THE INTERNATIONAL SOCIETY OF
PRECISION AGRICULTURE PRESENTS THE
13th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE
July 31-August 4, 2016 • St. Louis, Missouri USA

Translating Data into Knowledge – Precision Agriculture Database in a Sugarcane Production

Sanches, G. M.; Kolln, O. T.; Franco, H. C. J.; Duft, D. G., Magalhães, P. S. G.

Brazilian Bioethanol Science and Technology Laboratory (CTBE), Brazilian Center for Research in Energy and Material (CNPEM). Biomass Production Division. Campinas, São Paulo, Brazil.

**A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA**

Abstract. The advent of Information Technology in agriculture, surveying and data collection became a simple task, starting the era of "Big Data" in agricultural production. Currently, a large volume of data and information associated with the plant, soil and climate are collected quick and easily. These factors influence productivity, operating costs, investments and environment impacts. However, a major challenge for this area is the transformation of data and information (collected in the field) in applicable knowledge. Within the context of Precision Agriculture (PA), which comprises a set of tools and technologies for georeferenced data collection to understand and manage inherent spatial variability within crop fields, the Brazilian sugarcane industry lacks results to assist farmers. The hypothesis of this work is that with the knowledge of the spatial variability of soil fertility and crop productivity, through the application of data mining techniques, it is possible to assist sugarcane producer in the correct management of the crop. Two areas cultivated with sugarcane, with 10 and 30 ha, were monitored over the years 2012, 2013 and 2014. During this period, soil sampling was taking annually (117 and 107 points, respectively) and yield maps registered using a yield monitor. Using a computational environment created to support sugarcane agricultural research, data acquisition, formatting, verification, storage, and analysis of the principal component analysis (PCA) and decision trees for knowledge extraction were performed. The results show that a major factor for variation of sugarcane crops yield is related to texture, the amount of organic matter available and soil pH. Where there was an increase in the levels of organic matter from one year to another there was an increase in capacity cation exchange (CTC) and greater availability of Potassium and Phosphorus. Based on the knowledge rules by a decision trees analysis, it is possible to created specific management zones in the field that support the grower in a decision making. With the expanded dataset, we expect to recognize relevant patterns that are reproduced consistently across distinct experiments, assisting producers in the correct crop management to improving the profitability of production.

Keywords. *data mining; precision agriculture dataset; decision trees, principal component analysis*

Introduction

Precision Agriculture (PA) comprises some management practices to attempt increase productivity, profitability and improve environmental stewardship of rural areas. Essentially, the benefits are achieved by local treatment, considering the spatial variability. The main technologies available for PA users are yield monitors, remote and proximal sensing, Global Navigation Satellite Systems (GNSS) and Global Information Systems (GIS). However, these technologies are more advanced in cereals and grains, when compared to sugarcane industry (Silva et al, 2011). One of the factors for the weak advancement of precision agriculture (PA) in the Brazilian sugarcane industry is the lack of applicable knowledge to assist farmers in decision-making. The development proper decision-support systems for implementing precision decisions remains a major stumbling block to the PA adoption (McBratney et al, 2006). At the strategic and tactical level, assembled data on the performance of various farm management systems should be grouped by soil series to build a systematic database, allowing "quick and preliminary" evaluations of the effects of farm management strategies based on experiences obtained elsewhere on similar soils (Bouma et al, 1999). To overcome this challenge, agricultural information technology (AIT) has been broadly applied to every aspect of agriculture and has become the most effective means and tools for enhancing agricultural productivity and for making use of full agricultural resources (Yan-e, 2011). Within this context, the Precision Agriculture and Geoprocessing Team of Brazilian Bioethanol Science and Technology Laboratory (CTBE, Campinas, São Paulo, Brazil) has worked to contribute to the Precision Agriculture expansion in the sugarcane industry. The main objective of this paper was use a computational environment created to support sugarcane agricultural research, data acquisition, formatting, verification, storage, and data analysis (Driemeier et al, 2014) to assist sugarcane producers in the correct management by the extraction knowledge of the soil spatial variability and crop productivity through data mining techniques. From large volumes of data, obtained by different technologies, it is possible to extract relevant knowledge that can assist producers in production profitability, increasing the efficiency and sustainability of sugarcane industry.

Material and Methods

The data used in this study are from two experimental areas of precision agriculture projects in sugarcane fields. The first experimental area is located at Pedra Mill (PeM - 30 hectares - Sao Paulo - Brazil - 21°16'36.94"S 47°18'31.31"W - 583 m) and the second in Sao Joao Mill (SJM - 10 hectares - Sao Paulo - Brazil - 22°23'37.21"S 47°18'31.31"W - 640 m). The average slope of the areas is 10% and 2%, respectively, for the PeM and SJM. The sugarcane varieties in the experimental areas, chosen according to the weather conditions and local soil type, were CTC09 and SP80-3280 to PeM and SJM, respectively. The details in the management and initial objectives of the PeM and SJM experiments were reported in Magalhães et al. (2014) and Rodrigues et al. (2012), respectively. The main difference in the management used in the fields is related to soil fertilization. In PeM fertilizer were applied at variable rates yearly over the three cane cycles: Nitrogen (in accordance to the expected yield), phosphorus and potassium (in accordance to soil deficit determined by wet-chemical soil analysis), while in the SJM there received no fertilizer during the years of the experiment. The areas were sampled on a regular grid of 50x50 m and 30x30 m for PeM and SJM, respectively, with 107 and 117 sampling points (Figure 1). All soil samples were submitted to laboratory tests to characterize the macro and micronutrients, pH, organic matter and clay, silt and sand content at surface layer (0.00 to 0.20 m). For this study the soil attributes of greatest interest for sugarcane were analyzed: organic matter (OM), pH, phosphorus (P), Potassium (K), calcium (Ca), magnesium (Mg), hydrogen + aluminum (H + Al), sum of the bases (SEB), cation exchange capacity (CEC) and base saturation (BS). The cycles of cane corresponding to the evaluated years were: cane plant, 1st and 2nd ratoon for PeM; and 2nd and 3rd ratoon for SJM. The harvest of the experimental areas was monitored by yield monitor (SIMPROCANA, Enalta, São Carlos, Brazil). The yield data were reduced to the soil sampling grid using a regression in the buffer zone (Figure 1 - Detail) using the algorithm presented by Driemeier et al. (2014). As a first step, attributes of soil chemical composition were converted to the logarithm of concentrations. The logarithm scale reduced the skewness from concentration distributions positive and was additionally justifiable from a physical-chemical perspective (Atkins and Paula, 2002). The second step was to remove outliers from the data sets, which could cause detrimental bias to correlations and covariance. Any entry deviating from the mean by more than three standard deviations (for a given attribute) was treated the outlier (Driemeier et al, 2014).

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed for the purpose of simplifying the description, a set of interrelated variables using the dimensionality reduction and interpretation of components. This analysis does not discriminate variables as dependent or independent, as in the regression analyzes, with all attributes treated as variables. Thus, this technique can be understood as a method of transformation of the original variables into new uncorrelated variables, where each principal component (PC) is a linear combination of the original variables. The amount of data explained by each component is given by the variance, and the PC sorted in descending order, where the main component containing more information is the first, and so on. Algebraically, the principal components are linear combinations of *p* random variables X_1, X_2, ..., X_p. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with X_1, X_2, ..., Xp as the coordinate axes. The new axis is the direction of maximum variability and provides a simple description of the covariance structure (Johnson and Wichern, 2007). PCA was employed to reduce the dimensionality of the attribute space and observe the correlation structure between the different soil attributes evaluated and sugarcane yield. Prior to PCA, imputation of missing data was performed by the Expectation-Maximization algorithm associated with a multivariate normal model, as described in Johnson and Wichern (2007). At first time we used the PCA with the original data from the two areas in all evaluated cycles. At the second time we made data subtraction (Eq. 1) to obtain the variability of the attributes levels at the evaluated sample points over the years. The main objective of this analysis was to know the correlation structure between the soil attributes and yield, observing possible reasons of yield variation over the experimental fields. For Eq. 1, a positive value means an increase in the assessed attribute from one year to the next, while a negative value the opposite case. This analysis, by dimensionality reduction of the problem, allow the interpretation of the various parameters evaluated in a simpler and effective way, resulting in a robust application to identify the determining factors in the sugarcane yield.

$$C_{(N\boldsymbol{x},\boldsymbol{y})} = C_{(i\boldsymbol{x}+,\boldsymbol{y}1)} - C_{(i\boldsymbol{x},\boldsymbol{y})} \qquad (1)$$

where: $C_{(}{}^{N}{}_{x,y)}$ – new content attribute evaluated at coordinates (x,y) and i – evaluation year.

Decision Trees Analysis

Decision Trees belong to the class of supervised algorithms where a dependent variable is explained to the cost of *n* independent variables, measured on any scale. The Decision Tree consists in a set of rules that associate a set of conditions with a specific outcome. These rules can be also represented in an intuitive tree format, enabling the visualization of the relationships between the predictors and the output. One of the main features of a decision tree is its representation in the form of a hierarchical structure that translates an inverted tree that grows from the root to the leaves. This translates into a hierarchical data analysis representation of progression in order to perform a predictive or classification task. The principle in classification trees is "divide-to-achievement". Thus, at each level of a tree, a problem more complex of prediction/classification (in which there is greater heterogeneity of the target variable values) is decomposed into simpler sub problems. This fact is reflected in the descendants nodes, in which heterogeneity of the variable to predict (and explain) is attenuated and can make predictions with lower risks for each of nodes (Santos Rodrigues, 2005). It follows that an investigation level flow from "general" for the "specific" case, in that each new tree level descendant nodes is limited to the value of another explanatory attribute. Decision trees can be used for different purposes, according to the problem to be solved. For this study the decision tree was used to find the soil attributes (independent variables) that better explain the sugarcane yield variability (dependent variable). There are several algorithms used in decision trees. Here it was used the CHAID algorithm - Chi-square Automatic Interaction Detection (Kass, 1980). The method adopted by this algorithm is the recursive division of the observations set in subgroups. At each step, the algorithm determines a classification rule by selecting a variable, where a cutoff in the values of this variable is done to maximize the statistical difference of subgroups (in relation to the dependent variable). All soil attributes evaluated in this study were categorized, and the contents ranked from the lowest levels to the highest, according to Raij et al. (1997). The only exception was for the organic matter content, where we used the continuous values of this one. The yield (dependent variable) was classified into five levels: very high (≥110 Mg ha^{-1}); high (90≤y <110 Mg ha^{-1}); medium (70≤y <90 Mg ha^{-1}); low (50≤y <70 Mg ha^{-1}) and very low (<50 Mg ha^{-1}). The tree CHAID decision was executed by STATISTICA 13® software (StatSoft, Dell Software, Oklahoma, USA).

Results

The experimental areas are differences in the mean levels of clay and sand content (Fig. 2), PeM area being more clay (≈458 g kg^{-1} of clay) than SJM (≈232 g kg^{-1} of clay). Silt is on average equal (≈90 g kg^{1}) for both areas. The average levels of organic matter in the two experimental areas decreased over time (Fig. 3), the average levels exceeding 20 g dm^{-3} for the PeM area and less than 12 g dm^{-3} for SJM area. Phosphorus content increased over the years for both areas, with the SJM area contains approximately three times more phosphorus content compared PeM. Considering a desired minimum level of 16 mg dm^{-3} P, only the SJM maintained their mean levels within this range. Potassium decreased over time for the PeM, while for SJM increased in the third ratoon. PeM was richer in potassium and maintained the level within the desired range (K = 1.6 mmol$_c$ dm^{-3}). Calcium and magnesium levels were always higher than desired (according Raij et al, 1997) for both areas, although calcium concentrations decreased over time at PeM. The soil pH was, on average, always within the average range (5.1 <pH <5.5) for both areas, however the minimum values (data not show) to this element were as low as 4.4 (high acidity) for both areas. The cation exchange capacity increased at SJM, while decreased at PeM, following the trends of Ca and Mg elements. The BS remained, on average, more than 60% (desired level) in SJM and was lower at PeM. The average yield declined over the sugarcane cycles in both areas. The largest decrease rate from successive cycles occurred at PeM (94 to 60 Mg ha^{-1} from cane plant to 1st ratoon). The distribution of the raw yield data shows the absence of outliers in the samples for both areas (Fig 4), and the less data variability was registered in the second ratoon at PeM (CV = 8%). The highest yields were recorded for PeM (≈140 Mg ha^{-1} - Cane Plant), while the lowest for the SJM (≈37 Mg ha^{-1} - Third Ratoon). Through PCA of the soil attributes and yield, components one and two explained, together, approximately 67% of the total variability of the data. According to the PCA, it is possible to observe that yield was directly influenced by the levels of potassium, organic matter and H + Al contents in the

soil (Fig 5 - top - left). The main facts that demonstrate this came from PeM data (Fig 5 - top - right), with a clear distinction of SJM and PeM clusters. The Pearson correlation coefficient (data not show) between yield and OM, K and H + Al were significant at 5% probability for PeM (r = 0.60, 0.28 and 0.33, respectively, for OM, K and H + Al). For SJM correlation coefficients were significant only for the OM and H + Al (r = 0.20 and 0.30, respectively). On the other hand, analyzing the attributes in terms of the difference between the evaluated years, it is possible to observe that yield is directly related to changes in the soil pH (Fig 5 - bottom - left). This correlation was best expressed by data from the PeM (r = 0.48). For SJM there were no attributes with significantly correlation with yield variation (Table 1). The SEB was directly related to variations in Ca and Mg content, especially Ca (r = 0.99 and 0.97, at PeM and SJM, respectively). It is also possible to observe that variations in OM caused variations in the P, K and CTC contents, as the orientation of the vectors in the PCA. The variation in the OM caused significant variation in the P and K contents for SJM (r = 0.29 and 0.22, respectively) and CTC for the PeM (r =
0.34). The decision tree algorithm based on CHAID first divided data according to the experiment area, i.e. showed that the PeM and SJM sites are significantly different in terms of yield (dependent variable). From 520 yield data in the database for analysis by regression tree, most often, in categorical terms, it is for high yields (Fig 6). The global average for the two areas was 83.45 Mg ha^{-1}. The first division of the tree between the two areas assessed also shows that the highest frequency in the PeM and SJM data are for high yields, averaged over all evaluated cycles equal to 88.61 and 76.36 Mg ha^{-1}, respectively. After this first division, the organic matter attribute was the most significant in explaining yield for PeM, with divisive content equal to 23 g dm^{-3}. Contents above this value showed a higher frequency of high yields (M = 99.59 Mg ha^{-1}), while lower levels showed low yields (M = 69.81 Mg ha^{1}). On the other hand, the attribute that influenced the yield for SJM was soil pH, where high acidity led to a greater frequency of low yield (M = 73.96 Mg ha^{-1}) and low acidity led to high yields (M = 81.81 Mg ha^{-1}). After the relevance of OM in the soil, pH was the determining factor in yield for PeM, while BS was decisive in SJM. In places where there were higher levels than 23 g dm^{-3} of OM and where the acidity showed high, the high content of potassium (> 3.1 mmol$_c$ dm^{-3}) was an important factor to produce high yields (M = 107.2 Mg ha^{-1}). Through the most important factors in determining yield for the evaluated areas, established by rules created in the decision tree, it's possible create management zones that allow to guide farmers in soil management and decision making in sugarcane (Fig. 7).

Discussion

With the total clay content in soils, it is clear the textural difference between the areas evaluated in this study. The average content of clay, sand and silt in both areas, can be classified into sandy-clay-loam and sandy-loam (EMBRAPA, 1999), respectively, for PeM and SJM. The maximum levels of clay found in the soil show that PeM can be clayey in specific regions of the field, while SJM presented transition regions of sandy-loam to clay-sandy-loam texture. Common fact in the sugarcane fields, the present study also observed the decrease rate in average levels of organic matter content for both areas (Santos et al, 2008). The organic matter content in the soil also followed, as expected, the soil texture. The most clayey areas, such as PeM showed higher levels of organic matter, while SJM showed lower levels (< 12 g dm^{-3}). The values are as expected according Raij et al (1997), where sandy soils have lower levels of OM (<15 g dm^{-3}) and in the clayey soils the levels are between 16 and 30 g dm^{-3}. The yield variability data in the sampling grid, formatted according to the algorithm developed in Driemeier et al. (2014), demonstrate the model robustness to removing the noise and possible discrepancies in sugarcane yield monitors, as reported by Maldaner et al (2015). Because of different management methods adopted in the initial objectives of the experiments, in the PeM (where there were fertilizations with P and K) the average K levels decrease over the cycles, while for phosphorus there was an increase in the average content. In SJM, where no fertilization was done, opposite to expectations, there was an increase in K and P average contents. For SJM, one hypothesis of growth P and K contents in the soil can be observed by the principal component analysis. Through the PCA and the Pearson's correlation coefficient, it is possible to note that there is an evidence of increase in these levels in places where there was an increase in organic matter content of soil, where the inverse rule is also true, i.e., the decrease in OM contents reduced the availability P and K contents (Figure 5 - below). This was also reported by Nogueirol et al (2014) showing the importance of organic matter in the availability of macro and micronutrients in the soil. With a higher OM levels in the soil, PeM showed a direct correlation between the organic matter and production (r = 0.60), showing also that where this attribute is found in higher concentrations in the soil it is possible to provide greater amounts of nutrients, reach higher yields. The greater availability of OM in PeM caused a direct relationship with the CEC levels, showing again to be an important element in soil fertility (Landell et al, 2003). This fact is evidenced by the PCA with the original elements contents in the soil (Figure 5). Moreover, as expected, the greater availability of Ca and Mg promotes an increase in Sum of the Bases (SEB) for both areas. This show the robustness of the data and analysis used, since the Sum of the Bases is closely related to these elements (SEB = K + Ca + Mg + Na), where the sodium content is irrelevant compared to the other elements concentration in the soil (Raij et al, 2001). Aiming to establish rules that could highlight the applicable knowledge through dataset used here, we applied decision trees technique based on the CHAID algorithm. While it will require large amounts of data to ensure statistical differences (Santos Rodrigues, 2005), the CHAID algorithm has the advantage of not allowing the occurrence of overfitting, i.e., the tree overgrowth to reach the smallest details in data variability, making this an efficient and practical algorithm in targeting or tree growth. By applying this algorithm, it is evident the initial distinction between the experimental sites according to the yield. This fact shows the relationship between soil type and yield potential (Raij, 2011), one of the most decisive factors in establishment of management zones. Clayey soils present greater production potential compared to sandy soils, as evidenced in this paper. For this fact, sugarcane is managed according to production environments (Prado, 2005), where the soil texture is one of the determining factors for classification this environments "zones". PeM, more clayey, showed that the second most important factor in determining the productivity was the organic matter available in the soil, where the content of 23 g dm^3 distinguished productive local sites. The locals with higher level of this element produced, approximately, 30 Mg ha^{-1} more sugarcane biomass. On the other hand, in the SJM (where no fertilizer input, soil pH (expressed in terms of acidity) was an important factor to differentiate higher yield zones (M ≈ 82 t ha^{-1}) to the lower yield zones (M ≈ 74 Mg ha^{-1}). This fact contributes to Malavolta (1979) that show that the availability of nutrients to the plants occurs when the pH is at in lower acidity conditions, with the ideal range for the sugarcane is between 5.5 and 6.0 of pH (Raij et al., 1997). Thus, this work contributes towards to evidence that the pH must be handled properly in low productive potential sites. In PeM soil acidity was also divisor factor productivity after the organic matter content. In poorer locations of OM (<23 g dm^{-3}) and the lower acidity regions, it is possible observe higher yield zones (M ≈ 81 mg ha^{-1}). On the other hand, in regions of high OM contents (> 23 g dm-3), potassium is presented as an important factor in determining high yields (≈ 30 Mg ha^{-1} sugarcane biomass difference between locations where there were high levels of potassium in the soil compared to lower levels). This fact also contributes to evidence the availability of this element in the soil is a decisive factor for biomass production. By the rules established with this database, we can classify the areas in precision management zones (Figure 7) that allow help producers to increase the production profitability, increasing the efficiency and sustainability of sugarcane industry.

Conclusion

This work shows the importance of data analysis tools focused on agriculture. The database construction is extremely important in helping the knowledge extraction for producers, allowing greater production profitability. The organic matter content and soil pH are essential factors that must be managed properly to ensure higher yields and should be managed according to their spatial variability. With the expanded dataset, we expect to recognize relevant patterns that could be reproduced consistently across distinct experiments, assisting producers to perform the correct crop management and improve the production profitability. Our research team are working to expand the sugarcane Precision Agriculture database, adding data from different data acquisition technologies as well as data from other experiments (finished and ongoing).

Acknowledgments

FAPESP (São Paulo Research Foundation) for the financial support (Process: 2015/01587-0; 2013/50942-2 and 2008/09265-9), Pedra and São João Mill for field support.

References

Atkins, P., Paula, J. de, 2010. *Physical Chemestry*, 9th ed. Oxford University Press,434 Oxford.

Bouma, J., Stoorvogel, J., van Alphen, B. J., Booltink, W. G. (1999). Pedology, Precision Agriculture, and the Changing Paradigm of Agricultural Research. Soil Science Society of America, 63 (6) 1763-1768.

Driemeier, C. E., Ling, L. Y., Pontes, A. O., Sanches, G. M., Franco, H. C. J., Magalhães, P. S. G., Ferreira, J. E. (2014) Data Analysis Workflow for Experiments in Sugarcane Precision Agriculture. In: IEEE 10th International Conference on eScience (eScience), Sao Paulo.IEEE 10th International Conference on e-Science. p. 163.

EMBRAPA. Sistema brasileiro de classificação dos solos. Rio de Janeiro: Embrapa Solos, 1999. 412 p.

Johnson, R.A., Wichern, D.W., (2007). *Statistical Analysis*, sixth edit. ed. Pearson Pretice Hall, Upper Saddle River.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 (2), 119127.

EMBRAPA. Sistema brasileiro de classificação dos solos. Rio de Janeiro: Embrapa Solos, 1999. 412 p.

Landell, M.G.A.; Figueiredo, P.; Vasconcelos, A.C.M. (2003). O estado da arte da pesquisa em cana-de-açúcar na região Centro-sul do Brasil. In: REUNIÃO ITINERANTE DE FITOSSANIDADE DO INSTITUTO BIOLÓGICO – RIFIB, 9, 2003b, Catanduva, SP. Anais..., p.1-9.

Magalhães, P. S. G., Sanches, G. M., Franco, H. C. J., Driemeier, C. E., Kolln, O. T., Braunbeck, O. A. (2014). Precision Agriculture in Sugarcane Production. A Key Tool to Understand Its Variability. In: 12 International Conference on Precision Agriculture, 2014, Sacramento, CA. 12th ICPA Abstracts Book.

MALAVOLTA, E. *ABC da Adubação*. 4a edição. São Paulo SP, Editora Agronomia Ceres, 1979. 255 p.

Maldaner, L. F., Spekken, M., Eitelwein, M. T., Molin, J. P. (2015). Removal of errors on maps of productivity of sugarcane. In: X SBIAGRO – Congresso Brasileiro de Agroinformárica. 21 a23 out. Piracicaba-SP. 11 p.

McBratney, A., Whelan, B., Ancev, T. (2006). Future directions of precision agriculture. *Precision Agriculture*, 6, 7-23.

Noguierol, R. C., Cerri, C. E. P., Silva, W. T. L., Alleoni, L. R. F. (2014). Effect of no-tillage and amendments on carbon lability in tropical. *Soil and Tillage Research*, 143, 67-76.

Prado, H. (2005). Ambientes de produção em cana-de-açúcar. Available at: https://www.ipni.net/ppiweb/brazil.nsf/87cb8a98bf72572b8525693e0053ea70/7759ddc6878ca7eb83256d05004c6dd1/$FILE/Enc12-17-110.pdf. Acessed in: 05/30/2016

Raij, B. van. (2011). *Fertilidade do solo e manejo dos nutrientes*. Piracicaba: International Plant Nutrition Institute. 420 p.

Raij, B. van., Cantarella, H., Quaggio, J.A., Furlani, A.M.C. (1997). *Recomendações de adubação e calagem para o Estado de São Paulo*. Campinas. Instituto Agronômico; Fundação IAC., Campinas.

Raij, B. van. (1991). *Fertilidade do solo e adubação*. Piracicaba, Ceres/Potafos, 343p

Rodrigues Jr., F.A., Magalhães, P.S.G., Franco, H.C.J.., (2012). Soil attributes and leaf nitrogen estimating sugar cane quality parameters: Brix, pol and fibre. *Precis. Agric*. 14,484 270–289.

Santos, G. A., Silva, L. S., Canellas, L. P., Camargo, F. A. O. (2008). *Fundamentos da matéria orgânica do solo: ecossistemas tropicais e subtropicais.* 2 ed. Porto Alegre: Metrópole. 624 p.

Santos Rodrigues, M. A. (2004). Arvores de Classificação. Monografia. Universidade dos Açores – Departamento de Matemática. 34 p.

Silva, C. B., Moraes, M. A. F. D., & Molin, J. P. (2011). Adoption and use of precision agriculture technologies in the sugarcane industry of São Paulo state, Brazil. *Precision Agriculture*, 12(1), 67–81.

Yan-e, D. (2011). Design of Intelligent Agriculture Management Information System Based on IoT. In: Intelligent Computation Technology and Automation (ICICTA),.Guangdong, Shenzhen. v.1. 1045-1049.

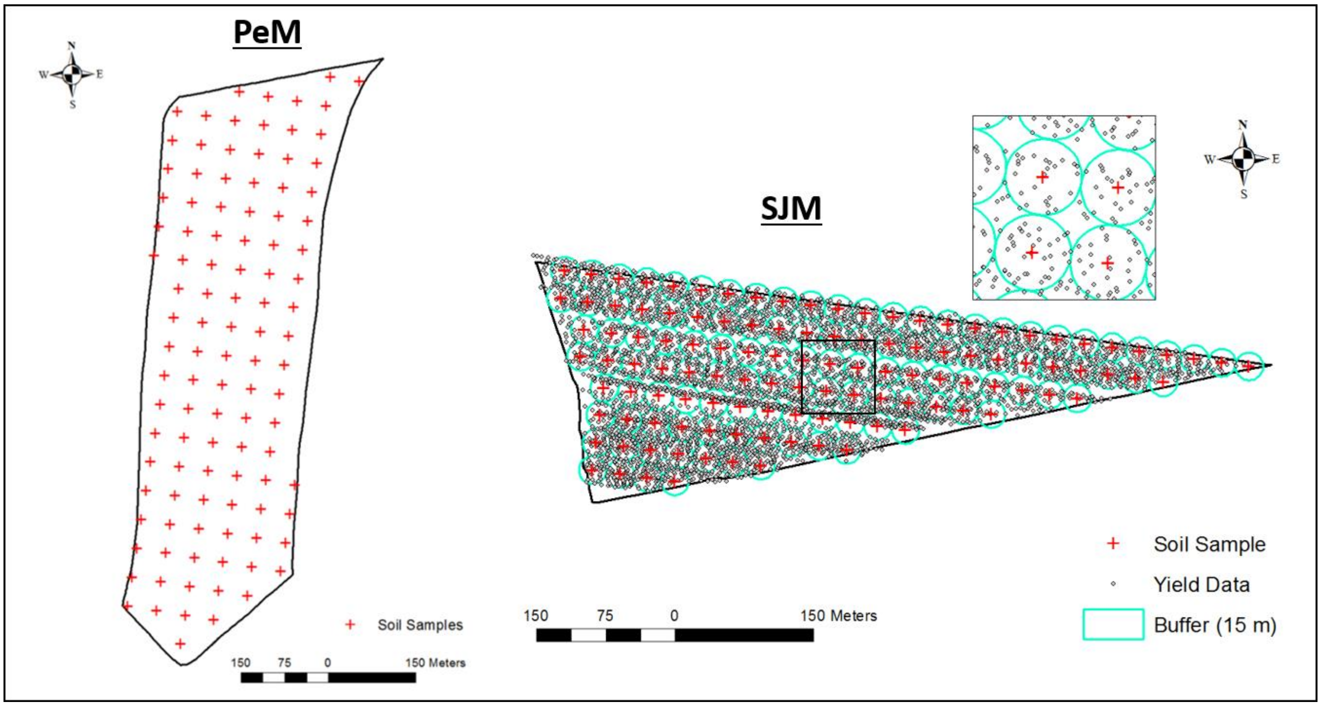

Fig 1. Soil samples location at field experiments of Pedra Mill (PeM) and Sao Joao Mill (SJM). Detail of buffer zone to convert yield data into soil sample grid.

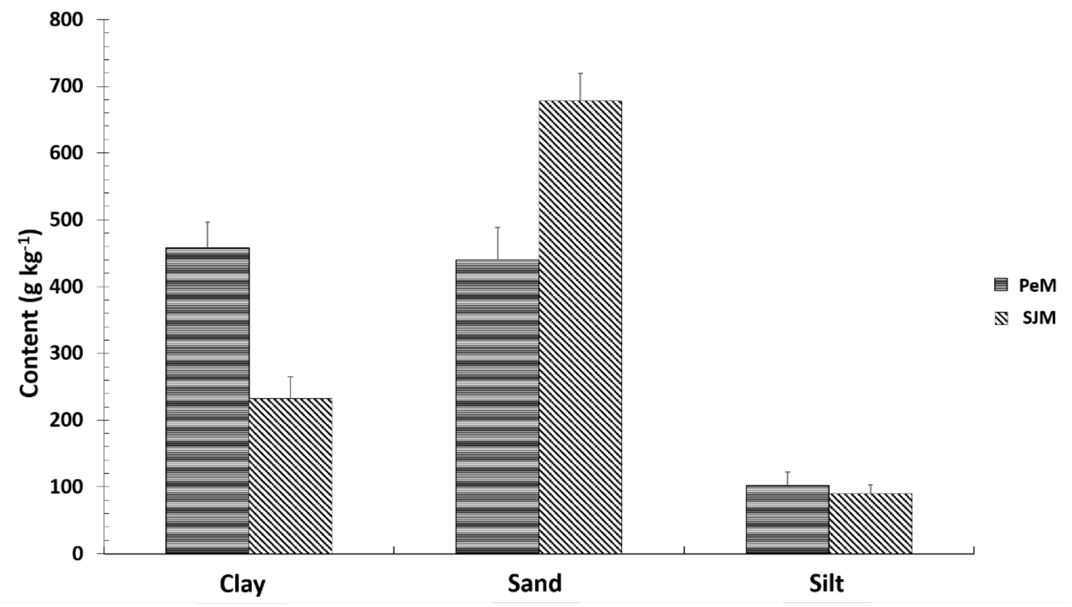

Fig 2. Clay, sand and silt content (g kg^{-1}) of Pedra Mill (PeM) and Sao Joao Mill (SJM).

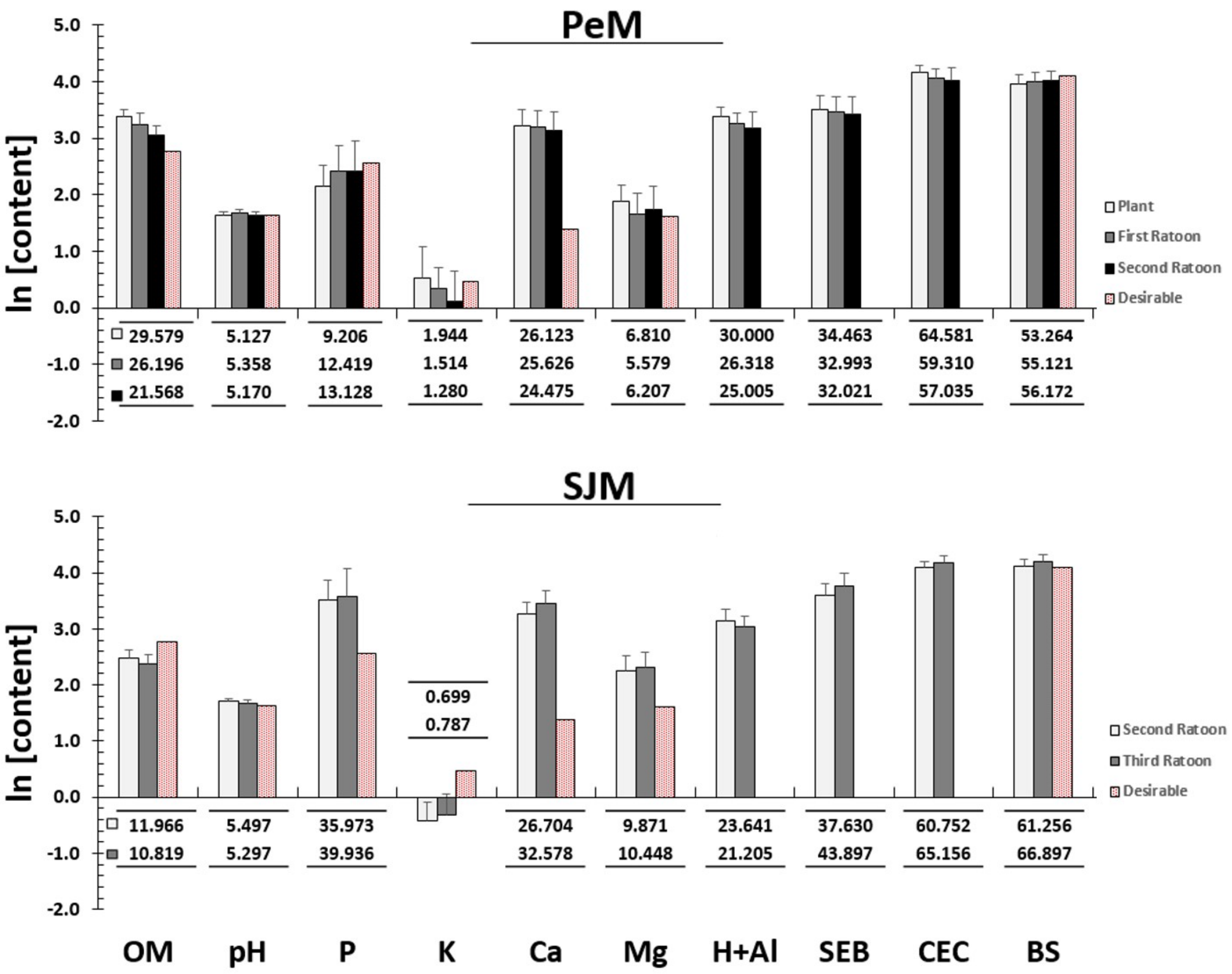
PeM
SJM
ln [content]
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
□ Plant
▨ First Ratoon
■ Second Ratoon
▨ Desirable
□ Second Ratoon
▨ Third Ratoon
▨ Desirable
□ 29.579 5.127 9.206 1.944 26.123 6.810 30.000 34.463 64.581 53.264
▨ 26.196 5.358 12.419 1.514 25.626 5.579 26.318 32.993 59.310 55.121
■ 21.568 5.170 13.128 1.280 24.475 6.207 25.005 32.021 57.035 56.172
0.699
0.787
□ 11.966 5.497 35.973 26.704 9.871 23.641 37.630 60.752 61.256
▨ 10.819 5.297 39.936 32.578 10.448 21.205 43.897 65.156 66.897
OM
pH
P
K
Ca
Mg
H+Al
SEB
CEC
BS

Fig 3. Natural logarithm (ln) of the soil attributes content at Pedra Mill (PeM – above) and Sao Joao Mill (SJM – bottom). The numbers represent the mean content of soil attributes. Red columns represent the desirable content according Raij et. al 1997. [Units]: [OM] – [g dm^{-3}]; [pH] – [at $CaCl_2$]; [P] – [mg dm^{-3}]; [K, Ca, Mg, H+Al, SEB and CEC] – [mmol$_c$ dm^{-3}]; [BS] – [%].

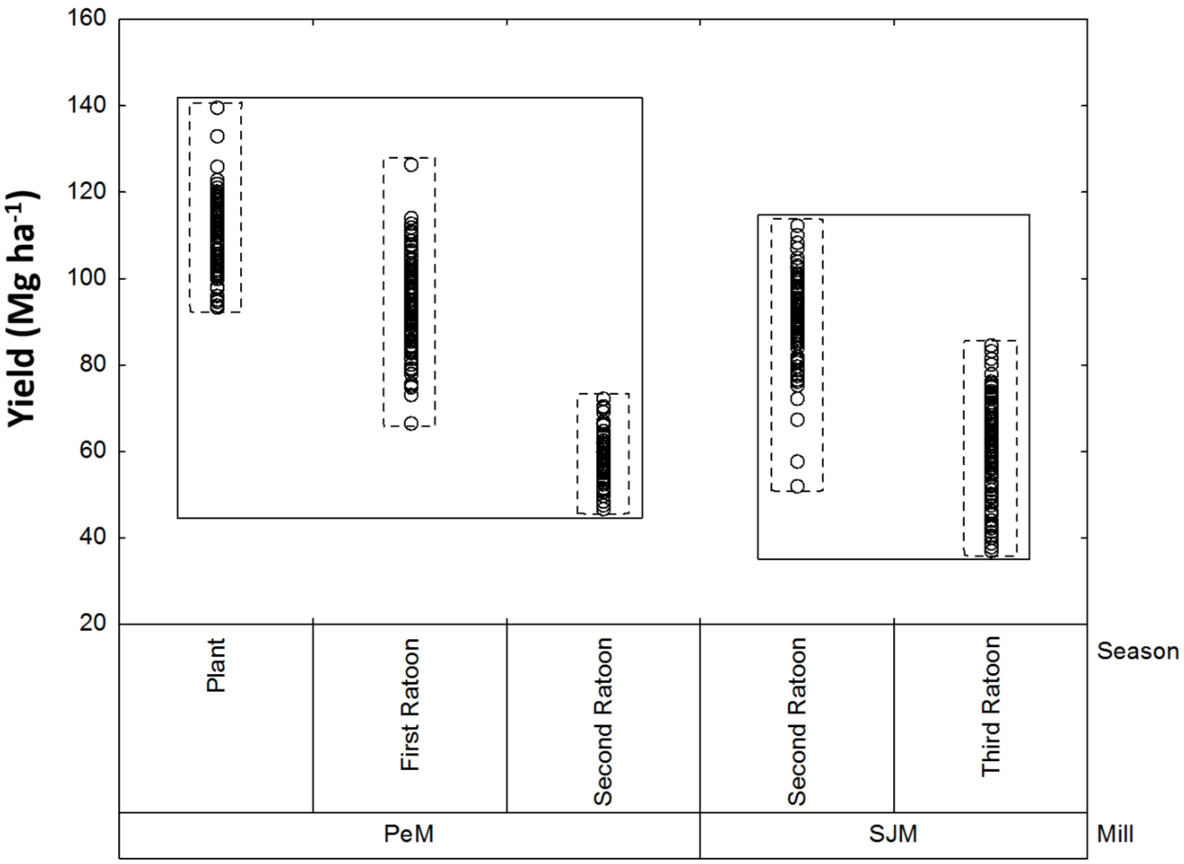

Fig 4. Variability plot of yield (Mg ha^{-1}) for Pedra Mill (Plant, first and second ratoon) and Sao Joao Mill (second and third ratoon).

Table 1. Pearson Coefficient Correlation of soil attributes and yield by difference between years for PeM (below of the main diagonal) and SJM (above of the main diagonal)

	OM	pH	P	K	Ca	Mg	H+Al	SEB	CEC	BS	Yield
OM	-	0.00	0.29*	0.22*	-0.05	0.17	0.05	0.01	0.03	-0.01	0.11
pH	0.14*	-	0.09	-0.02	0.16	0.32*	-0.55*	0.22*	0.05	0.43*	-0.03
P	0.13	-0.03	-	0.06	0.27*	0.18*	0.00	0.27*	0.26*	0.21*	0.06
K	0.02	-0.01	0.06	-	-0.13	0.16	0.07	-0.03	-0.01	-0.08	0.00
Ca	0.33*	0.24*	-0.01	0.11	-	0.56*	-0.04	0.97*	0.94*	0.79*	0.01
Mg	0.32*	0.04	0.07	0.15*	0.68*	-	-0.21*	0.74*	0.66*	0.68*	0.03
H+Al	0.08	-0.35*	0.13	0.10	0.04	-0.14*	-	-0.08	0.22*	-0.55*	0.08
SEB	0.34*	0.22*	0.01	0.19*	0.99*	0.78*	0.01	-	0.95*	0.84*	0.01
CEC	0.34*	0.06	0.06	0.22*	0.91*	0.65*	0.42*	0.91*	-	0.65*	0.04

BS	0.28*	0.42*	-0.05	0.16*	0.74*	0.71*	-0.53*	0.78*	0.49*	-	0.00
Yield	0.04	0.48*	0.07	0.05	0.07	-0.16*	-0.01	0.04	0.04	0.09	-

* Significant at 5%.

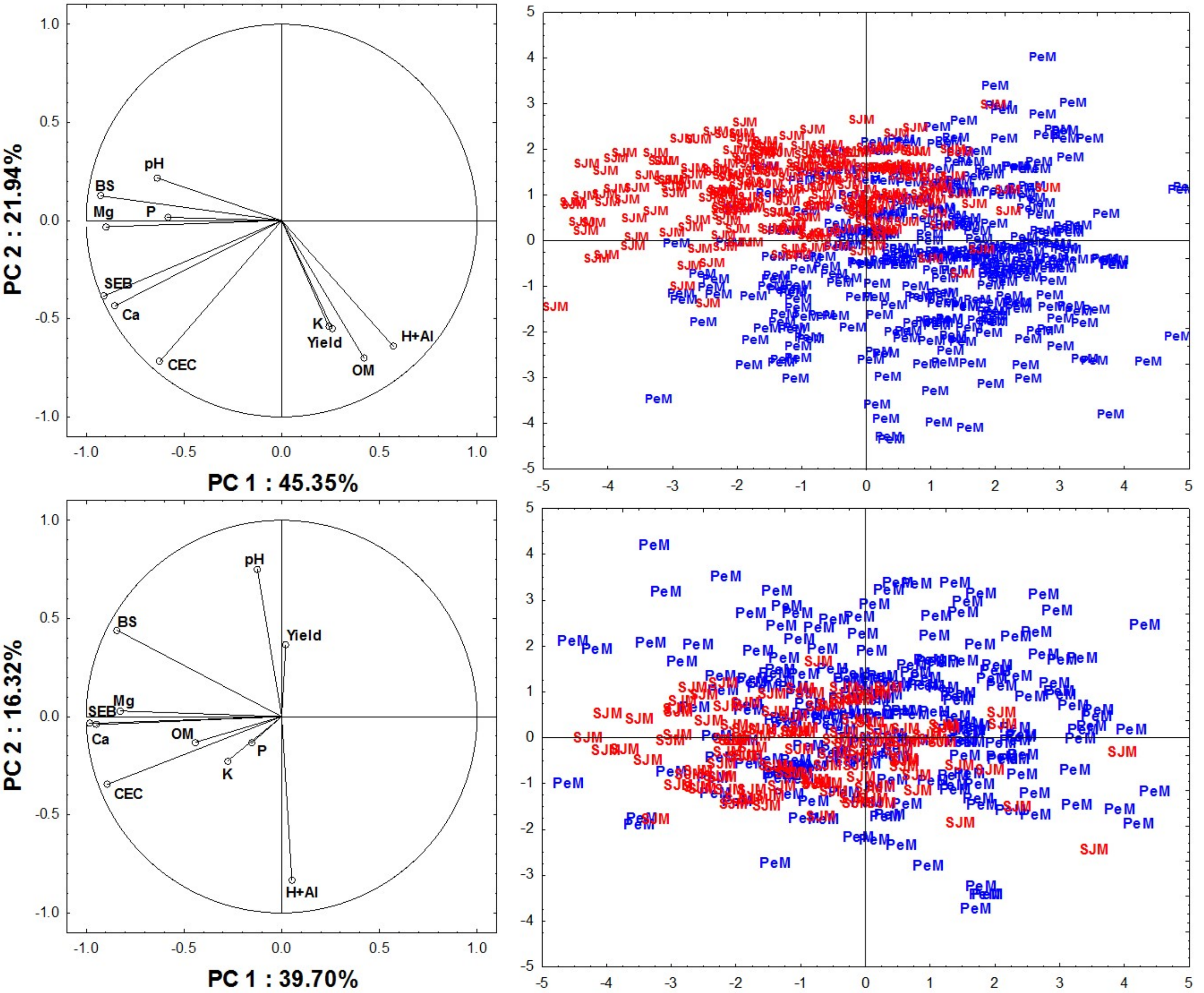

Fig 5. Principal Component Analysis (PCA) of soil and yield data for Pedra Mill (PeM – Blue) and Sao Joao Mill (SJM – Red). Original (above) and difference between years (below) for soil and yield data for all years evaluated. Projection of the variables (left) and cases (right) on the principal component-plane for the first and second principal components.

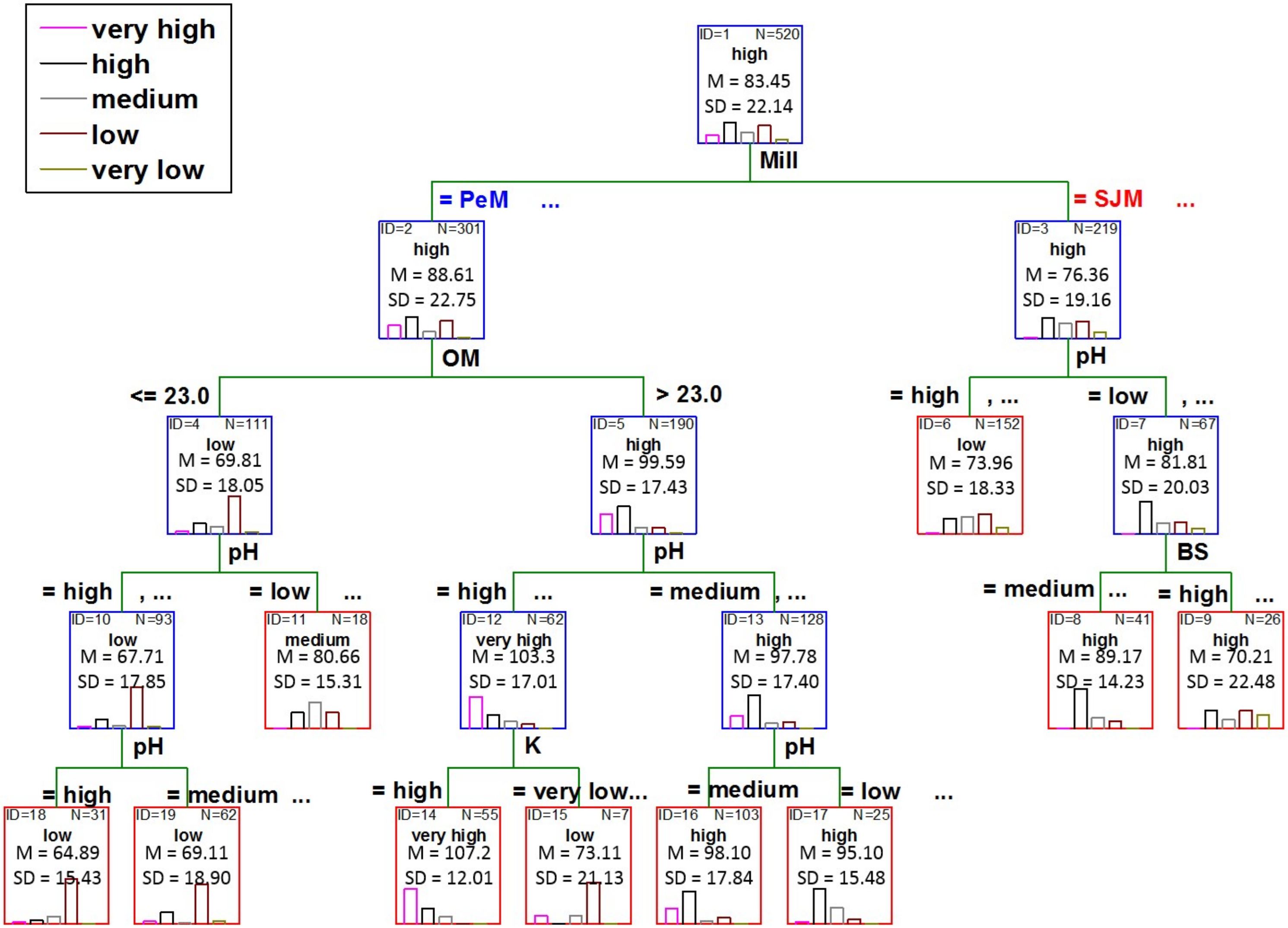
very high
high
medium
low
very low
ID=1 N=520 high M = 83.45 SD = 22.14
Mill
= PeM ...
= SJM ...
ID=2 N=301 high M = 88.61 SD = 22.75
OM
ID=3 N=219 high M = 76.36 SD = 19.16
pH
<= 23.0
> 23.0
= high , ...
= low , ...
ID=4 N=111 low M = 69.81 SD = 18.05
pH
ID=5 N=190 high M = 99.59 SD = 17.43
pH
ID=6 N=152 low M = 73.96 SD = 18.33
ID=7 N=67 high M = 81.81 SD = 20.03
BS
= high , ...
= low ...
= high ...
= medium , ...
= medium ...
= high ...
ID=10 N=93 low M = 67.71 SD = 17.85
pH
ID=11 N=18 medium M = 80.66 SD = 15.31
ID=12 N=62 very high M = 103.3 SD = 17.01
K
ID=13 N=128 high M = 97.78 SD = 17.40
pH
ID=8 N=41 high M = 89.17 SD = 14.23
ID=9 N=26 high M = 70.21 SD = 22.48
= high
= medium ...
= high
= very low...
= medium
= low ...
ID=18 N=31 low M = 64.89 SD = 15.43
ID=19 N=62 low M = 69.11 SD = 18.90
ID=14 N=55 very high M = 107.2 SD = 12.01
ID=15 N=7 low M = 73.11 SD = 21.13
ID=16 N=103 high M = 98.10 SD = 17.84
ID=17 N=25 high M = 95.10 SD = 15.48

Fig 6. CHAID results decision tree for soil (independent variable) and yield (dependent variable) data for Pedra Mill (PeM) and Sao Joao Mill (SJM) for all years evaluated. Yield was divided into five classes – very high (≥110 Mg ha^{-1}); high (90≤y<110 Mg ha^{-1}); medium (70≤y<90 Mg ha^{-1}); low (50≤y<70 Mg ha^{-1}) and very low (<50 Mg ha^{-1}). M – yield mean in the node (Mg ha^{-1}); SD – Standard Deviation of yield in the node (Mg ha^{-1}); N – Number of cases in the node.

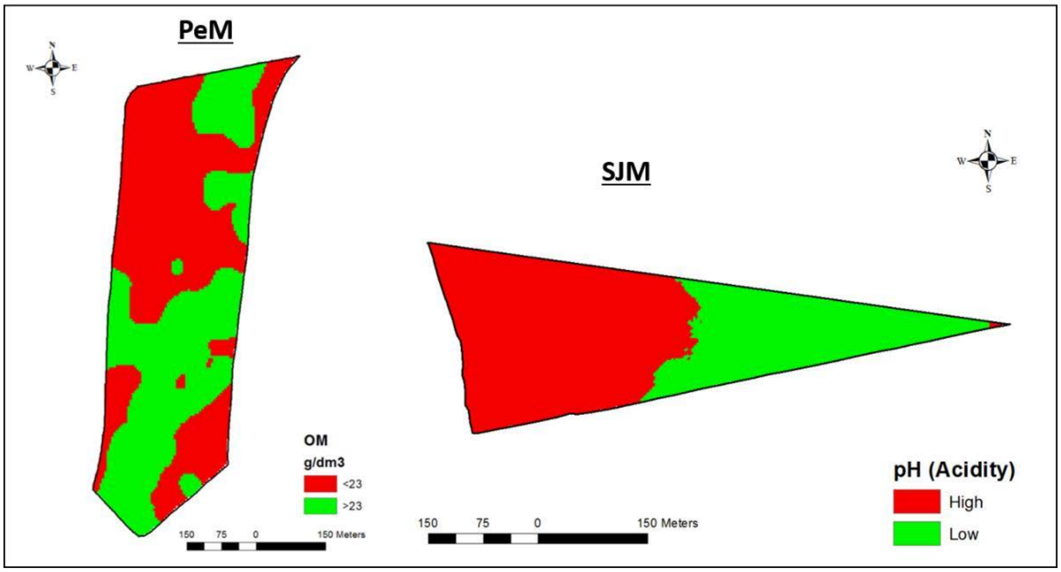

Fig 7. Experimental areas Pedra Mill (PeM) and Sao Joao Mill (SJM) divided into management zones according to the first rule established by CHAID decision tree.