A computational environment to support research in sugarcane agriculture

Carlos Driemeier a, Liu Yi Ling a, Guilherme M. Sanches a,b, Angélica O. Pontes a, Paulo S. Graziano Magalhães a,b,⇑, João E. Ferreira c

a Brazilian Bioethanol Science and Technology Laboratory (CTBE), National Center for Research in Energy and Materials (CNPEM), Caixa Postal 6192, CEP 13083-970 Campinas, São Paulo, Brazil

b School of Agriculture Engineering – FEAGRI, University of Campinas – UNICAMP, Campinas, SP, Brazil

c Institute of Mathematics and Statistics – IME, University of São Paulo – USP, São Paulo, SP, Brazil

Abstract

Sugarcane is an important crop for tropical and sub-tropical countries. Like other crops, sugarcane agricultural research and practice is becoming increasingly data intensive, with several modeling frameworks developed to simulate biophysical processes in farming systems, all dependent on databases for accurate predictions of crop production. We developed a computational environment to support experiments in sugarcane agriculture and this article describes data acquisition, formatting, storage, and analysis. The potential to support creation of new agricultural knowledge is demonstrated through joint analysis of three experiments in sugarcane precision agriculture. Analysis of these case studies emphasizes spatial and temporal variations in soil attributes, sugarcane quality, and sugarcane yield. The developed computational framework will aid data-driven advances in sugarcane agricultural research.

1. Introduction

Sugarcane is an important crop mainly in tropical and subtropical countries. Brazil is the largest sugarcane producer, with 9 Mha cultivated to produce 659 million Mg of sugarcane in the 2015/2016 season, resulting in 34.600 Mg of sugar and 29 billion L of ethanol (CONAB, 2015). In addition to sugar and ethanol, Brazil is today the country with the largest installed capacity of biomass-based electricity generation (IRENA, 2015). In 2015, the supply of electricity from biomass had estimated growth of 7%, with a total generation over 22 TW h, where sugarcane accounts for 80%.

Several modeling frameworks such as AUSCANE, QCANE, APSIM, MOSICAS and CANEGRO (Marin et al., 2011) are increasingly being employed to simulate biophysical processes in sugarcane farming systems. They are all dependent on databases, exemplifying the many ways in which agriculture is moving towards intensive data acquisition and processing. In addition, agriculture worldwide is witnessing a growing adoption of the so-called Precision Agriculture (PA), which comprises a set of tools to help farmers understand and manage soils and crops inherent spatial and temporal variability. PA relies on collection, analysis, processing, and synthesis of voluminous georeferenced data, which can be collected from a number of different technologies (Zamykal and Everingham, 2009). Research and technology on PA have advanced considerably in the past 20 years (Bramley, 2009). Due to its intense use of information, PA has grown and evolved to incorporate the best of multidisciplinary science and technology (Zamykal and Everingham, 2009), requiring farmers to look at their business from different perspectives (Srinivasan, 2006).

Sugarcane production system, however, differs substantially from major staple crops, affecting development and adoption of agricultural technologies. Comparison between a major cereal (e.g., wheat) and sugarcane highlights some key differences. Wheat area worldwide is 215 Mha, primarily in temperate zones, compared to 26 Mha of sugarcane, primarily in tropical developing countries, especially in Brazil. Furthermore, the harvested part of cereal crops is the grain, with mean yields of 3.2 Mg ha⁻¹ for wheat, compared to 71 Mg ha⁻¹ for harvesting the stalks of sugarcane in 2013 (FAO, 2015). Differences in area and location make sugarcane a small fraction of the global market for agricultural technology. In addition, the high tonnage of sugarcane requires dedicated...
technologies, such as tailored yield monitors (Magalhães and Cerri, 2007). Due to specificities of the sugarcane system, and despite rapid adoption of auto-steer in tractors and harvesters (Bramley and Trengove, 2013; Silva et al., 2011), PA is not yet adopted by the sugarcane-based sugar-ethanol industry as it is for other agricultural systems (Gebbers and Adamchuk, 2010). According to surveys conducted in Brazil (Anselmi et al., 2014; Avanzi et al., 2014; Silva et al., 2011) and Australia (Bramley and Trengove, 2013), low PA adoption can be explained by four factors: relative advantage (usefulness), compatibility, trialability and observability. For sugarcane production, perceived usefulness is correlated with increased crop yield, reduced costs, and improved management. On the other hand, high costs of equipment, lack of qualified staff and lack of information on PA technologies were pointed by sugarcane farmers as the main barriers (Silva et al., 2011).

In this context, efforts have been primarily dedicated to experiments aiming at establishing the scientific grounds and demonstrating the advantages of PA techniques applied to sugarcane (Portz et al., 2011; Rodrigues et al., 2012). Because of these goals, characterization of soil and plant attributes in experiments is much more comprehensive than the expected for large-scale PA practice. Furthermore, testing data acquisition technologies and contextualizing their outputs are important goals of the experimentation stage. Considering the above, treating the diversity of measurable attributes is a critical point in experimentation for sugarcane PA.

The data-driven character of PA has attracted the attention of the research community from many different areas. For instance, there are studies on clustering algorithm to delineate management zones (Tagarakis et al., 2012), data acquisition techniques with remote sensing (Mulla, 2013; Song et al., 2009), and software architecture for data analysis and integration of sensor based PA monitoring (Chen et al., 2015).

In this work, we present a computational environment created to support sugarcane agricultural research, including but not limited to research in PA. Data acquisition, formatting, verification, storage, and analysis are discussed. To demonstrate the applicability of the computational environment, data of soil chemistry, sugarcane quality, and sugarcane yield from three experiments are jointly analyzed and discussed.

2. Computational environment

2.1. Handling of raw data

Sugarcane agricultural experiments may include several sources of raw data, including data acquired by different analytical laboratories and by various types of sensors. The current version of the computational environment is able to process data in matrix formats. Processing of images (e.g., from unmanned aerial vehicles and satellite) is foreseen as a future upgrade for the system.

The database has an expandable set of allowed matrix formats – essentially one matrix format for each type of measurement. To assure that matrices are properly recorded in the database, we routinely handle raw data following the tasks presented in Fig. 1.

Using spreadsheets, raw data from sensors and laboratory files are converted into data matrices consistent with the predefined database formats. Such data matrices are verified and then inserted into the database. Importantly, data acquisition and formatting are performed by agricultural field scientist, while verification and insertion are performed by computational workers. Among other advantages, this division of tasks assures an independent verification of data veracity. Verifications include matrix formats, measurement units, and typical range of values acceptable for a certain measured attribute. Once verified, data matrices are inserted into the database using python-generated SQL scripts.

2.2. Database

A relational database for sugarcane agricultural experiments was created and named BDAgro – CTBE Database of Agricultural Experiments, as detailed in a Technical Memorandum (Pontes et al., 2014). BDAgro was constructed having PostgreSQL as relational database management system and pgAdmin as database administration and development platform.

BDAgro conceptual model, i.e. entity-relationship model (Elmasri and Navathe, 2010), includes entities associated to management and responsibilities (e.g., records of projects and responsible persons). Nevertheless, more relevant for the analytical purposes of the computational environment, BDAgro represents agricultural experiments through the following entities:

- **Experiment** is defined by a certain land area during a certain period of time. The land area is most often an open agricultural field, but may also be inside close environments such as greenhouses.
- **Event** is one important fact within one experiment. Events may be of three types: (i) **intervention**, associated with change in experimental land area (e.g., harvest, soil fertilization); (ii) **characterization**, associated with data acquisition without change in land area (e.g., characterization of soil attributes); and (iii) **planning**, representing a record associated with neither physical change in land area nor new data acquisition (e.g., nutrient prescription).
- **Static data** is data generated by events. It is termed static because each event is defined at a specific moment within one experiment. Static data has x and y spatial coordinates as attributes. Additional attributes depend on type of static data (i.e., on type of measurements). Soil attributes, sugarcane quality, and sugarcane yield are examples of types of static data.
- **Dynamic data** is data acquired continuously during the course of one experiment. Meteorological information is one example of dynamic data.

We will refer to these entities as the article follows. However, the analysis of the case studies does not yet include any dynamic data because of the complexity of agricultural analysis using fine temporal granularity.

2.3. Data analysis

We adopted the Work-Event-Data-flow (WED-flow) approach (Ferreira et al., 2010) as the methodology for modeling analysis.
workflows. WED-flow addresses the integration of three main flow paradigms: workflow, event-flow and data-flow. These three flow paradigms combine the concepts of workflow composition, transactions (i.e., activities), events, and data states. At an abstract level, the WED-flow is modeled as a SAGA (Garcia-Molina and Salem, 1987), composed from SAGA steps, each of which is enclosed in a transaction. The main advantage of WED-flow approach is the integration of three main flow paradigms supporting modeling of analysis workflows. More concretely, in this paper, we assume that: the flow of work is a set of analysis steps; the flow of analysis event is a set of constraints (that once satisfied triggers the following analysis step); and flow of data is a set of data states generated by each analysis step. An analytical module was created within BDagro by constructing a set of tables containing the data states resulting from analysis workflows. The statistical computing integrated to the database is performed with the R programming language.

3. Case studies

Data presented in this paper has been collected from 2007 to 2014 in different sugarcane fields. Sugarcane is a semi-perennial crop, which typically grows in cycles of four to six years, being harvested and fertilized annually.

3.1. Three Experiments

The first and second field experiments were conducted from 2010 to 2014 in two adjacent 50 ha areas in a commercial sugarcane field in Serra Azul – SP, Brazil (21° 16’ 41”S and 47° 32’ 10”W), which belongs to Pedra Mill. Before sugarcane planting, a survey of the area was carry out in November 2010 to establish the soil chemical and physical conditions and nutrient need for crop implementation. In the first field experiment (Fig. 2a), the area was divided into a regular 50 m grid with 204 sample points located using a differential global positioning system (DGPS) (Ag114™, Trimble, Navigation Ltd, Sunnyvale, CA, USA) and a yield monitor (Simpro- cana, Enalta, São Carlos, Brazil). Annually, after harvest, soil samples were taken again at same grid to diagnosis some deficiency and recommend fertilizer application using VRT when applicable (field 1 and 2). At each field, the experiment has been conducted for the whole sugarcane cycle, i.e. from planting, following through the consecutive ratoons, during 4 cycles.

3.2. Data acquisition techniques

Soil and crop data have been acquired with two types of techniques: (i) sampling at grid points, for physical and chemical soil attributes and (ii) scanning the area by soil (soil apparent electrical conductivity) and crop “on-the-go” sensors (crop spectrum reflectance), which are mounted on vehicles equipped with differential global positioning system (DGPS). Data acquisition with on-the-go sensors has relatively lower costs and is promising for large-scale PA practice. On the other hand, sampling at grid points is labor intensive, requiring manual collection of samples for offline laboratory analysis. Such sampling approach is likely of limited applicability for large-scale PA. However, for the purpose of experimentation, sampling-based techniques are employed because they allow measurement of much more diverse sets of soil and plant attributes, thus expanding the scope of the experiments.

Sample collection at grid points is recorded in the database as events of characterization, which generate static data. Measured attributes of soil chemistry include pH, contents of organic matter (OM), and concentrations of macronutrients (P and K) and micronutrients (B, Mg, S, Ca, Mn, Fe, Cu, and Zn). Soil samples were...
collected from two soil layers: 0–0.20 and 0.20–0.40 m deep in experiments 1 and 2; 0–0.20 and 0.20–0.50 m deep in experiment 3. Sugarcane quality parameters (Fiber, Brix, and Pol) were measured from sugarcane samples collected immediately before harvest. Sugarcane yield was measured by an on-the-go yield monitor that determines sugarcane yield during the harvesting operation (Magalhães and Cerri, 2007). In the database, yield is a type of static data associated with a characterization event. Each event of yield characterization is simultaneous to the intervention event of harvest (Section 2.2). In this article, we analyze sugarcane yield, whereas we omit data from other on-the-go sensors (of apparent electrical conductivity, crop vegetation indexes).

3.3. Analysis workflow

We developed a data analysis workflow aiming at providing synthetic views of (i) spatiotemporal variability in each measured attribute and (ii) correlations among soil attributes that are presumably associated with soil pH. The model of the analysis workflow is sketched in Fig. 3, representing flow of analysis (represented by boxes), flow of events (represented by arrows), and flow of data states (represented by databases). Note that results of analysis (Fig. 3) are correlations, spatial autocorrelations, and outputs from principal component analysis (PCA). In essence, all these results are based on correlations, which would be detrimentally affected by nonlinearities and outliers.

In order to avoid such issues, linearization and filtering steps are applied beforehand in the workflow. Linearization takes the logarithm of concentrations of soil components (OM, macronutrients and micronutrients), keeping other attributes unchanged. The logarithm scale reduced the positive skewness from concentration distributions and is additionally justifiable due to ubiquity of linear relations with log (concentration) found in physical-chemistry (Atkins and Paula, 2010). Outliers in data sets are removed by a filtering step. Any entry deviating from the mean by more than three standard deviations (for a given attribute) was treated as an outlier.

Following linearization and filtering, the next steps aim at representing all measurements of one experiment as a matrix of entries \( U = \{ u_{ij} \} \), where \( i \) is the index for grid points and \( j \) for measured attributes. Sampling at grid points straightforwardly generates data in this matrix format. Nevertheless, in case of proximal measurements, such as those performed in grid refinement points (Fig. 2), \( u_{ij} \) is calculated as the average of the proximal points. On the other hand, each attribute from on-the-go sensor is a function \( u_{i}(x,y) \), where \( (x,y) \) are the spatial coordinates of the sensor track on the field. The value of \( u_{ij} \) is estimated by employing linear regression to fit a plane to \( u_{i}(x,y) \) points within a circle of 50 m diameter centered at the grid point \( i \).

With attributes estimated at grid points, Moran’s I spatial autocorrelation, \( I_j \), is calculated by considering the connection matrix \( M \) whose terms equal one for neighbor grid points, and zero otherwise. The matrix \( L \) is obtained by normalizing \( M \) to have each row summing to unit. Then, \( I_j = z_j^T L z_j \), where \( z_j \) is the vector of the mean-centered normalized attribute estimated at grid points, \( z_j = (u_{ij} - \bar{u}_j)/\sigma_j \) (Cliff and Ord, 1973).

Measured attributes vary widely concerning their range of values as well as their data acquisition technologies. Therefore, precision in attribute measurement also vary widely. Typical attribute precisions were estimated as “noise” levels \( s_j \). For attributes determined by sampling at grid points, \( s_j \) was estimated as the standard deviation of measurements in proximal points. For attributes obtained from on-the-go sensors, \( s_j \) was estimated from the standard error of the linear regression coefficient that estimates \( u_{ij} \).

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the attribute space. Prior to PCA, missing data in matrix \( \{ u_{ij} \} \) was imputed by the Expectation-Maximization (EM) algorithm associated with a multivariate normal model (Johnson and Wichern, 2007), stopping EM iterations when imputed \( u_{ij} \) change way less than \( s_j \). This imputation approach preserves the data covariance structure, thus being well suited as data preparation method for PCA. PCA was performed from correlation matrixes, which means that each attribute contributes with normalized, unit variance.

4. Results and discussion

4.1. Spatial autocorrelations

The “null hypothesis” of PA is uniform management of the field (Whelan and McBratney, 2000). A corollary is that, to be relevant for PA, a measured attribute should present positive spatial autocorrelation. That is, an attribute \( j \) having spatial autocorrelation \( I_j \approx 0 \) is consistent with random spatial distribution and, therefore, cannot be used to justify any site-specific action on the field. On the other hand, greater values of \( I_j \) might justify site-specific interventions.

Considering the diversity of attributes characterized in the three experiments, spatial autocorrelation \( I_j \) is distributed between \( \approx -0.2 \) and \( \approx 0.8 \) (Fig. 4). Negative \( I_j \) is observed mainly in experiment 2 (Fig. 4B). Most likely negative \( I_j \) can be interpreted as \( I_j \approx 0 \) perturbed by statistical noise, which are greater in experiment 2 due to fewer grid points (24) (Fig. 2). Noteworthy, Webster and Oliver (2007) recommend that no fewer than 100 sampling points are necessary to estimate the variability of a certain field. Furthermore, it is important to note that, for the three experiments, many attributes are close to \( I_j \approx 0 \) so that such measured attributes should not be employed to support site-specific interventions on the field.
4.2. Spatiotemporal variability in soil attributes

Statistical noise tends to reduce the magnitude of all types of correlations, including spatial autocorrelation and correlation between sequential measurements of a given attribute. Fig. 4 plots inter-year correlation against spatial autocorrelation $I$. The inter-year correlation is an index for the temporal stability of the spatial variability of a given attribute. For attributes of soil chemistry (black labels), a clear trend is observed in experiment 1 (Fig. 4A), with attributes clustered near the diagonal dashed line ($y = x$). Indeed, inter-year correlation is positively correlated ($r = 0.73$) with spatial autocorrelation. This trend is consistent with inter-year correlation and spatial autocorrelation being both substantially reduced by statistical noise. The correlation between inter-year correlation and spatial autocorrelation is weaker but still positive in experiments 2 and 3 ($r = 0.54$, $r = 0.32$, respectively). This result indicates that statistical noise is also a major factor reducing inter-year correlation and spatial autocorrelations in experiments 2 and 3. In these two experiments, most attributes of soil chemical attributes are above the diagonal dashed line (Fig. 4B and C), i.e., for most attributes inter-year correlation is greater than spatial autocorrelation. This observation might result from inherently lower spatial autocorrelations in experiments 2 and 3, due to the greater spacing between grid points in experiment 2, and perhaps due to the more homogeneous, flat terrain of experiment 3.

4.3. Spatiotemporal variability in sugarcane quality

In experiments 1 and 2, attributes of sugarcane quality have spatial autocorrelation and inter-year correlation of approximately zero (blue labels in Fig. 4A and B, respectively). This observation indicates that sugarcane quality is consistent, at least approximately, with random spatiotemporal variations. Experiment 3 shows a different behavior. In year 2011, Brix and Pol (quality attributes associated with sugar content) have spatial autocorrelation $l_j \approx 0.3$ and correlation with previous year of $\approx 0.2$ (Fig. 4C). These are significant correlations, demonstrating that within-field spatial variability and temporal stability of Brix and Pol are possible issues to be managed by PA techniques. Correlations of even higher magnitudes are observed in sugarcane fiber contents (‘Fbr’ labels in Fig. 4C), reinforcing that spatiotemporal management of sugarcane quality is possible, at least in principle. Nevertheless, the fact that appreciable correlations and spatial autocorrelations are only occasionally observed in one (experiment 3) out of three experiments suggests that significant spatiotemporal variability in sugarcane quality may often remain undetected.

4.4. Spatiotemporal variability in sugarcane yield

Yield is certainly a major aim in crop management. In other crops, yield spatial variations are commonly found to be rather stable across several harvests (Godwin et al., 2003). In sugarcane, however, we have been observing a lack of temporal stability in yield spatial variability. In experiment 1, yield in 2014 has a positive ($r = 0.2$) correlation with yield in 2013. This level of correlation locates yield close to the diagonal trend observed in Fig. 4A, which would be consistent with temporal stability reduced mainly by statistical noise. On the other hand, yield in 2013 has a near zero correlation with yield in 2012 (Fig. 4A). This yield data is quite below the diagonal trend of Fig. 4A and the lack of inter-year correlation cannot be attributed to statistical noise. Even more extreme case is observed in experiment 2, where correlation between 2012 and 2011 yields is negative ($r = -0.6$).

A more detailed investigation indicated two major causes for the zero or negative inter-year correlations in yield. One cause is occasional sensor failure. Readings from yield monitor are very noisy (Fig. 5a), even after removal of outliers (Fig. 5b). The example of Fig. 5b has mean yield of 93 Mg ha$^{-1}$ and standard deviation of 47 Mg ha$^{-1}$. Typical precision ($s_p$) in yield is reduced to about 1–4 Mg ha$^{-1}$ due to multiple readings used to estimate $u_y$ by regression (Fig. 3). However, with such noisy signal, some sensor failures remain unfiltered, detrimentally affecting the precision of yield estimates and reducing associated correlations. The second cause of low inter-year correlations originates from real changes in yield, especially due to damages inflicted to the sugarcane crop. Harvest operation may damage the roots of the sugarcane plants, occasionally pulling roots out of the soil. The result of such damage is a localized yield reduction in the next harvests, thus reducing inter-year correlation. Such damages to roots may become visible as gaps in the sugarcane field. Indeed, field observations by agronomists have been suggesting that crop damaging during harvest is the major cause for declining yield across the multiyear crop cycles (Zamykal and Everingham, 2009).
The numerous soil and plant attributes exist in an abstract space of high dimensionality where visualization of the information is impossible. Hence, reduction of dimensionality is often desirable and PCA is a statistical technique to do it. Fig. 6 presents the PC1 & PC2 loadings from PCA applied to attributes judged to be potentially associated with soil pH. The attributes are pH itself, the main elements of lime (Mg and Ca, applied to soils to increase their pH), and organic matter (OM, whose decomposition is thought to decrease soil pH). PCA was applied separately for each experiment. Analyzed attributes span sequential years as well as the two soil layers that were characterized. Our interest in soil pH is due to the possibility of pH control using lime broadcasted on the soil surface with variable rate technology, which could reduce costs by applying lime prescribed according to site-specific demands.

As a first observation, experiments 1 and 3 show loadings that are clustered (Fig. 6A and C). In each experiment, there is one cluster for pH, one for lime elements (Mg and Ca), and one for OM. Such clustering indicates high correlation between soil top and bottom layers. The clustering also indicates that spatial variations of pH, Ca, Mg, and OM are mostly preserved along the years. This is instructive and perhaps surprising, considering that experiments 1 and 3 employed distinct strategies to control soil pH (Section 3.1). Loadings of OM are comparatively more spread than loadings of pH, Ca, and Mg (Fig. 6A and C). This observation is consistent with noisier OM measurements as well as with significant dynamics in spatial variability of soil OM. Experiment 2 does not present such clustering of loadings (Fig. 6B), which might be due to poorer statistics arising from fewer grid points (Fig. 2).

Experiments 1 and 3 also share the relative directions of loading clusters. The centroids of the pH clusters form an angle slightly greater than 90° with the centroids of OM clusters (Fig. 6A and C). This relative direction indicates a tendency to have slightly negative correlations between pH and OM, consistent with decreasing pH due to higher levels of decomposing OM. Furthermore, the loading clusters of Ca and Mg are observed in-between those of pH and OM (Fig. 6A and C). Proximity with loadings of pH is consistent with Ca and Mg causing increases in soil pH. On the other hand, the proximity between loadings of OM and of lime elements (Mg and Ca) is consistent with OM acting as storage medium for Mg and Ca available in soils. These relative directions in loadings plots demonstrate the possibility of finding common behavior across multiple sugarcane agricultural experiments.

### 5. Conclusions

This paper described a computational environment to support research in sugarcane agriculture. Data acquisition, formatting, and verification steps are performed prior to data insertion into a dedicated database named BDAgro. Data analysis is integrated to the database by recording data states generated through data analysis workflows. Such workflows can be tailored for the specific aims of each study.

The computational environment was employed to jointly analyze three experiments in sugarcane precision agriculture (PA),
These experiments comprised characterization of soil attributes, sugarcane quality, and sugarcane yield. Our analysis showed low spatial autocorrelations and low inter-year correlations for most of the sugarcane quality and yield attributes. This finding revealed important limitations in measurements of sugarcane quality, while yield seems to be primarily impaired by crop damages inflicted by the mechanical harvest processes, such as ratoon damage and soil compaction. Furthermore, our multivariate analysis of features associated with soil pH revealed common behavior in two experiments, evidencing the possibility of common underlying principles to be identified across multiple sugarcane agricultural environments. Overall, these case studies demonstrate the usefulness of the computational environment for supporting research in sugarcane agriculture, which will positively impact the sustainability and competitiveness of the sugar–ethanol industry.

Acknowledgements

Project funded by CNPq and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP grants 2011/01817-9 and 2015/01587-0).

References


