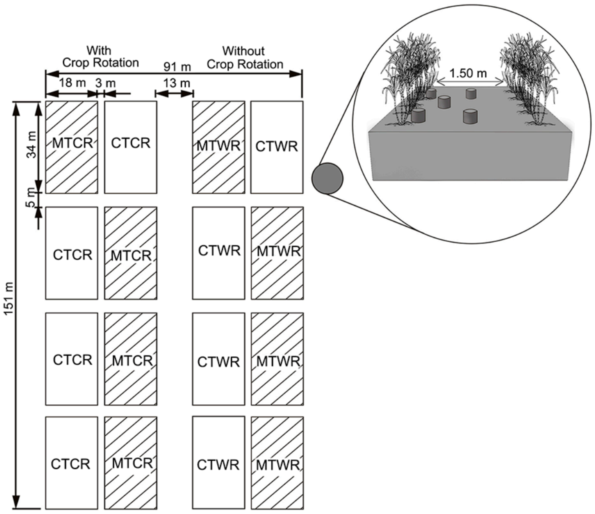

Figure 1 – Sketch of the data collection of soil respiration. CTCR = conventional tillage crop rotation; CTWR = conventional tillage without crop rotation; MTCR = minimum tillage with crop rotation; MTWR = minimum tillage without crop rotation.

Analysis of soil respiration, temperature and moisture

Temporal variability of soil respiration was evaluated using a portable system. This system consists of a closed chamber connected to PVC collars previously installed in the soil, with an internal volume of 991 cm^3 and a soil contact area of 71.6 cm^2. In the measurement mode, CO_2 concentration of the air inside the chamber is continuously measured by optical absorption spectroscopy in the infrared region - IRGA (Infra-red Gas Analyzer).

Evaluations were made in five PVC collars inserted 3 cm into the soil (diameter = 10 cm and height = 7 cm), distributed in 16 experimental plots (80 collars total). The collars were inserted 24 h before the first measurement, after tillage, in order to minimize disturbances due to placing the chamber into the soil. The measurements were always performed in the morning between 8h00 and 10h00 a.m., and each reading lasted 1 min and 30 s. Readings were performed daily in the first days after soil tillage. After the stabilization of CO_2 emissions, the readings were held weekly and later biweekly, and were conducted for 97 d after soil tillage in order to plant sugarcane. Soil respiration recorded in each of the five collars of each experimental unit (2 lines and 3 between the lines) was aggregated into a single measurement by the weighted average (assuming an area of 27 % for the line and 73 % between the lines). The daily soil respiration during the study period was estimated by the integral area method under the curve versus time.

Measurements the soil temperature and soil moisture were taken at the exact same time of the soil respiration measurements. Soil temperature (St) was evaluated at all points studied with the sensor, which is part of the portable CO_2 analysis system. This sensor consisted 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 CO_2 concentration by the TDR (Time Domain Reflectometer). The probe Thetaprobe ML2 is a tool that directly measures soil water content, corresponding to the volumetric moisture content, using the principle of wave generation that releases an electromagnetic pulse to a set of rods with reflection measured in the time domain (TDR).

After collecting CO_2, disturbed and undisturbed soil samples, with aluminum rings to determine soil chemical and physical variables, were collected at depths of 0.00-0.10 m, 0.10-0.20 m, 0.20-0.30 m and 0.30-0.40 m.

Soil physical variables

Total soil total porosity (TP), macroporosity (Macro), microporosity (Micro), and bulk density (Bd) were determined according to Brazilian Agricultural Research Corporation methodologies (Donagema et al., 2011). The resistant to penetration (RP) was obtained by the Stolf formula (1991) and water content in soil was determined by the gravimetric method in disturbed samples (Donagema et al., 2011). The mean diameter of the aggregates (MDA) was determined according to the method described by Kemper and Chepil (1965) and the calculation of the aggregate tensile strength (Ts) was performed as described by Dexter and Kroesbergen (1985).

Soil chemical variables

All samples were taken to the laboratory, air dried and then passed through a 2.0 mm mesh. It was determined soil pH ($CaCl_2$ 0.01 mol L^{-1}), exchangeable cations (Ca^{2+}, Mg^{2+} and K^+), phosphorus available in resin (P), organic carbon (wet oxidation), acidity potential, CTC potential and base saturation in accordance to the methodology proposed by Raij et al. (2001).

Soil microbiological variable

For the analysis of this variable, undisturbed soil samples were collected at the same depths for the chemical and physical variable (0.00-0.10 m, 0.10-0.20 m, 0.20-0.30 m and 0.30-0.40 m). The samples were collected and placed in a cooler and refrigerated during transportation to the laboratory where they were preserved in cold storage at 4 °C until the time of analysis. The microbial biomass carbon (MBC) was determined by the fumigation-extraction method proposed by Vance et al. (1987).

Data analysis and modeling

The original dataset was composed of 19 variables (one dependent (or response) variable and 18 independent (or explanatory) variables) (Table 1) that were added to the dataset totaling 1,552 observations. The variable-target refers to soil respiration and is the classification target.

The 18 independent variables were formed by nine soil physical variables (Sm, Ts, MDA, Macro, Micro, TP, Bd, RP Ts), eleven soil chemical variables (H + Al, Al^{3+}, Ca^{2+}, Mg^{2+}, K^+, P, C-org), a soil microbial variable (MBC) and a climatic variable (daily precipitation obtained at the weather station located at the sugarcane mill) (Table 1).

To identify different soil respiration levels, a discretization of the variable-target into categories was necessary, that is, continuous data was transformed into discrete data (intervals).

The categorization of the response variable intervals simplifies information, which facilitates the interpretation and decision making when analyzing a decision tree. However, it was not possible to produce rankings for different soil respiration levels in analyzing the literature, since this information is scarce or nonexistent. For that purpose, soil respiration values (g m^{-2} h^{-1}), present in the database, were ordered increasingly and equally divided into three soil respiration classes: low, medium and high (Table 2).

Table 1 – Description of 18 independent variables (physical, chemicals, microbiological and climatic) used in the composition of the database aimed at decision tree induction to predict different levels of soil respiration.

Variable		Description	Unit
Physical	Sm	Soil moisture	%
	St	Soil temperature	°C
	MDA	Mean diameter of the aggregate	mm
	Macro	Soil macroporosity	m^3 m^{-3}
	Micro	Soil microporosity	m^3 m^{-3}
	TP	Total porosity	m^3 m^{-3}
	Bd	Bulk density	kg dm^{-3}
	PR	Penetration resistance	Mpa
	Ts	Tensile strength of the aggregate	kPa
Chemical	H+Al	Acidity potential	$cmol_c$ dm^{-3}
	Al^{3+}	Exchangeable aluminum	$cmol_c$ dm^{-3}
	Ca^{2+}	Exchangeable calcium	$cmol_c$ dm^{-3}
	Mg^{2+}	Exchangeable magnesium	$cmol_c$ dm^{-3}
	K^+	Exchangeable potassium	$cmol_c$ dm^{-3}
	P	Exchangeable phosphorus	mg dm^{-3}
	C-org	Organic carbon	mg dm^{-3}
Microbiological	MBC	Microbial biomass carbon	µg C g^{-1} d^{-1}
Climatic	P	Precipitation	mm d^{-1}

Variable selection

To select the variables to be included in each repetition, the most common methods are based on information theory, such as information gain, which represents the expected reduction in entropy caused by partitioning the examples according to a variable (Han et al., 2011).

Entropy measures the amount of information carried by a variable, thereby characterizing the randomness or uncertainty of a set of examples (Shannon, 1949). Thus, given a set of examples for the relative target variable S of interest, and a C categorization of that variable in n classes S1, S2, ..., Sn, the entropy H (S) is defined by Equation 1, as follows:

$$H(S) = \sum_{i=1}^{n} -(pi).\log(pi) \qquad (1)$$

where: *pi* is the proportion of favorable cases of *Si* class.

Due to a large number of variables generated in the data preprocessing, a variable selection procedure was used to remove variables with low correlation values with respect to the target variable. Thus, four approaches for variable selection were evaluated: (i) No variable selection in which the use of all variables occurred, characterized by the absence of selection; (ii) Correlation-based feature selection (CFS), which searches the set of correlated variables in order to prevent reuse of the same information; (iii) The chi-square method (χ^2) is based on the concept of statistical independence; (iv) The Wrapper approach occurs in conjunction with the basic learning algorithm. In other words, this method creates a subset of variables, which is tested by the learning algorithm of interest. This process is repeated for each subset of variables until the given stopping criterion is reached. This approach evaluates variables using precision estimates provided by predetermined learning algorithms (Kohavi and George, 1997).

Induction and validation of the classification model (decision tree)

The induced models, with the variation of the number of objects (instances or observations) per leaf, were evaluated using the cross-validation method in 10 folds and through the following metrics: (i) accuracy rate; (ii) number of leaves (number of rules) generated, which are generally associated with model interpretability; (iii) Kappa coefficient that measures the agreement between observed and predicted classes of the classifier; (iv) precision per class, which can be thought of as a measurement of exactness, i.e., percentage of instances labeled as low, medium and high are actually such.

The difference between precision and accuracy is that the former is a description of random errors, a measurement of statistical variability, and the latter is a description of systematic errors, a measurement of statistical bias.

As a result of induction of the decision tree model, the known matrix of errors or matrix of agreements is calculated (Table 3), widely used in statistical analysis of agreement (Han et al., 2011).

In the *Total* column of Table 3, *P* is the total value of positive cases and *N* is 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 and *N*' the total number of cases classified as negative. The matrix of agreements allows 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 as in Equation 2.

$$Accuracy = (TP + TN)/(P+N) \qquad (2)$$

To describe the agreement measured between 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, the Kappa measurement is used (Equation 3). Their values vary from 0 to 1, representing bad and good classification results, respectively. The Kappa coefficient can be defined by the following equation (Witten et al., 2011):

$$K = Pr\ (a) - Pr\ (e)/1 - Pr\ (e) \qquad (3)$$

where: *Pr* (*a*) is the relative agreement observed for a given class in the matrix of agreements; *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 models performance from this measure was introduced by Landis and Koch (1977).

Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct and can be calculated using the equation (4):

$$P = TP/(TP + FP) \qquad (4)$$

Table 2 – Distribution of soil respiration (g $m^{2-1}h^{-1}$) according to low, medium and high classes and their limits, aimed at decision tree induction to predict different levels of Soil respiration.

Class	Limit
Low	[0.016; 0.124]
Medium	[0.125; 0.223]
High	[0.224; 3.21]

Table 3 – A 2 × 2 matrix of agreements. TP = true positives; FP = false positives; FN = false negative; TN = true negatives.

		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

The respective values of TP, FN, FP and TN for this study are described in Table 9.

For data classification, the decision tree method was used, available in the Weka 3.6 software. J48 was the induction algorithm used, widely known as C4.5, developed by Quinlan (1993). C4.5 builds decision trees from a set of training data using the concept of entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. Seeking to minimize a possible overfitting effect, the decision-tree pruning technique was performed to reduce the number of internal nodes, generating smaller trees, less complex, and thus easier to be understood.

Results

Figure 2 shows the temporal behavior of soil respiration evaluated for 97 days after soil tillage and sugarcane planting. These records were used to compose the dataset for decision tree induction to provide different soil respiration levels, where average values of soil respiration for low, medium and high classes were 0.09, 0.166 and 0.64 g m^{-2} h^{-1}, respectively. The greatest variations in soil respiration occurred due to changes in the soil water content, due to the occurrence of precipitation in the experimental area. Two major peaks occurred in soil respiration during this period, which coincide with precipitation events one day before.

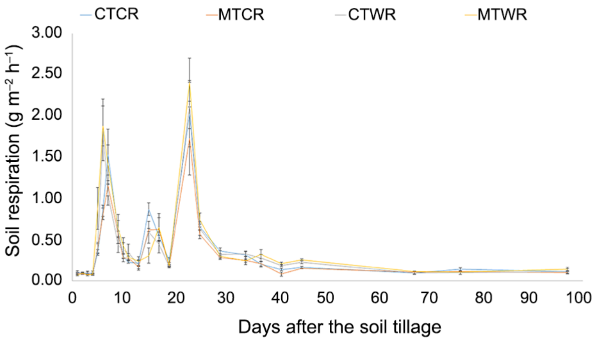

Figure 2 – Temporal variability of soil respiration evaluated after soil tillage at the time of sugarcane crop renewal and the standard errors. CTCR = conventional tillage crop rotation; CTWR = conventional tillage without crop rotation; MTCR = minimum tillage with crop rotation; MTWR = minimum tillage without crop rotation.

Table 4 shows the results of different methods for variable selection. This Table also shows rates of accuracy, Kappa coefficient and the number of rules generated in each tree for each variable selection method performed.

Of the 18 variables available in the dataset, the Wrapper method selected only five of them: soil temperature, precipitation, macroporosity, soil moisture and acidity potential (H + Al) (Table 4). This method works as a "black box" to find the best subsets of variables in order to find the set of variables that best suits the ranking algorithm.

In addition, this method showed good quality indicator metrics of the model, with accuracy of 81 %, a Kappa value of 0.72, indicating that the use of only these five variables is sufficient to achieve a very good classification, yielding a total of 90 rules.

The method without variable selection (used 18 variables available in the dataset) achieved accuracy of 81 %, Kappa index of 0.71 (Table 4), representing a very good correlation to the model; however, with a total of 100 rules.

The correlation-based feature selection method (CFS) selects variables considering those highly correlated with the response variable (soil respiration) and not correlated to each other. This method selected only two variables (temperature and precipitation) and had one of the worst results in the classification reaching an accuracy of 65 %, Kappa statistics of 0.47, which represents a good agreement for the model, with a total of 35 rules generated by the model.

On the other hand, the Chi-square method eliminated only one of the 18 variables, that is, soil organic carbon. This method had 81 % of accuracy with a Kappa index of 0.71, representing very good agreement for the model, with a total of 102 rules.

Table 5 shows the variables selected by the Wrapper method with their respective contributions in the information gain. The soil temperature variable showed information gain higher than the other selected variables. The order of variable influence in determining soil respiration according to information gain was as follows: soil temperature > precipitation > macroporosity > soil moisture > acidity potential.

Table 6 shows the per-class precision results obtained using the Wrapper method and demonstrates that the precision for low, medium and high classes ranged from 0.80 to 0.83, indicating an efficient model to classify the classes in this study, making it applicable to new cases.

In order to minimize the overfitting effects, a pruning procedure was applied in the selected decision tree

Table 4 – Rates of accuracy, Kappa statistics and the number of rules for different variable selection methods aimed at decision tree induction to provide different soil respiration levels.

Selection method	Number of variables select	Rate of accuracy	Kappa	Number of rules
With no variable selection	19	81	0.71	100
Correlation based feature selection	2	65	0.47	35
Chi-square	18	81	0.71	102
Wrapper	5	81	0.72	90

Accuracy rate of the model (%).

Table 5 – Information gain in descending order of variables selected by the Wrapper method aiming at decision tree induction to predict different soil respiration levels.

Variable contribution	Information Gain
Soil temperature	0.48
Precipitation	0.29
Macroporosity	0.08
Soil moisture	0.07
Acidity potential	0.04

Table 6 – Precision values for the low, medium and high classes of the generated model with the Wrapper method for variable selection.

Class	Precision
Low	0.83
Medium	0.80
High	0.82

Table 7 – Accuracy rates, Kappa coefficient, and the number of rules for different pre-pruning levels for the Wrapper method.

Number of objects per leaf	Rate of Accuracy	Kappa	Number of Rules
	%		
2	81	0.72	90
5	80	0.70	63
10	79	0.68	40
15	77	0.66	33
20	76	0.64	27
25	72	0.58	24
30	72	0.58	23

to make the learning model more comprehensive by decreasing the number of rules of the generated tree. Table 7 shows the accuracy rates, Kappa coefficient and the number of rules generated for different pre-pruning levels.

The results in Table 7 show that the decision tree model generated with the variables selected by the Wrapper method had accuracy of 76 % for a pre-pruning level equal to 20, with a Kappa coefficient, considered very good, yielding a total of 27 rules. It is possible to verify that for a pre-pruning level above 20 there is a reduction of the Kappa coefficient to a category regarded as good, without any significant reduction in the number of rules generated.

From the classification model generated by using the variables selected by the Wrapper method with the J48 algorithm for decision tree induction and pre-pruning level above 20 number of objects per leaf, a matrix of agreements was created, expressing the total accuracy and samples misclassified by the model (Table 8).

The main matrix diagonal contains the number of correct classifications, that is, the observed values coincide with the predicted ones for the analyzed dataset, while the off-diagonal elements represent the error obtained by the classifier. Analyzing the results, the proposed classification scheme was satisfactory, with a high accuracy rate for the target classes of classification, confirming the value of the Kappa index equal to 0.64 (Table 7), a very good rating (Landis and Koch, 1977).

Table 8 – Matrix of agreements of classification generated using the Wrapper method for variable selection and the J48 algorithm for decision tree induction.

		PREDICT		
		Low	Medium	High
Observed	Low	**398**	77	43
	Medium	96	**358**	63
	High	33	61	**423**

At this cutoff point, 27 rules were generated; 6 related to high soil respiration class, 11 to the medium soil respiration class, and 10 to low soil respiration class (Table 9).

By analyzing the rules generated based on this model (Table 9), the variables soil temperature, macroporosity, precipitation and soil moisture are essential to determine the high soil respiration class. In addition, the acidity potential (H + Al) only influenced the medium and low soil respiration classes, since the presence of this variable in the characterization of high soil respiration was not observed (Table 9).

Of all the rules generated by the classification model, 88 % covered acidity potential (H + Al) to determine a low soil respiration class and involved values of H + Al less than 5.8 $\text{cmol}_c\ \text{dm}^{-3}$.

Another interesting result was that 80 % of the rules to diagnose soil respiration belonging to the low class are subject to soil temperature below 17.92 °C (Table 9). On the other hand, 67 % of rules to classify soil respiration as high are conditioned to soil temperatures above 18.58 °C.

Discussion

The results of this study showed that data mining techniques allowed the development of an efficient model for soil respiration classification, by using the Wrapper method of selecting variables and the J48 algorithm for decision tree induction. The Wrapper method selected a subset efficient for soil respiration prediction with just five variables, namely soil temperature, precipitation, macroporosity, soil moisture, and acidity potential. The selected variables showed a great consistency with the literature, both for the selected variables and the decision tree rules.

In general, the first variables positioned in a decision tree (from the root) are the most important, since the lower entropy, the higher information gain. Thus, the most important variable represents the root of the decision tree. In this study, the most important variable corresponded to soil temperature (the root of the decision tree).

Many studies have reported that seasonal variations with increases in soil temperatures offer favorable conditions to maximize activity of microorganisms by increasing the organic matter degradation rate and

Table 9 – Rules generated by the pruned decision tree in which the minimum number of objects per leaf = 20 for the Wrapper method.

1	IF 18.58 < Temperature <= 20.84 AND Precipitation > 0 AND 11.9 < Moisture <= 16.96 THEN Low
2	IF Temperature <= 17.92 AND Macro <= 5.49 AND Moisture <= 20.55 THEN Low
3	IF 16.99 <= Temperature <= 18.34 AND Macro <= 5.49 AND Moisture > 20.55 THEN Low
4	IF Temperature <= 17.92 AND Macro > 5.49 AND H+AL > 5.8 AND Precipitation > 0.5 THEN Low
5	IF Temperature <= 17.92 AND Macro > 5.49 AND 5 < H+AL <= 5.8 AND Moisture <= 23.41 THEN Low
6	IF Temperature <= 16.86 AND Macro > 5.49 AND 5 < H+AL <= 5.8 AND Moisture > 23.41 THEN Low
7	IF Temperature <= 17.45 AND 5.49 < Macro <= 5.53 AND H+AL <= 5 THEN Low
8	IF Temperature <= 17.67 AND 5.49 < Macro <= 8.75 AND H+AL <= 5 THEN Low
9	IF Temperature <= 17.67 AND 8.75 < Macro AND H+AL <= 5 AND Moisture <= 20.11THEN Low
10	IF Temperature <= 17.67 AND 9.73 < Macro <= 10.19 AND H+AL <= 5 AND Moisture > 20.11THEN Low
11	IF 17.92 < Temperature <= 18.34 THEN Medium
12	IF 18.34 < Temperature <= 18.58 AND Macro > 5.49 THEN Medium
13	IF 8.58 < Temperature <= 20.84 AND Precipitation > 0 AND Moisture <= 16.96 THEN Medium
14	IF Temperature <= 17.92 AND Macro > 5.49 AND H+AL > 5.8 AND Precipitation <= 0.5 THEN Medium
15	IF 16.86 < Temperature <= 17.92 AND Macro > 5.49 AND 5 < H+AL <= 5.8 AND Moisture > 23.41 THEN Medium
16	IF Temperature <= 17.92 AND Macro > 10.59 AND H+AL <= 5 THEN Medium
17	IF 17.67 < Temperature <= 17.92 AND 5.49 < Macro <= 10.59 AND H+AL <= 5 THEN Medium
18	IF 17.45 < Temperature <= 17.67 AND 5.49 < Macro <= 10.59 AND H+AL <= 5 THEN Medium
19	IF Temperature <= 17.45 AND 5.53 < Macro <= 6.95 AND H+AL <= 5 THEN Medium
20	IF Temperature <= 17.67 AND 8.75 < Macro <= 9.73 AND H+AL <= 5 AND Moisture > 20.11THEN Medium
21	IF Temperature <= 17.67 AND 10.19 < Macro AND H+AL <= 5 AND Moisture > 20.11THEN Medium
22	IF 18.34 < Temperature <= 18.58 AND Macro <= 5.49 THEN High
23	IF Temperature > 18.58 AND Precipitation > 1 THEN High
24	IF Temperature > 18.58 AND 0 <= Precipitation <= 1 AND Moisture > 16.96 THEN High
25	IF Temperature > 18.58 AND Precipitation > 0 AND Moisture > 16.96 THEN High
26	IF 18.58 < Temperature <= 20.84 AND Precipitation > 0 AND 11.9 <= Moisture <= 16.96 THEN High
27	IF 16.99 < Temperature <= 17.92 AND Macro <= 5.49 AND Moisture > 20.55 THEN High

consequently soil respiration (Silva-Olaya et al., 2013; Flanagan et al., 2013; Nie et al., 2013). In Brazil, Tavares et al. (2016a) analyzed the spatial variability in green harvest system with a history of 10 years of implementation and observed that this system showed soil respiration linearly correlated to temperature ($R^2 = 0.80$). However, Bradford (2013) pointed that, seasonal patterns in responses from soil respiration to temperature are strongly dependent on substrate availability, for instance, the temperature must have minimal effects on respiration rates at times of the year when the substrate is depleted, and strong effects when substrate supply is abundant.

The results of this study showed that higher temperatures (> 18.58 °C) contribute to such high soil respiration classification. According to Flanagan et al. (2013), one of the main environmental factors driving soil respiration is soil temperature because of it controls and regulates physiological and various biogeochemical processes (i.e., decomposition of soil organic matter and root metabolism). Wallenstein et al. (2011) reported that soil warming typically accelerates soil microbial respiration rates due to increased soil enzymatic activities, which lead to decomposition.

The second most important variable in making the decision tree was rainfalls (precipitations) indicating, as well as the soil temperature, great effects on soil respiration and can lead to significant carbon loss in the form of CO_2. Figueiredo et al. (2014) evaluated the short-term soil respiration at the time of renewal of sugarcane (*Saccharum* spp.) in southern Brazil and observed increases in soil respiration caused by rainfall events. The authors highlighted that this fact is probably related to changes in soil water content, as precipitation causes increased microbial activity. In addition, there is the additional effect of the removal/displacement of air (CO_2) from soil porosity caused by water infiltration into the soil (Figueiredo et al., 2014). Tavares et al. (2016a) observed higher soil respiration in rainy periods in comparison to the dry period, related to greater microbial activity promoted by soil moisture and root activity during plant growth and development.

In their work with hot spots, hot moments, and spatio-temporal controls on soil respiration in a water-limited ecosystem, Leon et al. (2014) noted that monthly precipitation was the primary drive of the seasonal trend of soil respiration. Moreover, the authors reported that changes in water volumetric content in the soil caused by rainfall influenced the relationship between soil respiration and soil temperature. Therefore, the evaluation of soil temperature must be carefully analyzed, since temperature is influenced by soil moisture, which in turn is influenced by rainfall (Tavares et al., 2016a). In agreement, Sierra et al. (2015) point out that in soils, temperature and moisture covary at different spatial and temporal scales describing a trajectory in the x y plane.

It was observed that the rules generated from the decision tree model, which led to a high soil respiration class, involved the variables of soil temperature, macroporosity, precipitation and soil moisture, which have a high information degree. Several authors point out that soil respiration of agricultural soils is directly related to temperature and soil moisture conditions (Carbonell-Bojollo et al., 2012; Silva-Olaya et al., 2013).

Regarding the influence of precipitation to determine high-class soil respiration, Moitinho et al. (2015) characterized the spatial and temporal behavior of soil respiration and their relation to soil edaphoclimatic properties in sugarcane crops in Dourados, Mato Grosso do Sul, Brazil. They showed that the highest temporal variations of soil respiration were explained by changes in the soil water content, especially after rain. Silva-Olaya et al. (2013) observed the same trend in which higher soil respiration during the days with precipitation occurs probably due to increased soil moisture.

As for macroporosity, Tavares et al. (2015; 2016b) indicated that macroporosity and microporosity exhibit antagonistic behaviors, because macroporosity offers a less tortuous route to the CO_2 molecules, enabling the soil respiration in soil, as microporosity promotes lower linearity in the porous space and is associated with the most tortuous paths that hinder CO_2 transport from the soil to the atmosphere. Probably the joint action of good conditions and the decomposition of soil respiration caused by great temperature conditions, water content in the soil and microporosity favored high soil respiration.

The acidity potential (H + Al) only influenced the medium and low soil respiration classes. Probably this result may have impaired the growth and development of sugarcane plants, resulting in lower root respiration and, consequently, less soil respiration. Soil respiration is strongly linked to plant metabolism. Zhi-Min et al. (2013) observed that the average contribution of root respiration to total soil respiration was about 32 %. Li et al. (2011) pointed out mean contribution of root respiration to total soil respiration over 60 %. Another hypothesis may be related to the tendency to dissociate CO_2 to cause acidity in the soil and control the acidity potential, because the carbonic acid produced by dissolving CO_2 in water is an important acidifying agent in natural systems (Reuss and Johnson, 1986; Tossell, 2009).

Conclusion

Our results showed that data mining techniques allow the development of a model of soil respiration classification, with accuracy of 81 % using 27 rules to predict soil respiration. In addition, the Wrapper method of selecting variables selects a subset of only five variables out of 18 available in the original data set, and they have the following order of influence in determining soil respiration: soil temperature > precipitation > macroporosity > soil moisture > potential acidity.

References

Berry, J.A.M.; Linoff, G. 1997. Data Mining Techniques-for Marketing, Sales and Customer Support. Wiley, New York, NY, USA.

Bradford, M.A. 2013. Thermal adaptation of decomposer communities in warming soils. Frontiers in Microbiology 4: article 333.

Brito, L.F.; Marques Júnior, J.; Pereira, G.T.; Souza, Z.M. 2009. Soil CO_2 emission of sugarcane fields as affected by topography. Scientia Agricola 66: 77-83.

Campos, B.C.; Amado, T.J.C.; Tornquist, C.G.; Nicoloso, R.S.; Fiorin, J.E. 2011. Long-term C-CO_2 emissions and carbon crop residue mineralization in an Oxisol under different tillage and crop rotation systems. Revista Brasileira de Ciência do Solo 35: 819-832.

Carbonell-Bojollo, R.M.; Torres, M.A.R.R.; Rodríguez-Lizana, A.; Ordóñez-Fernández, R. 2012. Influence of soil and climate conditions on CO_2 emissions from agricultural soils. Water, Air, & Soil Pollution 223: 3425-3435.

Dexter, A.R.; Kroesbergen, B. 1985. Methodology for determination of tensile strength of soil aggregates. Journal of Agricultural Engineering Research 31: 139-147.

Dilustro, J.J.; Collins, B.; Duncan, L.; Crawford, C. 2005. Moisture and soil texture effects on soil CO_2 efflux components in southeastern mixed pine forests. Forest Ecology and Management 204: 85-95.

Donagema, G.K; Campos, D.V.B.; Calderano, S.B.; Geraldes, W.; Viana, J.H.M. 2011.

Manual of Methods of Soil Analysis = Manual de Métodos de Análise de Solo. 2ed. Embrapa Solos, Rio de Janeiro, RJ, Brazil (in Portuguese).

Fang, C.; Moncrieff, J.B. 2001. The dependence of soil CO_2 efflux on temperature. Soil Biology and Biochemistry 33: 155-165.

Figueiredo, E.B.; Panosso, A.R.; Donald, C.; Reicosky, D.C.; La Scala Jr, N. 2014. Short-term CO_2-C emissions from soil prior to sugarcane (*Saccharum* spp.) replanting in southern Brazil. GCB Bioenergy 7: 316-327.

Flanagan, L.B.; Sharp, E.J.; Letts, M.G. 2013. Response of plant biomass and soil respiration to experimental warming and precipitation manipulation in northern great plains grass land. Agricultural and Forest Meteorology 173: 40-52.

Han, J.; Kamber, M.; Pei, J. 2011. Data Mining: Concepts and Techniques. 3ed. Morgan Kaufmann, San Francisco, CA, USA.

Kemper, W.D.; Chepil, W.S. 1965. Size distribution of aggregates. p. 499-510. In: Black, C.A., ed. Methods of soil analysis: physical and mineralogical properties, including statistics of measurement and sampling. Part 1. American Society of Agronomy, Madison, WI, USA.

Kohavi, R.; George, H.J. 1997. Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.

Köppen, W.; Geinge, R. 1928. Klimate der erde. Gotha: verlag justus perthes. Wall-map.

Lal, R. 2001. Fate of eroded soil carbon: emission or sequestration. p. 173-181. In: Lal, R., ed. Soil carbon sequestration and the greenhouse effect. Soil Science Society of America, Fitchburg, WI, USA. (SSSA Special Publications, 57).

Lal, R. 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304: 1623-1627.

Landis, J.R.; Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Leon, E.; Vargas, R.; Bullock, S.; Lopez, E.; Panosso, A.R.; La Scala, N. 2014. Hot spots, hot moments, and spatio-temporal controls on soil CO_2 efflux in a water-limited ecosystem. Soil Biology and Biochemistry 30: 1-10.

Li, Z.G.; Wang, X.J.; Zhang, R.H.; Zhang, J.; Tian, C.Y. 2011. Contrasting diurnal variations in soil organic carbon decomposition and root respiration due to a hysteresis effect with soil temperature in a Gossypium s. (cotton) plantation. Plant and Soil 343: 347-355.

Marcelo, A.V.; Cora, J.E.; La Scala Júnior, N. 2012. Influence of liming on residual soil respiration and chemical properties in a tropical no-tillage system. Revista Brasileira de Ciência do Solo 36: 1-6.

Moitinho, R.M.; Padovan, M.P.P.; Panosso. A.R.; Teixeira, D.B.; Ferraudo, A.S.; La Scala Jr., N. 2015. On the spatial and temporal dependence of CO_2 emission on soil properties in sugarcane (*Saccharum* spp.) production. Soil and Tillage Research 148: 127-132.

Monard, M.C.; Baranauskas, J.A. 2002. Rule Induction and Decision Trees = Indução de regras e árvores de decisão. p. 115-140. In: Rezende, S.O. ed. Intelligent systems: fundamentals and applications = Sistemas inteligentes: fundamentos e aplicações. Manole, Barueri, SP, Brazil (in Portuguese).

Nie, M.; Pendall, E.; Bell, C.; Gasch, C.K.; Raut, S.; Tamang, S.; Matthew D.; Wallenstein, M.D. 2013. Positive climate feedbacks of soil microbial communities in a semi-arid grassland. Ecology Letters 16: 234-241.

Orchard, V.A.; Cook, F.J. 1983. Relationship between soil respiration and soil moisture. Soil Biology and Biochemistry 15: 447- 453.

Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, USA.

Raij, B.V.; Andrade, J.C.; Cantarella, H.; Quaggio, J.A. 2001. Chemical Analysis to Evaluate the Fertility of Tropical Soils = Análise Química para Avaliação da Fertilidade de Solos Tropicais. Instituto Agronômico, Campinas, SP, Brazil (in Portuguese).

Reuss, J.O.; Johnson, D.W. 1986. Acid Deposition and the Acidification of Soils and Waters: Ecological Studies. Springer, Berlin, Germany.

Shannon, C.E. 1949. The Mathematical Theory of Communication. University of Illinois Press, Chicago, IL, USA.

Sierra, C.A.; Trumbore, S.E.; Davidson, E.A.; Vicca, S.; Janssens, I. 2015. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. Journal of Advances in Modeling Earth Systems 7: 335-356.

Silva-Olaya, A.M.; Cerri, C.E.P.; La Scala Junior, N.; Dias, C.T.S.; Cerri, C.C. 2013. Carbon dioxide emissions under different soil tillage systems in mechanically harvested sugar cane. Environmental Research Letters 8: 1-8.

Smith, P.; Lanigan, G.; Kutsch, W.L.; Buchmann, N.; Eugster, W.; Aubinet, M.; Ceschia, E.; Beziat, P.; Yeluripati, J.B.; Osborne, B.; Moors, E.J.; Brut, A.; Wattenbach, M.; Sunders, M.; Jones, M. 2010. Measurements necessary for assessing the net ecosystem carbon budget of croplandas. Agriculture, Ecosystems and Environment 139: 302-3015.

Soil Survey Staff. 2014. Keys to Soil Taxonomy. 12ed. USDA-Natural Resources Conservation Service, Washington, DC, USA.

Stockmann, U.; Adams, M.A.; Crawford, J.W.; Field, D.J.; Henakaarchchi, N.; Jenkins, M.; Minasny, B.; Mc Bratney, A.B.; Courcelles, V.R.; Singh, K.; Wheeler, I.; Abbott, L.; Angers, D.A.; Baldock, J.; Bird, M.; Brookes, P.C.; Chenu, C.; Jastrow, J.D.; Lal, R.; Lehmann, J.; O'Donnell, A.G.; Parton, W.J.; Whitehead, D.; Zimmermann, M. 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agriculture, Ecosystems & Environment 164: 80-99.

Stolf, R. 1991. Theory and test of formulas for transforming impact penetrometer data in soil resistance. Revista Brasileira de Ciência do Solo 15: 229-235 (in Portuguese, with abstract in English).

Tavares, R.L.M.; Souza, Z.M.S.; La Scala Jr, N.; Castioni, G.A.F.; Souza, G.S.; Torres, J.L.R. 2016a. Spatial and temporal variability of Soil CO_2 flux in sugarcane green harvest systems. Revista Brasileira de Ciência do Solo 40: 1-14.

Tavares, R.L.M.; Siqueira, D.S.; Panosso, A.R.; Castioni, G.A.F.; Souza, Z.M.; La Scala Jr., N. 2016b. Soil management of sugarcane fields affecting CO_2 fluxes. Scientia Agricola 73: 543-551.

Tavares, R.L.M.; Souza, Z.M.; Siqueira, D.S.; La Scala Jr, N., Panosso, A.R.; Campo, M.C.C. 2015. Soil CO_2 emission in sugarcane management systems. Soil and Plant Science 65: 755-762.

Tossell, J.A. 2009. H_2CO_3 (s): a new candidate for CO_2 capture and sequestration. Environmental Science & Technology 43: 2575-2580.

Vance, E.D.; Brookes, P.C.; Jenkinson, D.S. 1987. Microbial biomass measurements in forest soils: the use chloroform fumigation-incubation method in strongly acid soils. Soil Biology and Biochemistry 19: 697-702.

Vicca, S.; Bahn, M.; Estiarte, M.; Van Loon, E.E.; Vargas, R.; Alberti, G.; Ambus, P.; Arain, M.A.; Beier, C.; Bentley, L.P.; Borken, W.; Buchmann, N.; Collins, S.L.; De Dato, G.; Dukes, J.S.; Escolar, C.; Fay, P.; Guidolotti, G.; Hanson, P.J.; Kahmen, A.; Kroel-Dulay, G.; Ladreiter-Knauss, T.; Larsen, K.S.; Lellei-Kovacs, E.; Lebrija-Trejos, E.; Maestre, F.T.; Marhan, S.; Marshall, M.; Meir, P.; Miao, Y.; Muhr, J.; Niklaus, P.A.; Ogaya, R.; Peñuelas, J.; Poll, C.; Rustad, L.E.; Savage, K.; Schindlbacher, A.; Schmidt, I.K.; Smith, A.R.; Sotta, E.D.; Suseela, V.; Tietema, A.; Van Gestel, N.; Van Straaten, O.; Wan, S.; Weber, U.; Janssens, I.A. 2014. Can current moisture responses predict soil CO_2 efflux under altered precipitation regimes? A synthesis of manipulation experiments. Biogeosciences 11: 2991-3013.

Wallenstein, M.; Allison, S.D.; Ernakovich, J.; Steinweg, J.M.; Sinsabaugh, R. 2011. Controls on the temperature sensitivity of soil enzymes: a key driver of in situ enzyme activity rates. p. 245-258. In: Shukla, G.C.; Varma, A., eds. Springer, Berlin, Germany.

Witten, I.H.; Frank, E.; Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3ed. Morgan Kaufmann, San Francisco, CA, USA.

Zhi-Min, Z.; Cheng-Yi, Z.; Yilihamu, Y.; Ju-Yan, L.; Jun, L. 2013. Contribution of root respiration to total soil respiration in a cotton field of Northwest China. Pedosphere 23: 223-228.