Photo: curto (AdobeStock)

Scientific computing in agriculture

5

Sônia Ternes | Maria Fernanda Moura | Kleber Xavier Sampaio de Souza | Glauber José Vaz | Stanley Robson de Medeiros Oliveira | Roberto Hiroshi Higa | Helano Póvoas de Lima | Celina Maki Takemura | Enilda Coelho | Francisco Ferraz Laranjeira Barbosa | Marcos Cezar Visoli | Gilberto Romeiro de Oliveira Menezes | Luiz Otávio Campos da Silva | Sandra Aparecida Santos | Silvia Maria Fonseca Silveira Massruhá | Urbano Gomes Pinto de Abreu | Balbina Maria Araujo Soriano | Suzana Maria Salis | Márcia Divina de Oliveira | Walfrido Moraes Tomas

Introduction

Digital technologies have advanced incredibly fast, and its recent development has stimulated the acquisition of large volumes of different types of data, from the most varied sources. In agriculture, along its value generation chain, this data can include: a) omics data (genomics, proteomics, transcriptomics and metabolomics); b) acquired physicochemical attributes with spatiotemporal location through sensors; c) aerial and satellite images with spatiotemporal location; d) socioeconomic data; among others.

Similar to the data from more traditional sources, the use of this large data volume is analyzed through models and algorithms capable of extracting useful information for the decision-making process. This can occur in the development of a new biotechnological asset, land use monitoring, or in the control of a production process. Thus, scientific computing is understood as a collection of techniques, tools, and theories that encompass mathematics, statistics, physics, and computing. It also covers specific knowledge of certain sub-areas, such as applied statistics, econometrics, applied mathematics, computational intelligence, scientific visualization and biometrics. These will continue to be central in the development of new agricultural technologies, now in the context of the emerging Digital Agriculture. In recent decades, scientific computing has been identified as the third pillar of scientific research, along with experimentation and theory (Souza et al., 2017).

In the following sections, we present examples of applications that use scientific computation algorithms and techniques for the solution of problems in the agricultural sector. Section 2 presents two applications. They are based on the observation of a large raw dataset in order to recognize embedded patterns and derive knowledge and actions from such patterns to be used by an Expert System. Section 3

presents three applications based on the construction of mathematical and statistical models. These can carry out predictions and analyses from simulation scenarios in order to assist public decision-making. Through these different applications, it is possible to see that scientific computing is a research area that is eminently transversal to others.

Artificial intelligence

Artificial Intelligence is a broad area that began in the second half of the 1940s, when an artificial neural network was conceived, and which described how human neurons should learn to perform calculations. This area has undergone many modifications and has intersected with other disciplines, especially statistical modeling and various pattern recognition methods. These intersections compose a group of techniques known as intelligent systems, which are based on machine learning.

A machine learning model is supported by previously observed data coming from either databases, experiments, images, or texts. Data has attributes, which need to be described for each observation. For example, if we collect data from a pasture at different locations on the property, we will have the same attributes for each data collection, such as: location, grass type, date, pasture status (degraded, non-degraded, in degradation), geographic location, percentage of soil cover, soil type, pasture height, etc. With these attributes and the data collected, a classification model of the pasture's status can be built. If the observed data were texts, the attributes could be the words in the texts; if they were images, the attributes could be such images divided into very small pieces, or pixels, and could consider, for example, the color of each pixel.

Items "Automatic Soil Classification" and " SiBCS-based Expert System" present, respectively, examples of automatic soil classification and exploration of information in texts, making use of different artificial intelligence techniques.

Automatic soil classification

To classify a soil profile, the Brazilian Soil Classification System (SiBCS) considers a wide range of morphological, physical, chemical, and mineralogical attributes in addition to environmental aspects such as climate, vegetation, relief, parent material, hydric conditions, external characteristics and soil-landscape relationships (Santos et al., 2013).

To assist in this laborious process, Embrapa Digital Agriculture and Embrapa Soils designed two intelligent tools for automatic soil classification. The first is an Expert System that uses the SiBCs rules for soil classification. The second is aWeb system (SoloClass) for soil profile classification through a committee of intelligent solutions based on machine learning algorithms. These smart tools were developed within the scope of the project "Use of smart mobile devices in the classification of Brazilian soils – SmartSolos", led by Embrapa Soils. Both tools are presented in the following subsections.

SiBCS-based expert system

The SiBCS rules-based expert system simulates the reasoning of a domain expert when performing the classification of soil profiles. Thus, it can be used to classify soil profiles not yet classified and to validate previously classified profiles (Vaz et al., 2018).

Vaz et al. (2019a) used the expert system to analyze soil data provided by IBGE. They showed that this is an important tool for the curation of Brazilian soil data, as it allows it to be executed more efficiently and with fewer errors, benefiting soil governance in Brazil.

The advantages of making the expert system available through an API and the importance of this tool to facilitate soil data curation, while guiding a more adequate data recording, were also shown in Vaz et al. (2019b). Figure 1 shows that by making the expert system available through the API, the user can obtain the soil profile classifications from the expert system and compare them with previously known classifications. Thus, possible errors in soil data can be analyzed and corrected, making it a powerful tool for improving the quality of soil data in Brazil.

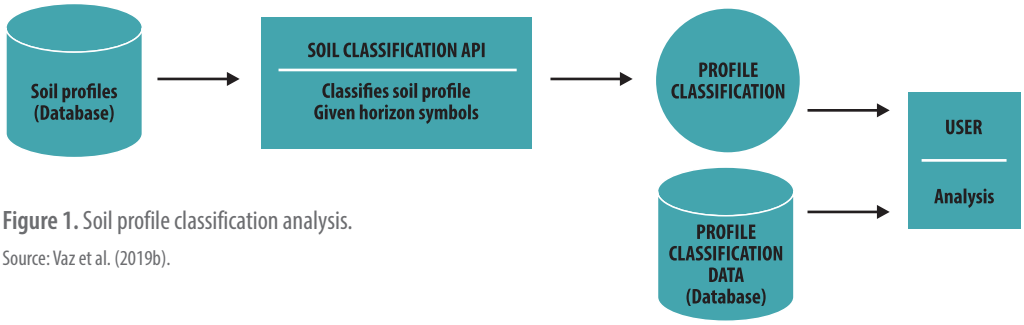

Figure 1. Soil profile classification analysis.
Source: Vaz et al. (2019b).

The great challenge of this system is codifying all SiBCS rules to treat its first four categorical levels. The classification taxonomy has more than a thousand classes between the first and fourth categorical level. In addition, the rules are quite complex, so joint work and great effort by computer and soil scientists are essential to make viable the development of such a system.

Although a specific application is being developed by Embrapa to use this soil classification API, partner institutions can also use it to create new solutions that rely on soil classification, provided their data is coded in accordance with the standards established by the system.

Regarding Brazilian soil data standards, there are different initiatives that seek to organize them. However, none of them has been consolidated as a standard, nor do they meet the needs of the expert system developed. As such, many observations could be made in relation to organizing these data in Brazil throughout this work. It is common, for example, to observe data redundancy in different fields, absence of fields necessary for recording important soil information for classification, and data representations that make difficult computational processing and data retrieval. The next step of this work is, therefore, to consolidate a series of recommendations for the structuring of Brazilian soil data in order to simplify computational manipulation, ensure higher quality of stored data, and facilitate the creation of new solutions that depend on them.

Future research is on the possibility of automating other processes that are normally time consuming or greatly increase the uncertainty of the data collected in the field. For example, color, texture, soil layer boundaries and other attributes are determined in a subjective way, according to personal interpretations made during fieldwork. The collection of this type of information can be facilitated and automated through computational tools that extract characteristics from images taken in the field.

Intelligent soil classification system

A promising alternative for automatic soil classification is the combination of machine learning (ML) algorithms with attribute selection methods. ML algorithms operate by building a model obtained from training samples to make data-driven predictions. Such data contain soil profile observations previously classified by pedologists. On the other hand, the attribute selection methods aim to find a subset of relevant variables related to the target task. It makes the learning process more efficient by simplifying the operating cost of the models, enabling to better understand the obtained results (Guyon; Elisseff, 2003).

SoloClass is an intelligent system developed for classifying soil profiles. This system allows a user to input a set of variables from one or more soil profiles, and then receives the classification of each profile according to SiBCS with a probability associated to the predicted class.

Five classes of ML algorithms were used for intelligent soil classification: a) symbolic: decision trees; b) based on instances: k-NN or k nearest neighbors; c) statistical learning: Support Vector Machines (SVM); d) bootstrap aggregation: Random Forest; and e) connectionism: Deep Neural Networks. All these algorithms were trained for the four categorical levels (orders, suborders, large groups, and subgroups) adopted by SiBCS.

The architecture of the SoloClass system is based on a classifiers committee, as shown in Figure 2.

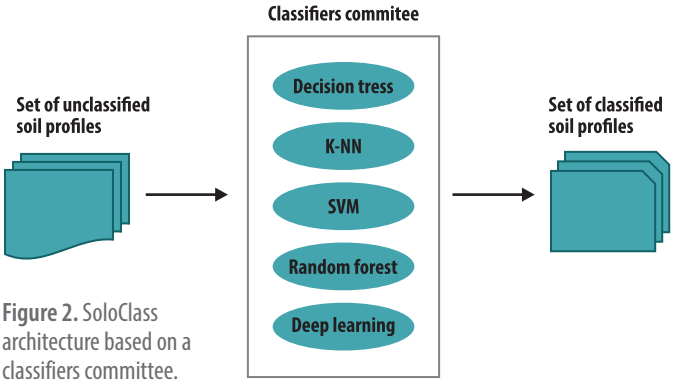

Figure 2. SoloClass architecture based on a classifiers committee.

Upon receiving a set of unclassified soil profiles, with different numbers of horizons, the user can select one or more classifiers that have been trained from a pre-classified database by pedologists (induction process). Subsequently, the system triggers the selected classifiers and stores the results presented individually. At the end of the deduction process, the classification committee (Figure 2) assigns the classification result to the soil profile by vote, that is, the classification associated with the profile is the one that obtained the highest frequency or majority vote.

This classifier committee-based architecture has some advantages such as: a) increase in the predictive power of the system due to the use of several classifiers adjusted to the data and combined for this purpose; b) reduction of variance and bias when compared to using only one machine learning method; c) extensible architecture, that is, other classifiers can be added.

The main benefits of the SoloClass system are: a) assisting national soil survey projects and programs, such as Pronasolos (Polidoro et al., 2016), acting as a facilitating tool in soil classification work; b) facilitate the understanding of soil classification for farmers, students, teachers, extension workers, and researchers; c) minimize possible human errors during the soil classification activity.

As it is a Web System, SoloClass[1] can be accessed through mobile devices or personal computers, without any operating system restrictions. This helps to expand access and the inclusion of a greater number of users. SoloClass has a responsive interface, that is, it can be characterized by the visual adaptation of a page or interface to any device on which it is viewed, without the need of a specific versions for each model.

[1] Available at: www.soloclass.cnptia.embrapa.br

Text mining in technical-scientific publications

The human learning process is based on observations, pattern formation, hypotheses, and inferences from these observations. Nowadays, there are plenty of observations, specifically, an excessive volume of data, both in databases and in published textual format. Data mining uses statistical analysis processes, in which algorithms are implemented in computer programs that can handle a large volume of data to find patterns and help formatting hypotheses and models that allow describing these patterns.

Text mining (TM) is a specialization of the data mining process. The main difference between the two processes is that, whereas conventional data mining works exclusively with structured data (pre-organized in databases or some representation, such as a spreadsheet), text mining inherently deals with unstructured data. Therefore, in TM, the first challenge is to structure the data with their respective attributes, based on the texts, so that data mining algorithms can be used.

The structuring of texts depends on the problem addressed. For example, if we want to know or relate which types of agricultural technologies are linked to the use of water in Brazilian agriculture, we can delimit a set of technical-scientific publications on the topic and extract this information. In this case, one option is the use of linguistic tools that allow identifying the vocabulary of interest (for example: irrigation, harvesting, water resources, pivot, etc.) and delimiting and disambiguating geographic locations (such as: São Francisco River, São Francisco Church, São Francisco City, etc.) in the texts.

Similar to this, in the methodology proposed by Moura et al. (2017), there is a semi-automated step-by-step process, which used software tools specifically developed for this purpose, and contained the following steps: 1) delimitation of publications of interest; 2) extraction and disambiguation of toponyms with the TopExtract tool (Takemura et al., 2013); 3) formatting a dictionary of terms of interest, manually performed by domain experts; 4) use of the ExtracTrans tool (Transaction Extraction tool) to: a) extract terms from texts by similarity and synonymy; b) creation of the transactions present in the texts (all the words of interest that appeared in the text); and c) elimination of redundant data, which does not contribute to the results; 5) pattern extraction, using machine learning algorithms, such as association rules, or even placing the results in an Excel spreadsheet or other similar software. For example, in Moura et al. (2017), 40 association rules were found for the Northeast region, among which:

If technologyClass = agricultural engineering & culture = grapes & cultureClass = fruit & region = NE → technology = irrigation.

Another application of very specific interest was to describe which quantitative and scientific computing methods were cited in Embrapa's scientific publications. We searched among those considered of the highest level, and according to the Qualis CAPES indicator (A1, A2, B1 and B2), between the years 2000 and 2018. Embrapa has its own cataloging system for its publications and technologies (Embrapa, 2020) where the metadata´s keywords are audited. This in itself indicates its very high quality. However, two major problems are present for the study: a) the large number of publications in this interval, approximately 22,000 articles; and b) the fact that the keywords in the articles cover agricultural terms, and not necessarily quantitative methods and scientific computation terms. In other words, the keywords of interest in this data analysis were not part of the conventional keyword repertoire of these articles and, therefore, could not be located only by search results, let alone by reading each of the 22,000, which would be a very extensive task.

Thus, the methodology by Moura et al. (2017) was adapted as follows: 1) the articles of interest were already selected; 2) the geolocation process was not necessary and instead, a process was created to download the articles and convert them to plain text; 3) the domain specialists, in quantitative methods and scientific computation organized the necessary dictionary of terms for the area; 4) a tool was adapted (from ExtracTrans tool) to extract the words of interest from the text collection by similarity, and subsequently, put the data in a spreadsheet; and 5) from the data sheets in Excel format, the techniques of crossing dynamic tables, aggregation of data from other sources, grouping, selection, and filters were applied to facilitate the data exploration in different views.

Some exploratory applications on a large volume of texts make use of a process similar to that used by search engines, such as Google, Yahoo, etc. The textual collection is indexed, in which each text (data) corresponds to a row of a table and each word (attribute) to a column, it is not always necessary to know the language in which the text is written, much less if there are dependencies between words. In each cell, the frequency of a word in the text, or some derived measure, is placed. Therefore, as this table has an exaggerated number of columns and many cells with zero value, we try to reduce the number of columns, selecting the most statistically significant words or word compositions.

There are many techniques to reduce the number of columns in a table, all of which depend on what one wants to answer in relation to the collection of texts. To format a collection of texts in a table like this, we have the I-PreProc tool (Pereira; Moura, 2015). A common application for this formatting is to group texts with similar content so that they must correspond to specific topics, that is, subdivided into more related subjects, as carried out in the Compilation and Retrieval of Technical-scientific Information and Induction to Knowledge (CRITIC@) project. This initiative, developed by Embrapa, also uses other tools such as the previously mentioned TopExtract application.

In Figure 3 on the left, it is possible to see that based on a search expression in the publications database, the search results are organized hierarchically into documents groups from where statistically significant terms found in the group are considered "topics". In the middle, there is the distribution of accumulated frequencies for the group over time. These are represented by "*Tractor, Effect, Term, Difference, Applied, Leaf, Pruning*". To the right, the locations mentioned in these documents. This result of data exploration gives us clues as to: a) how these documents could be organized according to groups; b) what the topics or set of keywords of this group would be, for example "tractor, pruning, pruning applied to the leafs", that is, what an expert in the area considers most important in the presented result; and c) geographic location, more specifically, where these groups appear most significantly.

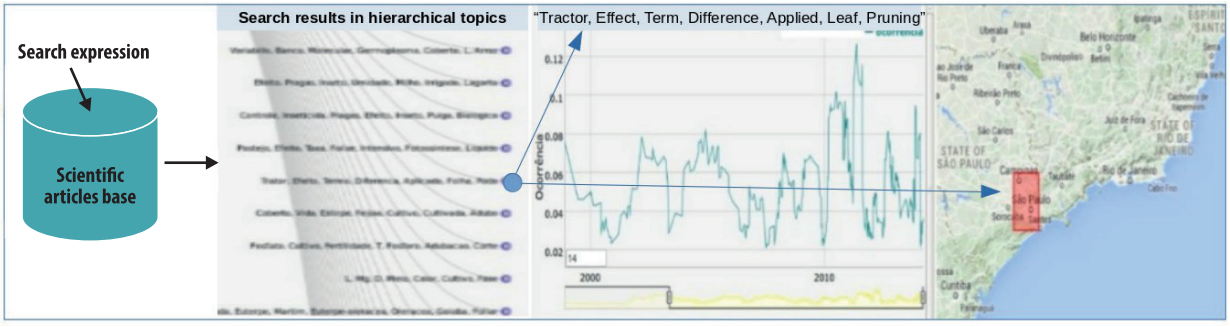

Figure 3. Example of a CRITIC@ project result.

As seen in the cited applications, text-mining processes, whatever the techniques, help the exploration and identification of information in a large volume of texts.

Mathematical and statistical modeling

Mathematical modeling is an even broader area than Artificial Intelligence. It uses a small framework of mathematical solutions, the same occurring with statistical modeling. The general idea for the modeling process is the simplified interpretation of a phenomenon, which is then described in mathematical language. Subsequently, it allows simulations to be carried out in a computer. Thus, the users of the model are positioned as experimenters in the real world, and based on the results of several computer simulations, they can understand details of the phenomenon in situations not experienced in practice. In agricultural research, for example, mathematical and statistical models are essential to complement biological experiments, allowing the study of disease dynamics in the field from computer simulations, that is, without environmental impact and with a great economy of resources. As an example, item "Modelagem da dinâmica de dispersão do HLB do citros" presents a simulation model for analyzing the intra-orchard dispersion of the disease known as citrus HLB.

In order to understand the difference between mathematical and statistical models, it can be considered a simple example, such as the mathematical equation that represents a straight line in a Cartesian plane (x, y), given by:

$$y = a + bx$$

In which a is the slope of this line and b the factor that correlates each value of x to exactly a value of y in this plane. On the other hand, if it is observed the weight and height values of a group of people, it is known a priori that the behavior of the observed points (weight, height) is linear, that is, it can be represented by a line, but it does not correspond exactly to the weight and height ratio of the entire population, that is, this set of points is just a sample of this population. A good sample should be randomized, so each person is randomly drawn for weight and height measures and has a statistically reasonable size. Thus, with this collected sample, the behavior of the population for the problem under study (weight, height) is estimated, which is a line composed of estimated values of the slope of the line (average of the observed values of weight), and the factor that correlates height with weight. This process considers model errors and estimates depend on probability distributions. In item "Genetic evaluation of livestock", a multivariate linear model is presented, that is, several dependent variables (which would replace weight) and several independent variables (which would replace height), and also: a) fixed effects, which are the of the factors that can be observed, such as the height in our example; and b) random effects, which are not observed in the sample collection, but need to be estimated by the model.

Another framework within mathematical modeling is inductive logic models, for example, "if A is a stable then A has horses". Among these models are fuzzy logic. For example, if we have a glass of water, it can be full, half full, half empty, or empty, according to the interpretation of each person looking at the glass. So, you can form rules, such as the glass is empty if it has 0 to 20 mL of water, it is half empty if it has 10 mL to 100 mL of water, it is half full if it has 50 mL to 160 mL of water, and it is full if you have more than 140 mL of water. To solve a classification of how each cup is, a system based on fuzzy rules is developed. Item "Sustainable Pantanal Farm" shows an application of systems based on fuzzy rules that assist decision making regarding sustainability in Pantanal farms, considering environmental, social, and economic values.

Modeling the citrus HLB dispersion dynamics

Brazil is the world's largest producer of oranges, with the 2020/2021 harvest being estimated at almost 288 million boxes (40.8 kg) (Fundecitrus, 2020). The disease known as huanglongbing (HLB) or greening, identified in the country in 2004, is currently the most important for the national citrus industry. Citrus HLB is caused by the bacterium *Candidatus Liberibacter asiaticus* and transmitted in Brazil mainly by the psyllid *Diaphorina citri*, which acquires the bacteria by feeding on the sap of infected plants, later transmitting them to healthy plants.

Due to its importance to the national economy, Embrapa has been developing biomathematical tools to assist in monitoring, sampling, detecting, and eradicating HLB from citrus since 2012. Initially, a deterministic compartmental mathematical model was developed (Vilamiu et al., 2013) to assess the impact on decreasing population levels of the insect vector *D. citri* in the Recôncavo Baiano region. More specifically in areas where citrus and alternative hosts are planted (orange jessamine – *Murraya paniculata*) in different proportions, aiming to collaborate with public policies for the sector. In this study, citrus and myrtle populations were divided into compartments (susceptible, exposed, infected and recovered plants), and the general characteristics of each compartment were expressed through mathematical equations in order to analyze HLB propagation temporal dynamics.

More recently, a new modeling approach based on simulation scenarios with different spatial configurations of orange jessamine and citrus was used to assess, among other aspects, the role of orange jessamine as a push or pull factor on vector insects in cultivated areas (Barbosa, 2015). For this purpose, individual-based modeling (IBM) (Grimm; Railsback, 2005) was used, and considers in the model the presence and particularity of each individual of the populations involved, while observing the final system as the result of interactions between the individuals of different populations. The IBM approach is adequate for the objectives of the study because it allows one to jointly explore the temporal and spatial aspects of the "host-insect vector-HLB" system in a more intuitive and flexible way than classical mathematical models such as those used in Vilamiu et al. (2013).

The IBM was developed in Python programming language and considers a standard agricultural landscape of the Recôncavo Baiano, containing 9 plots with 20 x 42 plants in each plot (total of 840 host plants per plot or 7,560 plants in the landscape), spacing between rows of 6 m and spacing between columns of 4 m, totaling an area of 120 m (width) x 168 m (length), just over 2 ha.

In order to analyze the intra-orchard dispersion of the insect vector and the propagation of HLB, 3 different landscapes were tested and compared: a) Scenario 1: only citrus; b) Scenario 2: citrus and myrtle around the entire area; c) Scenario 3: citrus and myrtle on the edges of each plot. Thus, the populations considered in the IBM and involved in the computer simulations are: a) main host plant (citrus); b) alternative host plant (myrtle) for testing the repulsion and attraction effect; c) *D. citri* insect vector in the nymph stage; d) adult vector insect.

In the execution of the model simulations, the user can choose different values for the following biological parameters obtained from studies and biological experiments conducted at the Embrapa Cassava & Fruits (Cruz das Almas, Bahia, Brazil) experimental fields:

- time of the disease incubation phase in the plant: 180 to 540 days;
- duration of the latency phase of the disease in the plant: 30 or 60 days;
- proportion of insects per plant: 0.41 to 5;

- simulation time: 1, 2, 5, 10 or 20 years;
- simulation mode: 1 (single) or 2 (multi);
- probability of primary infection (PIP), according to the incidence in the region: 0.01 (low), 0.15 (medium) or 0.30 (high);
- probability of detection of the disease in the field by the human inspection: 0 or 0.476.

The simulations start with all healthy plants and the arrival of a certain proportion of infective insects, according to PIP values. Populations evolve stochastically over time (according to the probability of occurrence) from processes such as birth and death of nymphs and adult insects, infection of host plants, acquisition of bacteria by nymphs and adult insects, reproduction, and flight of adult insects.

At the end of the computer simulations, two types of results are generated: a) "single" simulation type: at every 10 days of the simulation execution, a file with the status of populations in each position of the planting area is generated (type of host, infection status, number of insects in position); b) "multi" simulation type: at the end of 100 automatic executions (Monte Carlo process), graphs of the number of susceptible, infected and symptomatic plants are generated over time.

The MBI results are saved, and a software developed in Java language allows the visualization of the model results via Web. This is illustrated by the examples shown in figures 4, 5, and 6, related to the 3 simulation scenarios that represent different landscapes (different configurations and proportions of citrus and myrtle) for a "single" simulations type.

Figures 4, 5, and 6 show the dynamics of HLB spread, after a certain number of days from the start of the simulation, which occurs from the arrival of insects in random plants of the first two left columns of the

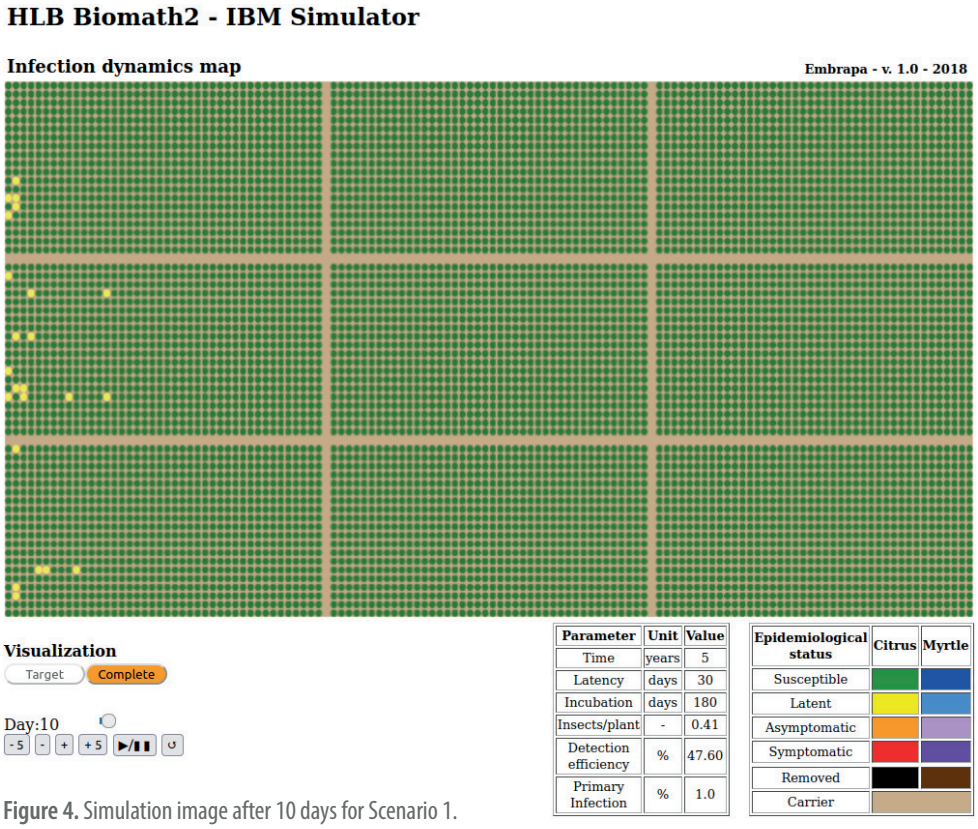

Figure 4. Simulation image after 10 days for Scenario 1.

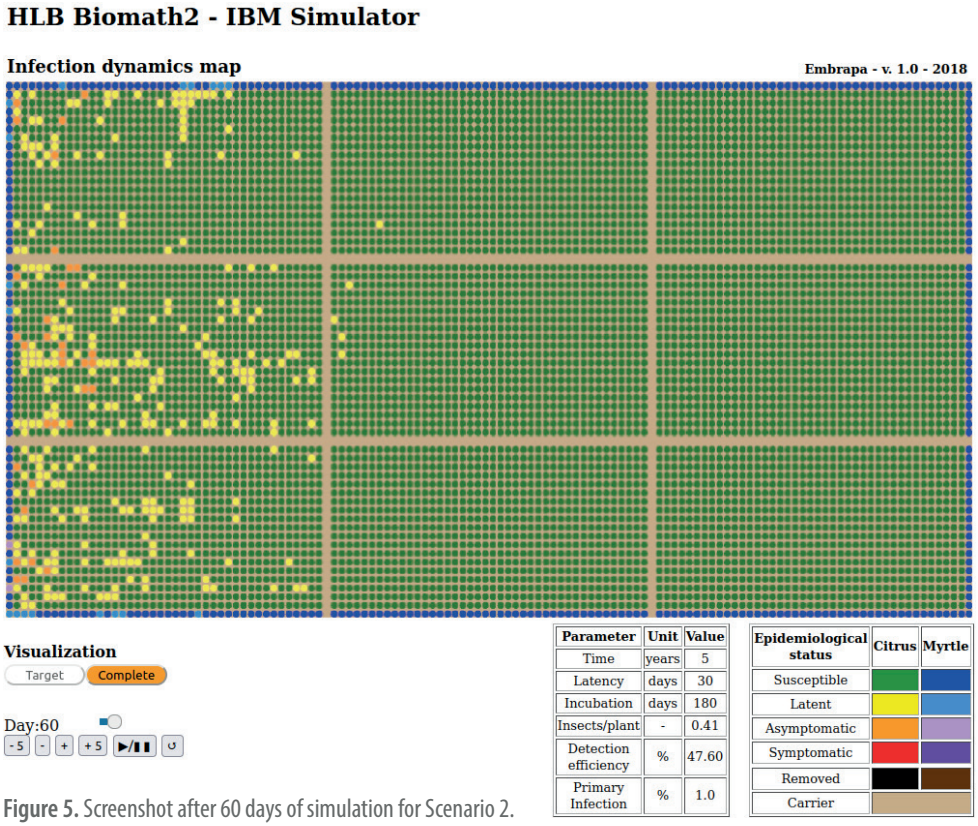

Figure 5. Screenshot after 60 days of simulation for Scenario 2.

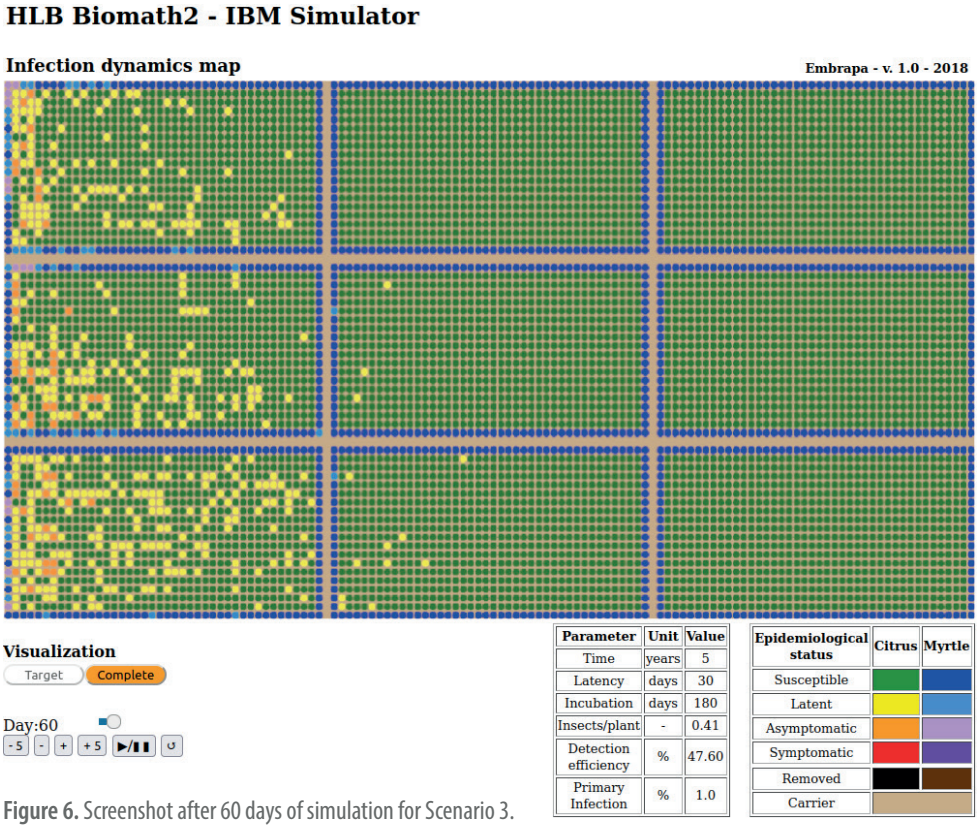

Figure 6. Screenshot after 60 days of simulation for Scenario 3.

fields. The amount of infective insects arriving in this area depends on the proportion of insects per plant and the PIP value chosen by the user. For example, for the proportion of 0.41 insects per plant, there are 1,033 insects at the beginning of the simulation, of which: a) for PIP = 0.1%: 1 infective insect; b) for PIP = 1%: 10 infective insects; c) for PIP = 15%: 154 infective insects.

Scenarios 1, 2, and 3 were tested separately in numerous combinations of the aforementioned parameters for the repulsion and attraction analysis. No visual differences were found in graphics generated by the "multi" execution, or the dynamics observed in the "single" executions. Following the analyses, scenarios were compared in a two by two scheme, and several simulations were performed for each scenario. Statistical tests were performed to compare the time of arrival of the disease in the target plot as well as in all comparisons between scenarios while considering different probabilities of primary infection. It was found that, statistically, there is no difference (comparisons between PIP equal to 1% and 15%) regarding the time of disease arrival in the target plot.

The results of the simulations prove observations made in field experiments: the primary infection has much more weight in the dynamics of disease propagation than the different spatial configurations of orange jessamine and citrus in the simulation scenarios .

Thus, the main conclusion obtained is that the simple presence of the alternative host (orange jessamine) does not significantly influence the epidemic process. This leads us to question how the interaction of the "HLB-insect vector-citrus" system would be with the use of vector population control methods, such as the application of insecticides, which could significantly affect the primary infection.

At the same time, the search for a threshold value for primary infection leads us to estimate the effort of regional management in order to stabilize the epidemic process. Furthermore, vector infectivity levels can be an indicator to be used in the future for the effectiveness of control measures in regional management. This indicator can be obtained more easily than extensive surveys with infected plants.

Currently (Barbosa, 2019) the MBI is evolving by the inclusion of new alternative hosts to evaluate in repulsion and attraction configurations, as well as testing periodic insecticide control strategies which minimize the effect of primary infection on landscapes.

From the spatiotemporal dynamics observed in the citrus HLB represented in the model, it is possible to simulate complex dissemination scenarios and perform the selection of more promising repulsion and attraction configurations to control the spread of the vector insect. This may be tested in future experiments along with obtaining indicators of effectiveness, with potential for more detailed studies in other projects.

Genetic evaluation of livestock

Animal breeding programs aim to genetically improve the population in terms of economic characteristics demanded by the market, adopting appropriate indices for the production system. In short, they consider the identification and genetic discrimination of individuals in the population, the selection of those with superior traits for replacement, either male or female, and the mating between them. An integral part of these programs are the genetic evaluation processes, which consist of continuously and cumulatively collecting biometric and genealogical data from the population undergoing improvement and periodically using a genetic-statistical model to predict the genetic values of each animal. The data include observation on the expression of physical or behavioral attributes of interest to the market. These attributes are called phenotypes and pedigree data, which in other terms means the relationships that define the genealogy of the population.

Currently, the methodology used in genetic evaluations of animals is based on the theory of mixed models (Henderson, 1963), known as BLUP (Best Linear Unbiased Prediction). It basically consists of the prediction of genetic values, adjusting the data simultaneously for fixed effects and an unequal number of observations per class (Lopes, 2005). Among the advantages of a genetic evaluation using BLUP are the inclusion of complete family information through a kinship matrix; comparison of individuals with different levels of fixed effects; and simultaneous evaluation of sires, females, and progenies. Lastly, there is the evaluation of individuals without observations, missed observations and with observations in only some characteristics (Lopes, 2005). BRBLUP (Higa, 2020) is a software for genetic evaluation of animals developed by Embrapa, based on the Python programming language and associated scientific computing libraries called Scipy/Numpy and PyTables. It supports the specification of mixed model equations so that different genetic-statistical models can be specified, including the multivariate animal model (MAM), which simultaneously evaluates fixed and random effects for a set of quantitative phenotypes while taking correlations between phenotypes or random effects into account, such as genetic origin effects.

As an example to illustrate the use of BRBLUP, (Example 5.4 of Mrode (2014) an animal model with two phenotypes (bivariate) is considered: a) FAT1: fat yield in lactation period 1; b) FAT2: fat yield in the lactation period 2. Associated with each phenotype is the presence of a fixed effect referring to herd-year-season (HYS1 and HYS2). The data set is shown in Table 1: there are eight animals, numbered from 0 to 7, and only those that have observed phenotypes (animals 0, 1 and 2) appear in the pedigree (columns Sire and Dam). The residual variances are 65 for the FAT1 phenotype and 70 for the FAT2 phenotype, with the covariance between them equal to 27; the genetic variances are 35 for the FAT1 phenotype and 30 for the FAT2 phenotype, with the covariance being equal to 28.

Table 1. Dataset (columns 0, 1, 2: pedigree – columns 0, 3, 4, 5, 6: observed data).

Animal	Father	Mother	HYS1	HYS2	FAT1	FAT2
3	0	1	0	0	201	280
4	2	1	0	1	150	200
5	0	4	1	0	160	190
6	2	3	0	0	180	250
7	0	6	1	1	285	300

Source: Adapted from Mrode (2014).

To solve the model, the BRBLUP software is executed through a command line, passing a configuration file as a parameter with the model specification. The result is stored in an output file.

Table 2 presents the contents of the generated output file. It contains 4 columns (Trait: column in the data file corresponding to a phenotype; Effect: specified effect in the model; Level: level of the effect in the data file; Sol: obtained solution). In this example, the first line of the file means that the solution for level 0 of effect 1 (HYS1) for the phenotype in column 3 is equal to 175.73126996362862. The seventh line means that level 1 (animal 1) for effect 0 (genetic value) for the phenotype in line 3 is equal to -2.999142788478058.

Accuracy values, which represent the reliability of the solution obtained for the genetic value, were not presented in this example, but are always used together. Finally, another aspect not addressed refers to the fact that, currently, animal genetic improvement programs are making efforts to include genomic information in the genetic evaluation process. This has direct implications for the construction and resolution of the genetic-statistical model used.

Sustainable Pantanal Farm

In recent decades, given the globalization of the economy and the creation of competitive markets, pressures to increase the productivity of farms in the Pantanal have intensified, compromising the sustainability of their production systems due to the fragility of its ecosystems. Given this scenario, a multidisciplinary group of researchers from Embrapa Pantanal, using previous experience on the characterization of Pantanal farms (Santos et al., 2017), developed a project in partnership with Embrapa Digital Agriculture. They aim to develop a tool to assess the sustainability of beef cattle production systems in complex and dynamic regions, such as the Pantanal, so that it would be possible to verify the system's weaknesses in order to seek good sustainability management practices.

Table 2. Result of genetic evaluation.

Trait	Effect	Level	Solution
3	1	0	175.73126996362862
3	1	1	219.61329398893875
4	2	0	243.23908674216108
4	2	1	240.54972646633607
3	0	0	8.969159144237393
4	0	0	8.840288629082728
3	0	1	-2.999142788478058
4	0	1	-2.7772802747175986
3	0	2	-5.970016355758499
4	0	2	-6.063008354365654
3	0	3	11.75424243135119
4	0	3	11.657587566164255
3	0	4	-16.252956614066754
4	0	4	-15.823507978243187
3	0	5	-17.31429689333114
4	0	5	-15.719126003080525
3	0	6	8.690473723985185
4	0	6	8.137644915235219
3	0	7	22.702139483291525
4	0	7	20.930688340763133

The Pantanal biome is located in the Midwestern region of Brazil (80%), also covering part of Bolivia and Paraguay. It constitutes an extensive neotropical wetland that is seasonally flooded, with a temporal and spatial variability of diversity, which is controlled by the flood pulse. This makes the region a complex, dynamic and uncertain system (Santos et al., 2017). Because it has extensive areas of natural grasslands with a predominance of forage, the Pantanal has a vocation for the extensive beef cattle ranching with low use of external inputs which has contributed to its conservation for more than two centuries. This has been the main economic activity on Pantanal farms, making it an important socioeconomic sector at the regional and national level. Considering that farms comprise about 95% of the Pantanal plain, the main challenge for decision makers is to define beef cattle production systems that do not cause major environmental impacts while bringing economic and social benefits to the local population and ensuring the conservation and sustainable use of natural resources.

In order to understand Pantanal farms holistically, aspects and indicators were defined in a hierarchical manner at both ranch and regional level and assess the beef cattle production system (Figure 7). These aspects and indicators were selected due to their practicality in representing and simplifying complex and systemic phenomena. Some of

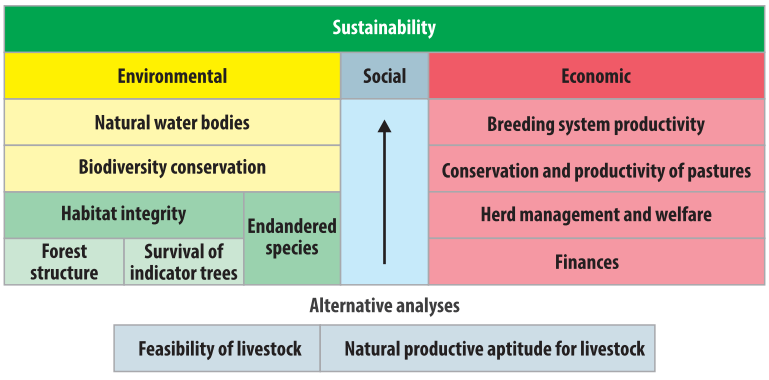

Figure 7. Hierarchical structure of the Sustainable Pantanal Farm.

the indicators were based on scientific studies carried out by the multidisciplinary team, while others were determined through several participatory workshops involving decision makers to validate the indicators. Some of these indicators must be evaluated directly in the field, while others can be studied through image analysis and mathematical calculations, or defined within the inference system adopted. To guide the field assessment and the collection of information necessary for the calculations, several protocols were developed and published (Soares et al., 2014; Santos et al., 2014a, 2014b, 2015; Abreu et al., 2015; Amâncio et al., 2016). This hierarchical process (Figure 7) enables assessing each aspect of sustainability individually and simultaneously.

The Sustainable Pantanal Farm software

Some problems arise in sustainability assessment, and it is necessary to take the level of abstraction involved in the concept into account, as well as the existence of natural variability in some phenomena. The synthesis provided by the indicators for a given "degree of sustainability" requires a robust methodology to deal with uncertainties, express complex interrelations, while at the same time, being interpretable and transparent to guarantee confidence in the assessment.

A mathematical and computational framework capable of dealing with these difficulties come from fuzzy set theory (FS), fuzzy logic, and fuzzy rule-based systems – FRBS. Such systems have been applied in areas such as engineering, modeling, and control, among others. Historically, its success is due to the ability to model knowledge based on natural language and good generalization capacity as well as the remarkable competence of FRBS in explaining the elaboration of the result based on the input values provided.

The Sustainable Pantanal Farm software was built as a decision support system based on models expressed in FRBS. Sustainability is evaluated by the environmental, economic, and social dimensions, at both ranch and regional level. Models were defined for each assessment (Figure 7), while input variables were the indicators themselves, with their scales defined in natural language (such as Good, Moderate and Bad). The relationships between indicators are expressed as a set of rules defined by domain experts. The evaluation results (indices and sub-indices), in addition to providing a comparative numerical value (1 to 10), have a corresponding qualitative output. Each model (index) feeds the more general models hierarchically further down, culminating in the farm's sustainability model.

The Sustainable Pantanal Farm interface to Internet[2] (Figure 8) is an interactive system where the user, given the indicator values, is able to infer qualitative concepts and numerical values, as well as compare how good these values are in relation to what is desired. It is also possible, through graphics, to visualize which indicators had more influence on the result. The rules that were used for the conclusion are shown to the user, ensuring interpretability and transparency. The system also allows the user to simulate scenarios in order to plan which ones lead to the level of sustainability one wants. The Sustainable Pantanal Farm software also has a second interface, aimed at mobile devices such as tablets and smartphones (Figure 9) using the Android operating system (available on the Google Play app store). Essentially, it provides the same functionalities, and it is based on the same mathematical models. Given a regional restriction of the Pantanal and farms in general, this version does not need an internet connection, as it has its own inference engine built into the application.

The Sustainable Pantanal Farm tool can be adopted by several decision makers (researchers, owners, technicians, politicians, legislators, certifiers, among others). Its main use is the diagnosis (degree of sustainability) of the beef cattle production system in the Pantanal through the assessment of environmental, social, and economic impacts of this activity, thus assisting in efficient management

[2] [2] Available at: https://www.fps.cnptia.embrapa.br

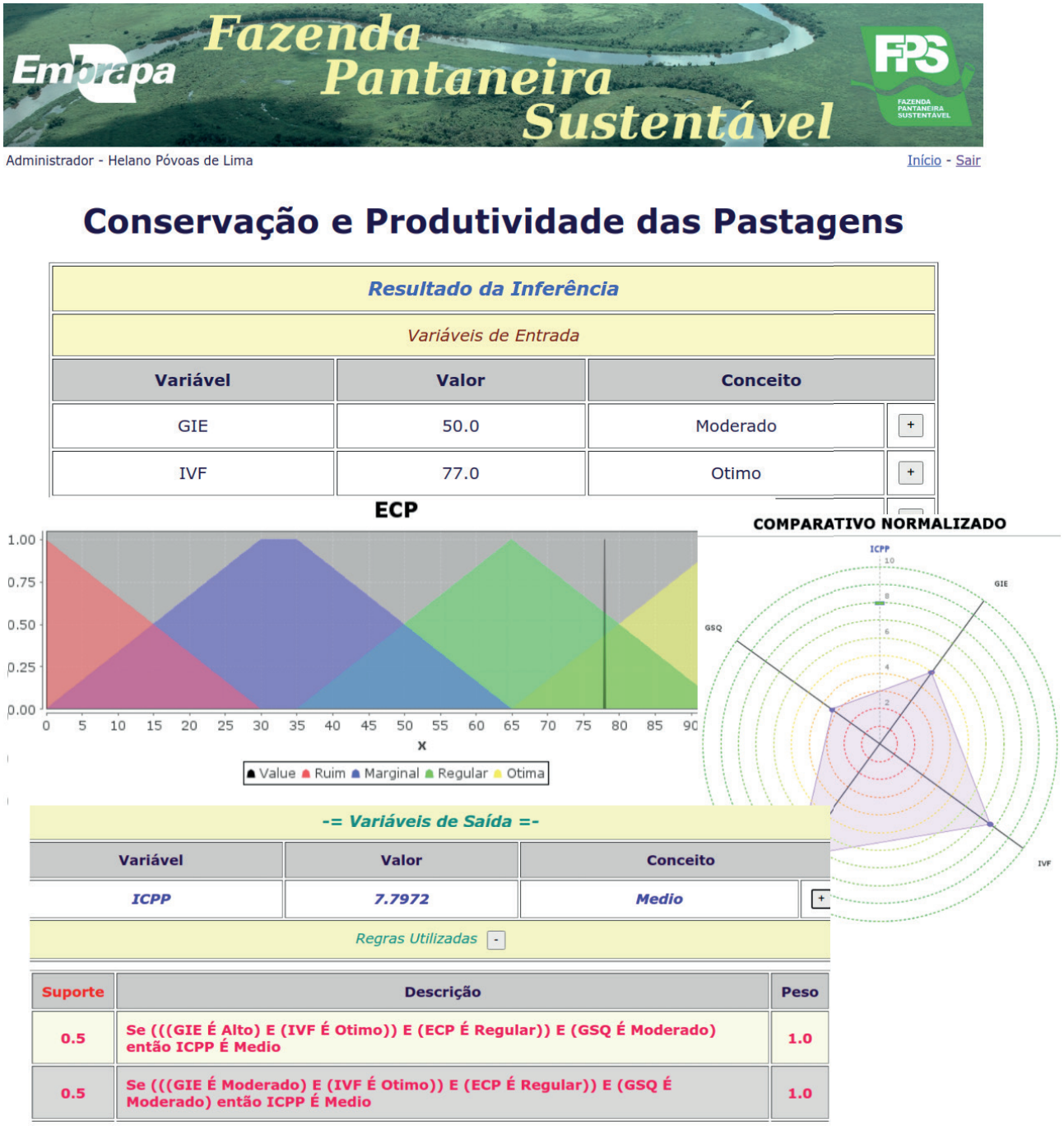

Figure 8. Internet Sustainable Pantanal Farm software interface elements.

Source: Sustainable Pantanal Ranch (2020).

through technology selection and good management practices. However, its application can be much broader for financing subsidy programs, certification, and marketing strategies that value products from the region. It may also offer necessary subsidies for the reformulation of current legislation and public incentive policies for sustainable production in the region. It is intended to insert the aspect of multifunctionality and ecosystem services in the future, something essential for the sustainability of production systems. The tool is being implemented in 15 farms in the Mato Grosso Pantanal with support from other agribusiness regional institutions, such as FAMATO, ACRIMAT, SENAR, IMEA, and rural unions,

Figure 9. Sustainable Pantanal Farm Android app interface.

Available at: https://play. google.com/store/apps/details?id=br.embrapa.cnptia.fps

as well as in six farms in the Pantanal of Mato Grosso do Sul, with support from FAMASUL, SENAR, and rural unions. Improvements will be incorporated over time, together with technicians, producers, and researchers.

Final considerations

In this chapter, several scientific-computing techniques applied in solving problems in the agricultural sector were presented. In the area of artificial Intelligence, classical logic techniques were applied to the development of an expert system for soil classification. The same problem was also addressed by a completely different technique using machine learning algorithms, which are fundamentally linked to statistics. Statistical analysis is also the basis of text mining techniques used to group documents with similar content in the agricultural area.

Another area of scientific computing, mathematical modeling, was explored in three different ways. In the first, Individual-Based Model provided a fully computational tool through a simulation system to compare three citrus and orange jessamine planting configurations in order to evaluate propagation control strategies for HLB in citrus. In the second application, linear predictor models, composed of classical mathematical equations, were used to assess the genetic values of livestock, with the objective of discovering which of them reinforce characteristics desired in the market. In the third model, the

mathematical calculations were internally performed in a fuzzy logic inference-based system in order to assess sustainability in Pantanal farms. In this case, the advantage of fuzzy logic is combining natural language in the construction of a logical model where the answer is explainable to the decision maker.

Scientific computing techniques are essential for analyzing the large volume of data produced in this process of agricultural digital transformation. Through these techniques, it will be possible, from the collected data, to extract information and knowledge that will assist in the decision-making process in all links of the production chains, becoming central in the development of new agricultural solutions and technologies in Digital Agriculture. The applications presented in this chapter illustrate the variety of problems that can be addressed by the scientific computing methodological framework, including mathematical and statistical modeling, classical and fuzzy logic systems, simulation models, and machine learning models.

Considering these applications, it is worth emphasizing that the constant growth in data availability, technological advances and the expansion of the dimension and complexity of the demands of Brazilian society pose enormous challenges and opportunities for research and development in scientific computing applied to agriculture.

References

ABREU, U. G. P.; LIMA, H. P.; SANTOS, S. A.; MASSRUHÁ, S. **Protocolo**: Índice Financeiro (IF) para a Fazenda Pantaneira Sustentável (FPS). Corumbá: Embrapa Pantanal, 2015. 12 p. (Embrapa Pantanal. Documentos, 134).

AMÂNCIO, C. O. G.; ARAÚJO, M. T. B.; SANTOS, S. A.; NARCISO, M.; OLIVEIRA, M. D. **Protocolo**: Índice de Bem-Estar Social (IBS) para a Fazenda Pantaneira Sustentável (FPS). Corumbá: Embrapa Pantanal, 2016. 16 p. (Embrapa Pantanal. Documentos 139).

BARBOSA, F. F. L. **HLB BioMath fase 2**: abordagem biomatemática como suporte a defesa fitossanitária e avaliação ex-ante de tecnologias de manejo. Cruz das Almas: Centro Nacional de Pesquisa de Mandioca e Fruticultura, 2015. 26 p. (Embrapa. Macroprograma 2 - Código SEG 02.13.03.007.00.000).

BARBOSA, F. F. L. **HLB BioMath fase** 3: biomatemática aplicada à otimização de tecnologias de interposição de barreiras, modificação microambiental e exclusão para manejo do huanglongbing dos citros. Cruz das Almas: Embrapa Mandioca e Fruticultura, 2019. 40 p. (Embrapa. Tipo II - Código SEG 20.18.03.044.00.00).

EMBRAPA. **Bases de Dados da Pesquisa Agropecuária**. Available at: https://www.bdpa.cnptia.embrapa.br. Accessed on: 19 May 2020.

FUNDECITRUS. **Sumário executivo**: estimativa da safra de laranja 2020/21 do cinturão citrícola de São Paulo e Triângulo/ Sudoeste Mineiro: cenário em maio de 2019. Available at: https://www.fundecitrus.com.br/pdf/pes_relatorios/2020_05_11_ Sumario-Executivo- da-Estimativa-da-Safra-de-Laranja-2020-2021.pdf. Accessed: 19 May 2020.

GRIMM, V.; RAILSBACK, S. F. **Individual-based modeling and ecology**. Princeton: Princeton University Press, 2005. DOI: 10.1515/9781400850624.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157-1182, 2003.

HENDERSON, C. R. Selection index and expected genetic advance. **Statistical Genetics and Plant Breeding**, v. 982, p. 141-163, 1963.

HIGA, R. H. **Tutorial**: introdução ao software brblup. Campinas: Embrapa Agricultural Informatics, 2020. (Embrapa Agricultural Informatics. Documentos, 168). 2020.

LOPES, P. S. **Teoria do melhoramento animal**. Belo Horizonte: FEPMVZ, 2005. 118 p.

MOURA, M. F.; TAKEMURA, C. M.; SILVA, I. L. C.; TÁPIAS, L. M.; OLIVEIRA, C. T. de; BASSOI, L. H.; OLIVEIRA, S. R. de M. Metodologia para a construção de portfólios tecnológicos agrícolas a partir de publicações técnico-científicas. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 11., 2017, Campinas. **Ciência de dados na era da agricultura digital**: anais. Campinas: Ed. Unicamp: Embrapa Agricultural Informatics, 2017. p. 537-546. SBIAgro 2017.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3rd ed. CABI, 2014. DOI: 10.1079/9781780643915.0000.

PEREIRA, R. G.; MOURA, M. F. I-Preproc: uma ferramenta para pré-processamento e indexação incremental de documentos. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA AGRICULTURAL INFORMATICS, 11., 2015, Campinas. **Resumos expandidos**. Brasília, DF: Embrapa, 2015. p. 17-23.

POLIDORO, J. C.; MENDONÇA-SANTOS, M. de L.; LUMBRERAS, J. F.; COELHO, M. R.; CARVALHO FILHO, A. de; MOTTA, P. E. F. da; CARVALHO JUNIOR, W. de; ARAUJO FILHO, J. C. de; CURCIO, G. R.; CORREIA, J. R.; MARTINS, E. de S.; SPERA, S. T.; OLIVEIRA, S. R. de M.; BOLFE, E. L.; MANZATTO, C. V.; TOSTO, S. G.; VENTURIERI, A.; SA, I. B.; OLIVEIRA, V. A. de; SHINZATO, E.; ANJOS, L. H. C. dos; VALLADARES, G. S.; RIBEIRO, J. L.; MEDEIROS, P. S. C. de; MOREIRA, F. M. de S.; SILVA, L. S. L.; SEQUINATTO, L.; AGLIO, M. L. D.; DART, R. de O. **Programa Nacional de Solos do Brasil (PronaSolos)**. Rio de Janeiro: Embrapa Solos, 2016. 53 p. (Embrapa Solos. Documentos, 183).

SANTOS, H. G. dos; JACOMINE, P. K. T.; ANJOS, L. H. C. dos; OLIVEIRA, V. A. de; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A. de; CUNHA, T. J. F.; OLIVEIRA, J. B. de. **Sistema brasileiro de classificação de solos**. 3. ed. rev. ampl. Brasília, DF: Embrapa, 2013. 353 p.

SANTOS, S. A.; CARDOSO, E. L.; CRISPIM, S. M. A.; SORIANO, B. M. A.; GARCIA, J. B.; BERSELLII, C. **Protocolo**: Índice de Conservação e Produtividade das Pastagens (ICPP) para a Fazenda Pantaneira Sustentável (FPS). Corumbá: Embrapa Pantanal, 2014a. 18 p. (Embrapa Pantanal. Documentos, 130).

SANTOS, S. A.; LIMA, H. P. de; BALDIVIESO, H. P.; OLIVEIRA, L. O.; TOMÁS, W. M. GIS-fuzzy logic approach for building indices: regional feasibility and natural potential of ranching in tropical wetland. **Journal of Agricultural Informatics**, v. 5, n. 2, p. 26-33, 2014b. DOI: 10.17700/jai.2014.5.2.140.

SANTOS, S. A.; LIMA, H. P. de; MASSUHÁ, S. M. F. S.; ABREU, U. G. P. de; TOMÁS, W. M.; SALIS, S. M.; CARDOSO, E. L.; OLIVEIRA, M. D. de; SOARES, M. T. S.; SANTOS JR., A. dos; OLIVEIRA, L. O. F. de; CALHEIROS, D. F.; CRISPIM, S. M. A.; SORIANO, B. M. A.; AMÂNCIO, C. O. G.; NUNES, A. P.; PELLEGRIN, L. A. A fuzzy logic-based tool to assess beef cattle ranching sustainability in complex environmental systems. **Journal of Environmental Management**, v. 198, part 2, p. 95-106, Aug 2017. DOI: 10.1016/j.jenvman.2017.04.076.

SANTOS, S. A.; OLIVEIRA, L. O. F.; LIMA, H. P.; ABREU, U. G. P.; OLIVEIRA, M. D.; ARAÚJO, M. T. B. D. **Protocolo**: Índice de Manejo e Bem-Estar do Rebanho (IMBA) para a Fazenda Pantaneira Sustentável (FPS). Corumbá: Embrapa Pantanal, 2015. 20 p. (Embrapa Pantanal. Documentos, 135).

SOARES, M. T. S.; OLIVEIRA, M. D.; CALHEIROS, D. F.; SANTOS, S. A.; LIMA, H. P. **Protocolo**: Índice de Conservação de Corpos de Água Naturais (ICA) para a Fazenda Pantaneira Sustentável (FPS). Corumbá: Embrapa Pantanal, 2014. 22 p. (Embrapa Pantanal. Documentos, 128).

SOUZA, K. X. S.; TERNES, S.; OLIVEIRA, S. R. M.; MOURA, M. F.; BARIONI, L. G.; HIGA, R. H.; FASIABEN, M. C. R. A perspective study on the application of Data Science in agriculture. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 11., 2017, Campinas. **Ciência de dados na era da agricultura digital**: anais. Campinas: Ed. Unicamp: Embrapa Agricultural Informatics, 2017. p. 537-546. SBIAgro 2017.

SUSTAINABLE PANTANEIRA FARM. Available at: https://www. fps.cnptia.embrapa.br. Accessed on: 19 May 2020.

TAKEMURA, C. M.; MOURA, M. F.; MACHADO, L. S. C. **TopExtract – toponym extraction and disambiguation tool**: componente de software para extração e desambiguação de topônimos. Campinas: Embrapa Monitoramento por Satélite, 2013. 1 CD-ROM.

VAZ, G. J.; SILVA NETO, L. de F. da; LIMA, R. N.; MARQUES, F. A. M; SANTOS, J. C. P. dos; OLIVEIRA, S. R. de M. Curadoria de dados de solos brasileiros por meio de um sistema especialista de classificação de solos. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 37., 2019a, Cuiabá. **Intensificação sustentável em sistemas de produção**: resumos. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2019.

VAZ, G. J.; SILVA NETO, L. de F. da; LIMA, R. N.; OLIVEIRA, S. R. de M. Uma API para a classificação de solos do Brasil. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 12., 2019, Jaguariúna. **IoT na Agricultura**: anais. Campinas: Embrapa Agricultural Informatics, 2019b.

VAZ, G. J.; SILVA NETO, L. de F. da; OLIVEIRA, S. R. de M.; BOTELHO, F. P.; ARAUJO FILHO, J. C. de. Development of an expert system for classification of Brazilian soil profiles. In: WORLD CONGRESS OF SOIL SCIENCE, 21., 2018, Rio de Janeiro. **Soil science**: beyond food and fuel: abstracts. Viçosa, MG: SBCS, 2018. Não paginado. WCSS 2018. Also published in: WORLD CONGRESS OF SOIL SCIENCE, 21., 2018, Rio de Janeiro. Soil science beyond food and fuel: proceedings. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2019. v. 1, p. 56-57.

VILAMIU, R. G. d'A.; TERNES, S.; LARANJEIRA, F. F.; SANTOS, T. S. Modelling the effect of an alternative host population on the spread of citrus Huanglongbing. **AIP Conference Proceedings**, v. 1558, p. 2504-2508, 2013.