DIEGO FERNANDES TERRA MACHADO

ACCURACY AND UNCERTAINTY OF DIGITAL SOIL MAPPING APPROACHES TO EXTRACT AND TRANSFER SOIL INFORMATION FROM REFERENCE AREA

LAVRAS – MG

2017

DIEGO FERNANDES TERRA MACHADO

ACCURACY AND UNCERTAINTY OF DIGITAL SOIL MAPPING APPROACHES TO EXTRACT AND TRANSFER SOIL INFORMATION FROM REFERENCE AREA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência do Solo, área de concentração em Recursos Ambientais e Uso da Terra, para obtenção do título de Mestre.

Profa. Dra. Michele Duarte de Menezes
Orientadora

Prof. Dr. Nilton Curi
Coorientador

LAVRAS - MG
2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Machado, Diego Fernandes Terra.
Accuracy and uncertainty of digital soil mapping approaches to extract and transfer soil information from reference area / Diego Fernandes Terra Machado. - 2017.
114 p. : il.

Orientador(a): Michele Duarte de Menezes.
Coorientador(a): Nilton Curi .
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2017.

Bibliografia.
1. Digital Soil Mapping. 2. Knowledge-driven. 3. Data-driven. I. de Menezes, Michele Duarte. II. , Nilton Curi. III. Título.

DIEGO FERNANDES TERRA MACHADO

ACCURACY AND UNCERTAINTY OF DIGITAL SOIL MAPPING APPROACHES TO EXTRACT AND TRANSFER SOIL INFORMATION FROM REFERENCE AREA

(ACURÁCIA E INCERTEZA EM ABORDAGENS DE MAPEAMENTO DIGITAL DE SOLOS PARA EXTRAIR E TRANSFERIR INFORMAÇÕES PEDOLÓGICAS A PARTIR DE ÁREA DE REFERENCIA)

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência do Solo, área de concentração em Recursos Ambientais e Uso da Terra, para obtenção do título de Mestre.

APROVADA em 29 de setembro de 2017

Dra. Michele Duarte de Menezes	UFLA
Dr. Sérgio Henrique Godinho Silva	UFLA
Dr. Marcos Bacis Ceddia	UFRRJ

Dra. Michele Duarte de Menezes
Orientadora

Prof. Dr. Nilton Curi
Coorientador

LAVRAS – MG
2017

À memória de minha amada vozinha Olimpia Oliveira Terra, de meu tio Joaquim e de minha amiga Juliana Mara de Oliveira.

Dedico

AGRADECIMENTOS

Agradeço à minha mãe Iolanda Oliveira Terra pelo amor e suporte nesta minha jornada científica e ao meu Pai, Roberto Fernandes Machado.

Aos meus tios e tias, por todo afeto. Vocês são para mim um porto seguro, um lugar para onde voltar. (Agradeço em especial aos meus tios Mizael, Fábio e Júlio e as tias Lia (Zetinha) e Isabel).

À Prof[a]. Michele Duarte de Menezes, por ter confiado em mim, pela paciência, por todo apoio e pelos ensinamentos nestes últimos dois anos.

Ao Prof. Sérgio Henrique Godinho Silva pelas contribuições ao longo da pesquisa.

À Yasmmin, pelo companheirismo de sempre, por ter encarado mais esse desafio comigo, com muito amor e amizade, como nos últimos 6 anos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de estudos; ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Aos Professores, técnicos e demais funcionários do DCS-UFLA.

Aos amigos da salinha, em especial Franciele, Rayner, Polyana e Juliana (*in memoriam*).

Aos meus amigos do DCS, Luis Renato, Mariana, Cristiane, Sara, Aline, Carol, Fernanda, Bruna, Ferreirinha, Maria Jéssica, Joana, Renata, Otávio, Diego, Jacqueline, Emerson, Rodrigo, e demais colegas.

MUITO OBRIGADO!

one hundred failures and one success is still success

ABSTRACT

Contradicting the need for detailed maps, we currently experience scarcity of investments on soil surveys in Brazil. In this sense, it is necessary to resort to techniques that allow the expansion of the mapped areas, at relatively lower costs. From this perspective, this work focused on the investigation of procedures and tools for the retrieving and extrapolation of soil type information from a reference area to its surroundings. The objectives included: (i) retrieving information from a detailed soil map of a reference area; (ii) to evaluate the transferability of information to a larger area, which preserves similar environmental characteristics similar to those of the pilot area; (iii) evaluate the accuracy and uncertainty of the inference models. From a Digital Elevation Model, a series of topographic indexes were calculated, which were correlated with the soil classes, represented by mapping units of the legacy map. The objective was to infer from the soil-landscape relationship of the pilot area, the distribution of soil types in the extrapolation area. For that duty, three inference procedures were applied, one data-driven (Random Forest (RF)) and two others, based on knowledge (Rule-based reasoning and Case-based reasoning - ArcSIE). Regarding RF, 52 models were graded from a routine of tuning and different combinations of training data. Although considered a robust predictor, RF demonstrated sensitivity to training strategies. Most of the models presented low accuracy. However, at least one model with more than 80% of global accuracy was obtained. Regarding RBR and CBR procedures, only the former resulted in a map with good precision. The advantage of using knowledge-based systems like RBR is to make explicit the soil-landscape relationship through a systematic set of rules. By accessing the uncertainty of the predictions, in addition to evaluating the behavior of the models, it was possible to observe the complexity of the soil-landscape relationship of Oxisols and Inceptisols, characteristic of tropical environments. This is particularly important for model review and sampling planning in the search for more accurate maps.

Keywords: Pedology. Digital soil maps. Soil surveys.

RESUMO

Contrapondo a necessidade de mapas mais detalhados, atualmente enfrentamos a escassez de recursos destinados ao levantamento de solos. Neste sentido, é preciso recorrer a técnicas que viabilizem a expansão das áreas mapeadas, a custos relativamente mais baixos. Sob essa perspectiva, este trabalho focou na investigação de procedimentos e ferramentas para a extrapolação de informações sobre classes de solo de uma área de referência para seu entorno. Os objetivos incluíram: (i) recuperar informações de um mapa de solos detalhado de uma área de referência; (ii) avaliar a transferibilidade da informação para uma área 15 vezes maior, que preserva características de paisagem semelhantes às da área piloto; (iii) avaliar a acurácia e a incerteza dos modelos de inferência. A partir de um Modelo Digital de Elevação, calculou-se uma série de índice topográficos, os quais foram correlacionados com as classes de solo, representadas por unidades de mapeamento do mapa legado. O objetivo, portanto, foi inferir, a partir da relação solo-paisagem da área piloto, a distribuição dos tipos de solo na área de extrapolação. Para tanto, foram aplicados três procedimentos de inferência, um baseado em dados (Random Forest (RF)) e outros dois baseados no conhecimento (Rule-based reasoning-RBR e Case-based reasoning-CBR - ArcSIE). Com relação a RF, foram grados 52 modelos a partir de uma rotina de ajustes e diferentes combinações de dados de treinamento. Embora considerado um preditor robusto, a RF demonstrou sensibilidade as estratégias de treinamento. Grande parte dos modelos apresentou baixa precisão, contudo, obteve-se ao menos um modelo com mais de 80% de acurácia global. Em relação aos procedimentos RBR e CBR, apenas o primeiro resultou em um mapa com boa precisão. A vantagem da utilização de sistemas baseados no conhecimento com o RBR é o de tornar explicita a relação solo-paisagem através de um conjunto sistematizado de regras. Ao acessar a incerteza das predições, além de avaliar o comportamento dos modelos, foi possível observar a complexidade da relação solo paisagem, característica de ambientes tropicais. Este aspecto é particularmente importante para revisão de modelos e planejamento de coletas de solo na busca por mapas com maior acurácia.

Palavras-chave: Pedologia. Mapas digitais de solos. Levantamento de solos.

SUMARRY

FIRST PART

DIGITAL SOIL MAPPING: INFERENCE MODELS AND UNCERTAINTY OF PREDICTIONS

1 INTRODUCTION

1.1 General introduction

Soil is essential for life, playing a role in providing food, fibers and fuel, serving as basis for human infrastructure and it is also related to other environmental such aspects as climate regulation, water security and biodiversity protection (FAO, 2015). Thus, knowing about the soil distribution allows a proper planning by defining the appropriate land use for each location.

Soil surveys are inventories of the morphological, physical, chemical and mineralogical characteristics of soils, as well as their geographic distribution, whose representations are given through maps and reports (DALMOLIN et al., 2004). The soil survey is based on understanding how the soils evolve for identifying their spatial distribution pattern. Hudson (1992) described it as a paradigm-based science. According to the author, by comprehending the soil forming factors (JENNY, 1941) and identifying the soil-landscape relationship (HUDSON, 1992), a soil scientist can accurately discriminate boundaries between different soil types.

Traditionally, soil mapping has been carried out in an empirical approach, based on tacit knowledge and mental models, with manual delineation of mapping unit boundaries. Although delivering very precious results, it is very onerous and time-consuming (KEMPEN et al., 2012), which are pointed out as important factors for the worldwide lack of soil spatial data information (MCBRATNEY; SANTOS; MINASNY, 2003). Also, it is argued that the cost / benefit ratio, which is poorly understood and difficult to estimate, also makes it difficult to raise funds (GIASSON; INDA-JÚNIOR; NASCIMENTO, 2006). Moreover, the strong reliance on tacit knowledge hinders the transfer of knowledge (HUDSON, 1992).

In Brazil, most soil maps were designed with low level of detail (LEPSH, 2013; SANTOS et al., 2013), in addition to being mostly in press (paper-based format), making their refinement more difficult (SILVA, 2016). The need for more detailed information contrasting with low investments in soil surveys resulted in a scenario that “we need to permit ourselves to consider alternative and possibly less costly approaches” (MALONE et al., 2016, p. 243).

An economical alternative to obtaining soil spatial information is the use of soil legacy data and Digital Soil Mapping (DSM) techniques. They allow retrieving information of already mapped areas and extrapolate the information for non-mapped areas. In this regard, the concepts of homosoil (MALLAVAN; MINASNY; MCBRATNEY et al., 2010), reference area (LAGACHERIE; LEGROS; BURROUGH, 1995), predictive soil mapping (SCULL et al.,

2003), and the *SCORPAN* model (MCBRATNEY; SANTOS; MINASNY, 2003) have an important role, serving as procedural parameters to create an information system and identify, through similarity assessment, sites which may be possible to extrapolate the information from a detailed area to another with sparse data (MALONE et al., 2016).

The increasing access to new techniques and tools for data acquisition and processing has significantly changed the way of soil information has been handled (ARROUAYS; LAGACHERIE; HARTEMINK, 2017). So, some alternatives have emerged to circumvent those limitations. In the last decades, DSM has been widely discussed and its feasibility has been reported by numerous researchers (BUI; MORAN, 2001; DOBOS et al., 2000; HENGLE et al., 2015; PANDARIAN; MINASNY; MCBRATNEY, 2017).

DSM can be described as the computer-assisted creation of spatial soil information systems by means of mathematical and statistical models combining field and laboratory observation, expert knowledge, and also correlated environmental features (DOBOS et al., 2006; LAGACHERIE; MCBRATNEY, 2007). Somehow, both traditional and digital approaches are similar on needing an input data on soil and covariates that describe the soil forming environment. It is important to highlight that the field work is still very needed and important. The main difference is related to how both models derive the spatial prediction from the input data (DOBOS et al., 2006).

DSM techniques can be grouped in two broad categories: statistical/geostatistical approaches (Data-driven) and knowledge based approaches (Knowledge-driven) (ASHTEKAR; OWENS, 2014; SHI et al., 2009). As a result, the digital maps represent estimates of soil types or properties spatial distribution. These estimations also involve different levels of uncertainty (STUMPF et al., 2017). Despite the great advances in DSM, some methodologies are unsettled as the evaluation of soil predictions uncertainty (ARROUAYS; LAGACHERIE; HARTEMINK, 2017).

1.2 Objectives

This study had three main objectives: (i) to retrieve information from a detailed scale soil map of a reference area; (ii) to evaluate the transferability of the information to a larger area with similar environmental characteristics using three different approaches (Random Forest; Rule-based Reasoning and Case-based Reasoning); and (iii) to assess the uncertainty of predictions. Thus, two scientific articles were presented:

a) Soil type spatial prediction from Random Forest: different training datasets, transferability, accuracy and uncertainty assessment. This study aimed to evaluate the use of Random Forest for extracting and extrapolating soil type information. For this duty, different training datasets (point and polygon derived data) and combinations of predictor variables were tested. A total of 52 models were evaluated by means of error of models itself, prediction uncertainty and external validation. It was also investigated the modeling behavior by reducing the amount of training data and the number of predictor variables to its main components by means of Principal Components Analysis, and Mean Decrease in Accuracy (for variables only);

b) Transferability, accuracy, and uncertainty assessment of different Knowledge-based approaches for soil type mapping. The main objective was to evaluate the efficiency of Rule-based and Case-based approaches on retrieving soil spatial information of a reference area and extrapolate it to surroundings with similar soil-environment characteristics. The methodology contains three main processes: i) knowledge acquisition; ii) soil inference procedures; iii) validation and uncertainty assessment.

2 REVIEW

2.1 Digital Soil Mapping basis

Digital Soil Mapping can be de defined as the creation, and population of spatial soil information systems by the use of field and laboratory observational methods, coupled with spatial and non-spatial soil inference models. Thus, DSM presents three main components: the input data (soil and environment information); the inference system (a set of techniques used to predict and populate the information system); and the output (a spatialized soil information system along with the uncertainty of prediction) (LAGACHERIE; MCBRATNEY, 2006).

2.1.1 Input data

The conception of predicting soil distribution over the landscape is to infer and spatialize data that we do not know based on measurements that we more-or-less know (MCBRATNEY et al., 2002). The inference is based on the soil-landscape relationships, and the measurements correspond to the input for the inference systems. The input data includes legacy soil data, soil maps and often, new samples, combined with information of related environmental features.

The expansion of access to Geographic Information Systems, collaborative programming communities (e.g. Software R), and free sources of environmental covariates as Digital Elevation Models (DEM), Remote Sensing data and climate databases have favored DSM research in the last decades (ARROUAYS; LAGACHERIE; HARTEMINK, 2017). Brevik et al. (2016) also related to the increasing of proximal sensing techniques in soil spatial models development.

Regarding soil information, in recent years, legacy data has become even more important, given its potential as input data for DSM. The term legacy data is applied to all information that was raised to characterize or mapping soil through traditional techniques of its time (OMUTO; NACHTERGAELE; ROJAS, 2013) serving as available knowledge about soils for a given area of interest. The major sources are soil maps and reports in a paper-based format (SILVA, 2016), but it is often found on shapefile format as point-data (profiles descriptions and prospections) and polygon-data (soil mapping units). As for the environment covariates, the effort in storing, coding and harmonizing and the access to soil databases have been contributing for DSM advances (ARROUAYS; LAGACHERIE; HARTEMINK, 2017).

2.1.1.1 Reference area

In locations where the soil-landscape relationships are known, in terms of rules or implicit in detailed soil maps, there is possible to extrapolate this information, from these reference area, for physiographically similar regions where these relationships are not yet known (ARRUDA et al., 2016).

The term, reference area, were presented by Favrot (1989), and it is related to a small natural region where the main soil classes and its soil-environment relationships are well known and stablished in terms of mapping rules. These approach assumes that, it is possible to delimit areas considered representative of a finite number of soil classes and its occurrence patterns on the landscape. Once these patterns are repeating and identifiable, "consequently, a purposely chosen reference area could be sufficient to identify all the soil classes of the larger area and to determine their spatial relations" (LAGACHERIE; LEGROS; BURROUGH et al., 1995, p.284).

> The first stage consists of a detailed survey in a small but representative area of a small natural region, called the reference area. It defines the main soil classes of the whole region and establishes their mapping rules. This first stage facilitates and accelerates the following stage in which new soil surveys are carried out in the same region. These surveys consist of identifying, at purposely chosen observation points, the soil classes previously defined during the reference area survey and then delineating their boundaries with the pre-established mapping rules. (LAGACHERIE; LEGROS; BURROUGH, 1995. p. 284)

The knowledge acquired by searching in a representative small natural region or inherited in terms of legacy data can be used to facilitate and accelerate the soil survey in other areas in the same natural region (ARRUDA et al., 2016). However, "this advantage was mostly observed when the further soil surveys were carried out by the surveyor of the reference area himself" (LAGACHERIE et al., 1995, p. 284). In fact, the use of the reference area is attached to tacit knowledge, nonetheless, the advances in data mining methods can provide solutions to assist in the knowledge retrievement (SILVA et al., 2016).

The use of the reference area associated with predictive methods for soil mapping can be found in studies such as Grinand et al. (2008) - classification trees; Yigini and Panagos (2014) - regression-kriging; Arruda et al. (2016) – artificial neural networks; Silva (2016) – Random Forest; McKay et al. (2010) – fuzzy logic/ArcSIE.

2.1.2 Inference models

The advances in computer science, powerful hardware and software that can handle with large datasets, in addition to the increase in number of spatial inference models, it is pointed out by numerous soil scientist as some of the main factors that made operational results in DSM (MINASNY; MCBRATNEY, 2016; MCBRATNEY; SANTOS; MINASNY, 2003; ARROUAYS; LAGACHERIE; HARTEMINK, 2017).

An inference model works on associating soil observations with their forming factors (environmental features) to understanding the spatial distribution of soils and coevolving landscapes to infer about soil type and properties by means of expert knowledge, mathematical and statistical functions (GRUNWALD, 2006; ARROUAYS; LAGACHERIE; HARTEMINK, 2017). These soil inference models are tools for environmental soil-landscape modeling (GRUNWALD, 2006), composed of a source, an organizer, and a predictor (MCBRATNEY et al., 2002).

The soil-landscape modeling is largely influenced by the attribute statistical type (Boolean, categorical, continuous…); the content of attributes or feature type (soil, relief, parent material, organisms, climate, age…); observations (total number of observations, density of observations, sampling design, sample support); and the geographic extent of observation (GRUNWALD, 2006). In this sense, the choice of the technique applied for the soil inference is directly related to the data set characteristics, in addition to the researcher knowledge about the available approaches.

Digital soil mapping techniques can be divided into two main types: Data-driven and Knowledge-driven (SHI et al., 2009). The Data-driven approaches are based on statistics, geostatistics, machine learning and data mining techniques. It is often recommended when there is availability of relatively large datasets and dense sampling schemes in combination with ancillary environment features information (MENEZES et al., 2013). Once Data-driven approaches are strongly data dependent, it becomes expensive and not practical when the data is not readily available (ASHTEKAR; OWENS, 2014). Hengl et al. (2007) identified at least four distinct groups of data-driven approaches for DSM: (I) Pure classification techniques, which is based on remote sensing images; (II) Pure regression techniques, when applying regression and data mining methods (e.g. Random Forest; Neural Networks; Support Vector Machines); (III) Pure geostatistical techniques stated on kriging methods; IV) and Hybrid statistical/geostatistical approaches. An overview of predictive soil mapping based on these approaches can be found in Scull et al. (2003), McBratney et al. (2000) and Grunwald (2006).

The Knowledge-driven approaches are based on tacit knowledge, and are composed of a Knowledge base (the expert knowledge about the soil variation), an environmental dataset (data of spatial environmental features) and an inference engine that combines the data and the knowledge base to infer logically valid conclusions about the soil (SKIDMORE et al., 1996). In this methods, the values derived from the environmental features analysis are incorporated into the scientist knowledge, and contrarily to the Data-driven techniques, the inferences are structured in the decisions made by the expert, constructing the rule base, not by pure statistics predictions (ASHTEKAR; OWENS, 2014; SCULL et al., 2003). A preexisting knowledge of the soil-landscape soil relationships is crucial, thus, in many cases, they also require extensive work on data preparation, e.g. to design classification rules and to adjust the final outputs (HENGL et al., 2007).

It is worth pointing out that both approaches, knowledge-driven and data-drive, are not mutually exclusive, in fact, in some applications, combining two ways would be ideal. According to Rossiter (2005), the soil variability can be divided into “regional factors”, which can be explained by knowledge-driven models, and its “residuals”, product of our inability to explain the phenomena, in addition to those inherently random events. “A mixed approach is to explain what can be explained by knowledge of soil-forming factors, and then see if the remaining unexplained variability has any (geo)statistical relation which can be used to improve the prediction” (ROSSITER, 2005, p.4).

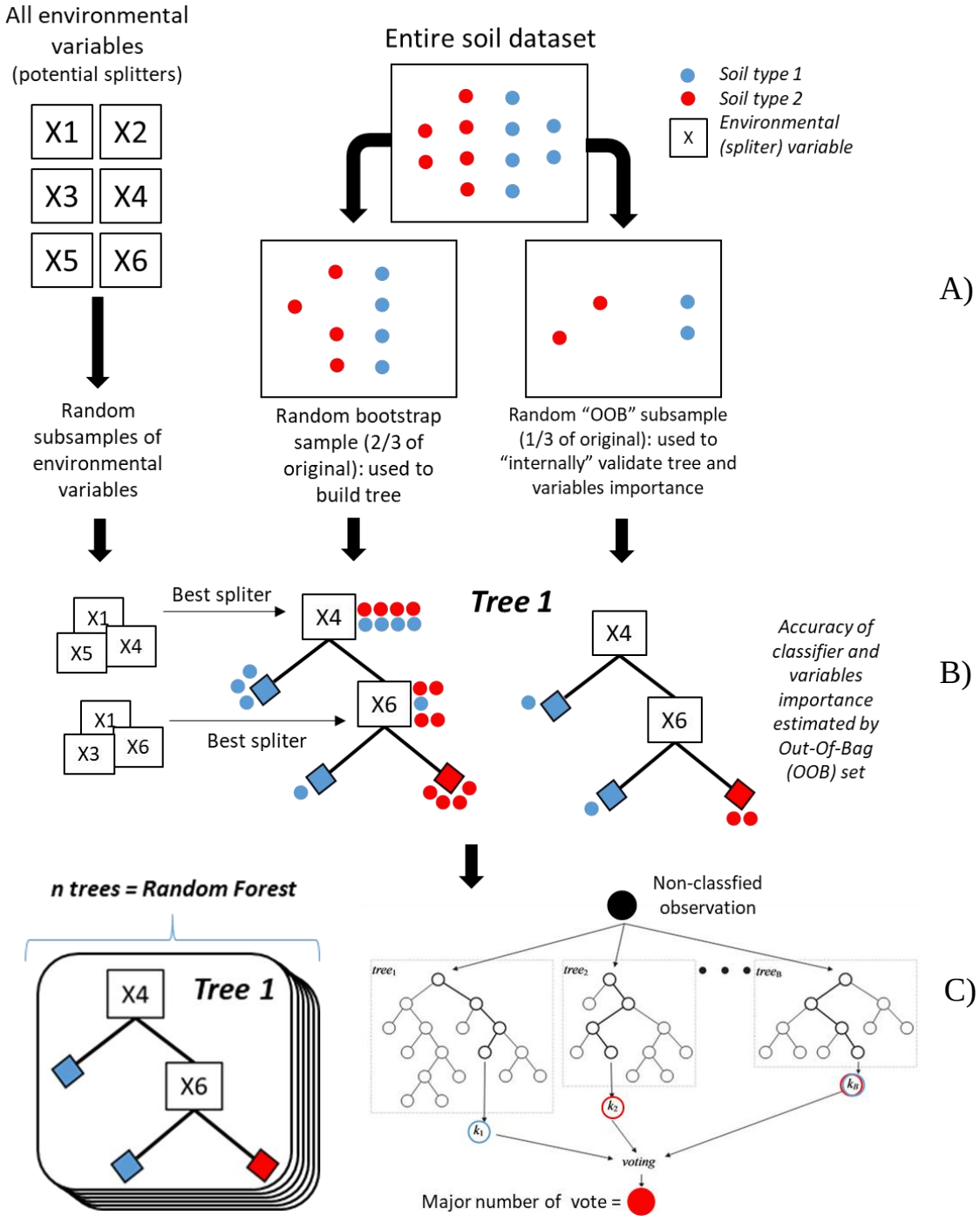
2.2 Random Forest

A Random Forest (RF) algorithm is an ensemble learning classifier, used to understand the relationship between a dependent variable and an ensemble of predictor variables. It consists of creating an independent collection of classification trees, from random vectors, sampled independently, and with the same distribution for all trees in the forest. Furthermore, at each node a random subset of variables is sorted, and the best among these is chosen as a “splitter variable”. The classification of a new input vector is a combination of the results reached by each tree in the forest. The output corresponds to the modal classification overall trees (BREIMAN, 2001) as seen at Figure 1.1. If the response is a factor, RF performs classification; if the response is continuous (that is, not a factor), RF performs regression.

The model demands an input training data. Each tree is grown based on a bootstrap sample. For each tree, about one-third of the observations are left out. This is called out-of-bag

data (OOB). This OOB data are used to get unbiased estimates of generalization error and variables importance (BREIMAN, 2001).

Figure 1.1 - The random forests algorithm is as follows: A) *ntree* bootstrap samples from the original data; B) for each of the bootstrap samples, grow an unpruned classification or regression tree; C) Predict new data by aggregating the predictions of the *ntree* trees.



Source: The author

There are basically two adjustments parameters that may be set to fine-tune the model for particular situations, which are; *mtry*, (the number of randomly selected predictor variables chosen at each node) and *ntrees* (the number of trees in the forest). "The only one of these parameters to which Random Forests is somewhat sensitive appears to be *mtry*" (CUTLER; CUTLER; STEVENS, 2012, p. 11).

In an extensive evaluation of different families of classifiers (179 classifiers arising from 17 families and 121 data sets), Fernández-Delgado et al. (2014, p. 1) related that "the classifiers most likely to be the bests are the random forest versions". RF is currently one of the most promising methods and its use is increasing in DSM.

On DSM literature and whose applied RF for classification/regression problems, others (BREIMAN, 2001; CUTLER; CUTLER; STEVENS, 2012; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MARMION et al., 2008; PRASAD; IVERSON; LIAW, 2006) some advantages and disadvantages of RF are often highlighted, which are:

a) advantages:

- insensitive to missing data (Missing value imputation);
- to the inclusion of irrelevant predictors and outliers;
- flexible with various types of datasets;
- less susceptible to over-fitting and it provides better error measurement in comparison with decision trees;
- incorporates randomness into its predictions (bootstrap sampling and randomized variable selection);
- naturally handle both regression and (multiclass) classification;
- are relatively fast to train and predict;
- depend only on one or two tuning parameters;
- have a built in estimate of generalization error; robust error estimates by using the OOB data;
- can be used directly for high-dimensional problems;
- can easily be implemented in parallel;
- measures of variable importance;
- differential class weighting;
- visualization;
- outlier detection;

- unsupervised learning;
- generates an internal unbiased estimate of the generalization error as the forest building progresses.

b) disadvantages or limitations:

- the stability of an RF classifier can be poor when it was used in an evaluation area that was different from the calibration area;
- difficult to provide guidelines for parameters adjustments to achieve good performance;
- for data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels;
- unlike decision trees, the classifications made by random forests are difficult for humans to interpret (black box). Results of learning are incomprehensible, compared to a single decision tree, or to a set of rules, they don't give you a lot of insight;
- it tends to return erratic predictions for observations out of range of training data.

2.3 Soil Inference Engine (ArcSIE) and Fuzzy Logic

The Soil Inference Engine (ArcSIE) is an expert knowledge-based inference engine powered with fuzzy logic. The inference method is based on fuzzy membership functions or similarity curves, which indicates the similarity between a local soil and a typical case. It is reasoned in partial membership concept. The similarity values vary from 0 (which means that soil is very different from the given soil type) to 1 (which means that local soil is exactly the same with the given soil type) (SHI et al., 2004), values in between this range express different degrees of similarity to the central concept (MENEZES et al., 2013). Fuzzy logic is particularly appropriate for soil distribution or representation due to the continuous and complex nature of soil-landscape relationships (SCULL et al., 2003).

There are two main types of knowledge supported by ArcSIE: rules and cases. For Rule-based reasoning approach, the membership functions are adjusted directly with the soil surveyor specifications. For Case-based reasoning approach, the membership functions correspond to a group of cases, defined in geographical space represented by polygons, lines or points. Case-

based reasoning aims to use the knowledge retrieved from specific cases to help solve a problem in a different area (SHI et al., 2004).

2.4 Uncertainty of predictions

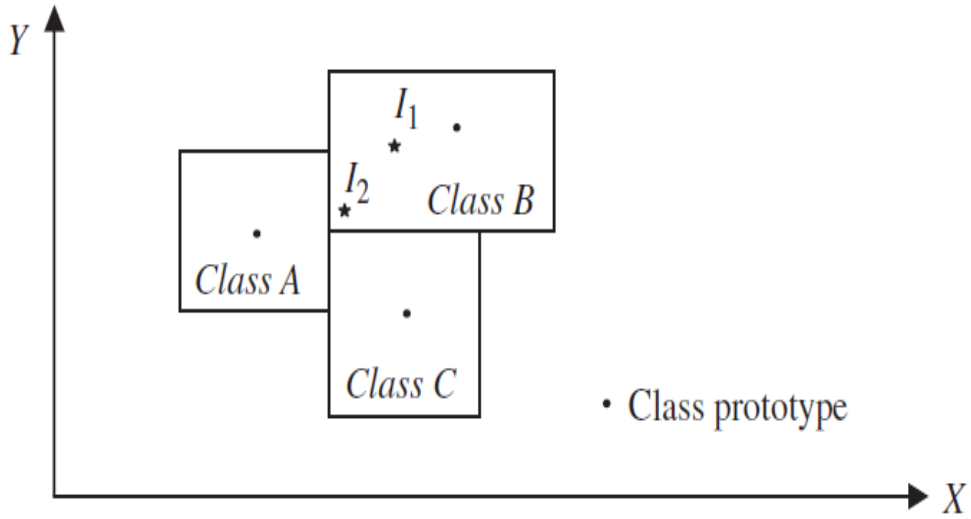
The uncertainty of predictions is strongly related to the knowledge domain about the study phenomena and the quality of the data base (HARROWER, 2003) also, the complexity of the area and soil variability may influence in this aspect. Qi and Zhu (2011) listed a series of causes for uncertainty related to predictive soil mapping, grouped in two kinds of errors: modeling error (generalization or over-simplification of transitional zones and non-avoidable inclusions) and mapping error (misplacement of class boundaries; mislabeling of polygons and avoidable inclusions). The uncertainty related to mapping errors are commonly modeled with stochastic methods, on the other hand, the uncertainty associated with modeling errors has been commonly studied using fuzzy logic (QI; ZHU, 2011).

For expert-fuzzy systems, like those used in soil inference engine approaches, the generalization of the continuous soil-landscape to discrete polygons map implies in a pixel-to-pixel conversion, in which a local object has its classification assigned due to its similarity to the soil class prototype (ZHU, 1997). In this sense, the uncertainty can be assessed in two different ways. The first one is the exaggeration uncertainty, which revels the “distance” for a given object classification in comparison with its prototype. For example, in Figure 1.2, by labeling as class B the objects I1 and I2, they assume the prototype properties by exaggerating the similarities between these instances and those of the class prototypes. It results in exaggeration uncertainty. The second one, the Ignorance uncertainty, also called Entropy, expresses the degree of certainty in a pixel classification, in which membership is concentrated in a particular class, rather than spread over a number of classes. In Figure 1.2, it is ignored that both objects l1 and l2 may present partial similarity to other prototypes (A and C) Qi and Zhu (2011).

Both entropy and exaggeration values do not indicate if a prediction itself is correct, however, they are appropriate for illustrating the uncertainty of predictions, to access the classifiers behavior, the quality of the modeling, and also as input in sampling planning for updating models (KEMPEN et al., 2009; STUMPF et al., 2017). Although the use of entropy for uncertainty assessment is most commonly used in fuzzy inference systems, its use in

ensemble-modeling like Random Forest has been shown to be appropriate (KEMPEN et al., 2009).

Figure 1.2 - Soil type distribution and assignments for objects l1 and l2.



Source: Qui and Zhu (2011).

REFERENCES

ARROUAYS, D.; LAGACHERIE, P.; HARTEMINK, A. E. Digital soil mapping across the globe. **Geoderma Regional,** [S.l.], v. 9, p.1-4, jun. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.geodrs.2017.03.002.

ARRUDA, G. P. de et al. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, Piracicaba, v. 73, n. 3, p.266-273, jun. 2016. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/0103-9016-2015-0131.

ASHTEKAR, J. M.; OWENS, P. R. Remembering Knowledge: An Expert Knowledge Based Approach to Digital Soil Mapping. **Soil Horizons,** Madison, v. 54, n. 5, p.1-6, 2013. Soil Science Society of America. http://dx.doi.org/10.2136/sh13-01-0007.

BREIMAN, L. Random forests. **Machine Learning**, [S.l.], v. 45, n. 1, p. 5–32, 2001.

BREVIK, E. C. et al. Soil mapping, classification, and pedologic modeling: History and future directions. **Geoderma,** Amsterdam, v. 264, p.256-274, fev. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2015.05.017.

BUI, E. N.; MORAN, C. J. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. **Geoderma**, Amsterdam, n. 103, p.79-94, fev. 2001.

CUTLER, A.; CUTLER, D.; STEVENS, J. R. Random Forests. In: ZHANG, C.; MA, Y. (Ed.). **Ensemble Machine Learning:** Methods and Applications. New York: Springer, 2012. cap. 5, p. 157-175. DOI: 10.1007/978-1-4419-9326-7_5

DALMOLIN, R.S.D.; KLANT, E.; PEDRON, F.A.; AZEVEDO, A.C. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. **Ciência Rural**, Santa Maria, v. 34, n. 5, p. 1479-1486, 2004.

DOBOS, E. et al (Ed.). **Digital Soil Mapping as a support to production of functional maps.** Luxemburg: Office For Official Publications Of The European Communities, 2006. 68 p.

DOBOS, E. et al. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. **Geoderma**, Amsterdam, v. 97, p. 367-391, 2000.

FAVROT, J. C. Une stratégie d'inventaire cartographique à grande échelle: la méthode des secteurs de référence. **Science Du Sol**, Montpellier, v. 27, n. 4, p.351-368, 1989.

Food and Agriculture Organization of the United Nations. 2015. Soil functions. 1 Infographic. < http://www.fao.org/3/a-ax374e.pdf> (accessed 19.09.17).

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? **J. Mach. Learn. Res.**, Brookline, v. 15, p. 3133–3181, 2014.

GIASSON, E.; INDA-JÚNIOR, A. V.; NASCIMENTO, P. C. Estimativa do benefício econômico potencial de dois levantamentos de solos no Estado do Rio Grande do Sul. **Ciência Rural**, Santa Maria, v. 36, n. 2, p.478-486, abr. 2006.

GRINAND, C. et al. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, Amsterdam, v. 143, n. 1-2, p.180-190, jan. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2007.11.004.

GRUNWALD, S. What do we really know about the space-time continuum of soil-landscapes? In: GRUNWALD, S. (Ed.). **Environmental Soil Landscape Modeling, Geographic Information Technologies and Pedometrics.** New York: Taylor and Francis, 2006. Cap. 1. p. 4-34.

HARROWER, M. Representing uncertainty: does it help people make better decisions? In: UCGIS WORKSHOP: GEOSPATIAL VISUALIZATION AND KNOWLEDGE DISCOVERY WORKSHOP, National Conference Center. **Proceedings…** Landsdowne, VA., November, 2003. p. 18 – 20.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random Forests. In: HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2009. Cap. 15. p. 1-745.

HENGL, T. et al. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **Plos One,** San Francisco, v. 10, n. 6, p.1-26, 25 jun. 2015. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0125814.

HENGL, T. et al. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. **Geoderma,** Amsterdam, v. 140, n. 4, p.417-427, ago. 2007. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2007.04.022

HUDSON, B.D. The soil survey as a paradigm-based science. **Soil Science Society of America Journal**, Madison, v. 56, p. 836-841 1992.

JENNY, H. **Factors of Soil Formation:** A System of Quantitative Pedology. New York: McGraw-Hill, 1941. 191 p.

KEMPEN, B. et al. Updating the 1: 50,000 Dutch soil map using legacy soil data. **Geoderma,** Amsterdam, v. 151, n. 3-4, p.311-326, jul. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2009.04.023.

KEMPEN, B. et al. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. **Soil Science Society of America Journal,** Madison, v. 76, n. 6, p.2097-2115, 2012. Soil Science Society of America. http://dx.doi.org/10.2136/sssaj2011.0424.

LAGACHERIE, P.; LEGROS, J.P.; BURROUGH, P.A. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. **Geoderma**, Amsterdam, v. 65, p. 283–301, 1995.

LAGACHERIE, P; MCBRATNEY, A. B. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: LAGACHERIE, A. B.; MCBRATNEY, A. B.; VOLTZ, M. (Ed.). **Digital soil mapping:** an introductory perspective. Amsterdam: Elsevier, 2007. Cap. 1. p. 3-24.

LEPSCH, I. F. A necessidade de efetuarmos levantamentos pedológicos detalhados no Brasil e de estabelecermos as séries de solos. **Revista Tamoios**, Rio de Janeiro, v. 9, n. 1, p. 03-15, 2013.

MCKAY, J. et al. Evaluation of the transferability of a Knowledge-Based Soil-Landscape Model. In: BOETTINGER, J. L. et al (Ed.). **Digital soil mapping:** Bridging Research, Environmental Application, and Operation. Dordrecht: Springer, London, 2010. Cap. 14. p. 165-178.

MALLAVAN, B.P.; MINASNY, B.; MCBRATNEY, A.B. Homosoil: a methodology for quantitative extrapolation of soil information across the globe. In: BOETTINGER et al. (Eds.), **Digital Soil Mapping**: Bridging Research, Environmental Application, and Operation. Springer, London, 2010. Cap. 12. p. 137–150.

MALONE, B. P. et al. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. **Geoderma**, Amsterdam, v. 262, p. 243-253, jan. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2015.08.037.

MARMION, M. et al. A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland. **Earth Surface Processes And Landforms,** Hoboken, v. 33, n. 14, p.2241-2254, 13 may 2008. Wiley-Blackwell. http://dx.doi.org/10.1002/esp.1695.

MCBRATNEY, A.B. et al. An overview of pedometric techniques for use in soil survey. **Geoderma**, Amsterdam, v. 97, p. 293-327, 2000.

MCBRATNEY, A.B. et al. From pedotransfer functions to soil inference systems. **Geoderma**, Amsterdam, v. 109, p. 41-73, 2002.

MCBRATNEY, A.B.; SANTOS, M.L.M.; MINASNY, B. On digital soil mapping. **Geoderma**, Amsterdam, v. 117, p. 3–52, 2003.

MENEZES, M.D. de et al. Digital soil mapping approach based on fuzzy logic and field expert knowledge. **Ciênc. Agrotec**, Lavras, v. 37, n. 4, p. 287-298, 2013.

MINASNY, B.; MCBRATNEY, A. B. Digital soil mapping: A brief history and some lessons. **Geoderma,** Amsterdam, v. 264, p.301-311, feb. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.geoderma.2015.07.017.

OMUTO, C.; NACHTERGAELE, F.; ROJAS, R.V. State of the art report on global and regional soil information: Where are we? Where to go? Global Soil Parthership Technical Report. Roma, FAO, 2013. 69p.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A.B. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. **Geoderma Reg**, [S.l], v. 9, p. 17–28, 2017.

PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. **Ecosystems,** [S.l.], v. 9, n. 2, p.181-199, mar. 2006. Springer Nature. http://dx.doi.org/10.1007/s10021-005-0054-1.

QI, F.; ZHU, A-X. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. **Computers & Geosciences,** [S.l.], v. 37, n. 9, p.1425-1436, set. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.cageo.2010.10.016.

ROSSITER, D.G. Digital soil mapping: towards a multiple-use soil information system. In: SEMANA DE LA GEOMÁTICA, 2005, Santa Fé de Bogotá, Colombia, *Anais...* Santa Fé de Bogotá: Instituto Geográfico Agustín Codazzi, 2005. p. 7-15. http://www.css.cornell.edu/faculty/dgr2/Docs/CoGeo2005/PaperRossiterGeomatica2005.pdf (accessed 19.02.17).

SANTOS, H. G. dos et al. Distribuição Espacial dos Níveis de Levantamento de Solos no Brasil. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 34., 2013, Florianópolis, SC. *Anais*....Florianópolis: SBCS, 2013. p. 1 – 4. <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/88902/1/distribuicaoespacial. pdf>. (accessed 15. 07.16).

SCULL, P. et al. Predictive soil mapping: a review. **Progress in Physical Geography**, Thousand Oaks, v.27, p. 171-197, 2003.

SHI, X. et al. Integrating Different Types of Knowledge for Digital Soil Mapping. **Soil Science Society of America Journal**, Madison, v.73, n.5, p. 1682–1692. 2009.

SHI, X. et al. A case-based reasoning approach to fuzzy soil mapping. **Soil Science Society of America Journal**, Madison, v. 68, p. 885-894, 2004.

SILVA, S. H. G. **Digital soil mapping:** Evaluation of sampling systems for soil surveys and refinement of soil maps at lower cost using legacy data. 2016. 104 p. Thesis (Doctorate degree) - Curso de Ciência do Solo, Departamento de Ciência do Solo, Universidade Federal de Lavras, Lavras, 2016.

SKIDMORE, A.K. et al. An operational GIS expert system for mapping forest soil. **Photogrammetric Engineering and Remote Sensing**, Bethesda, v. 62, p. 501-511, 1996.

STUMPF, F. et al. Uncertainty-guided sampling to improve digital soil maps. **Catena,** Amsterdam, v. 153, p.30-38, jun. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.catena.2017.01.033.

YIGINI, Y.; PANAGOS, P. Reference Area Method for Mapping Soil Organic Carbon Content at Regional Scale. **Procedia Earth and Planetary Science**, Amsterdam, v. 10, p.330-338, 2014. Elsevier BV. http://dx.doi.org/10.1016/j.proeps.2014.08.028.

ZHU, A-X. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. **Photogrammetric Engineering & Remote Sensing**, Bethesda, v. 63, p. 1195-1202, 1997.

SECOND PART – ARTICLES

2. ARTICLE 1. Soil type spatial prediction from Random Forest: different training datasets, transferability, accuracy and uncertainty assessment

* Article prepared according to the rules of Scientia Agricola

ABSTRACT

Different uses of soil legacy data as dataset training as well as the selection of soil environmental covariables could drive the accuracy of machine learning techniques. Thus, this work evaluated the performance of the Random Forest algorithm to predict soil classes from different training datasets and extrapolate such information to similar area. The following training datasets were extracted from legacy data: a) point data composed by 53 soil observations (small trenches and soil profiles); b) 30 m buffer around the soil samples (Buffer-Point); soil map polygon with two exclusion zones: c) 20 m from boundaries; d) 30 m from boundaries. These four datasets were submitted to principal component analysis (PCA) for sampling pixels reduction. A total of 52 models were evaluated by means of error of models itself, prediction uncertainty and external validation. The best result was obtained by reducing the number of predictors ensemble with the PCA along with information from buffer around the points. Although Random Forest has been considered a robust spatial predictor model, it was very clear it is sensitive to different strategies of selecting training dataset. Effort was necessary to find the best training dataset for achieving suitable accuracy of spatial prediction. To identify a specific dataset seems to be better than using a great number of variables or a large size of training data. The efforts made allowed the accurate acquisition of a mapped area 15.5 times larger than the reference/legacy area.

Keywords: Digital soil mapping; soil survey; legacy data.

2.1 INTRODUCTION

Soil mapping is an important tool for soil management and planning. However, there is still a need for maps in detailed scale, particularly in countries with sparse areas (Lagacherie and McBratney, 2007) and few financial resources, such as Brazil. The traditional way of mapping soils, based on Pedologists' empirical mental model associated with manually delineating mapping unit boundaries is very time-consuming and onerous (Kempen et al., 2012), although delivering precious maps. Such soil legacy has a potential to be source of training data in machine-learning techniques (Pelegrino et al., 2016), formalizing soil-landscape relationships and applying the information in areas with similar environmental conditions, providing gain in mapped areas with less time and costs (Silva et al., 2016). This is an important strategy of mapping in Brazil, due to the detailed soil surveys being mostly restricted to small areas (Mendonça-Santos and Santos, 2007).

Machine-learning is a computer-based statistical set of tools that could be used to figure out the relationship between soil type and environmental covariables (McBratney et al., 2003; Hastie et al., 2009) that represent soil forming factors (Jenny, 1941). In this context, Random Forest (RF) is one of the most promising techniques regarding digital soil mapping (Chagas, et al., 2016; Rudiyanto et al., 2016; Hengl et al., 2015; Heung et al., 2016; Heung et al., 2017; Souza et al., 2016), in which the way of using legacy data should be investigated for providing a suitable source of data in Random Forest, either from points or polygons.

Considering the influence of soil forming factors in the area of this study, relief is the main driver of soil variability (Menezes et al., 2009). In this sense, several types of digital terrain maps can be generated in Geographical Information System, thus, there has been growing interest in understanding how the characteristics of the environmental covariates influence the accuracy of digital soil mapping (Samuel-Rosa et al., 2015). The choice of effective auxiliary maps (best set of variables) should be sought.

Thus, this work aimed to extract soil information from a reference area (Favrot, 1989; Lagacherie et al., 1995) and extrapolate it to areas with similar soil-landscape relationships. The use of the reference area associated with predictive digital soil mapping approaches can be found in studies such as Grinand et al. (2008) - classification trees; McKay et al. (2010) – fuzzy logic; Arruda et al. (2016) – artificial neural networks; Silva et al. (2016) – Random Forest. The following sequence was implemented and evaluated using Random Forest: a) comparison between point and polygon as source of data to compose training dataset; b) evaluation of the

effects of reducing the number of predictor variables and training-data by principal component analysis on the accuracy of the predicted maps; c) assessment of maps uncertainty.

2.2 MATERIAL AND METHODS

2.2.1 Study Area

The study area is divided into reference area, named Vista Bela Creek watershed, where from the legacy data was extracted (175 ha) for model training, and extrapolation area (2,719 ha), where a new soil map was generated (Figure 2.1). Both areas are located in Minas Gerais state, Brazil, between the coordinates UTM 553781 and 581138 m, 7598766 and 7597100 m, 23K, datum WGS 1984, with an elevation range of 924-1342 m. The relief at the area was modeled through intense dissection provided by fluvial erosion, resulting in hilly features with convex to tabular summit and convex slopes, interspersed by elongated crests. There is predominance of gneisses and biotite-schists of Carrancas sequence and biotite-gneiss and amphibolite of Serra do Turvo sequence. According to Köppen classification, the climate is Cwa, with dry winter and rainy summer. The mean annual temperature varies between 18 and 22°C, presenting an annual precipitation average of 1,450mm (Menezes et al., 2009).

The main soil types that occur in the area are, classified according to US Soil Taxonomy (Soil Survey Staff, 2014), are Udept, Hapludox, Acrudox, and Fluvent (Menezes et al., 2009) according to Soil Taxonomy (Soil Survey Staff, 2014). Orthent are also found, but, occurring as inclusions associated with rock outcrops, in an intricate landscape pattern with Inceptisols, which may hinder its individualization, and consequently, the knowledge transferability.

The soil legacy data is composed by a soil map in detailed scale (1:10,000) with simple mapping units (Menezes et al., 2009). Table 2.1, soils are referred to as classified by the Soil Taxonomy classification (Soil Survey Staff, 2014). The soil map was produced by an experienced team in a traditional basis: analysis of aerial photography and manual delineation of soil mapping units, along with intensive fieldwork (total of 53 soil profiles). This watershed is considered as a reference area (Favrot, 1989; Voltz et al., 1997), since it comprises the whole soil-landscape relationships occurring in the region that can be extrapolated to areas with similar physiographic conditions. This map was used as the source of information for training Random Forest models.

Figure 2.1 - Study areas location: Vista Bela Creek Watershed (soil legacy - reference area) (Menezes et al., 2009) and the area to which information was extrapolated in Minas Gerais state, Brazil.

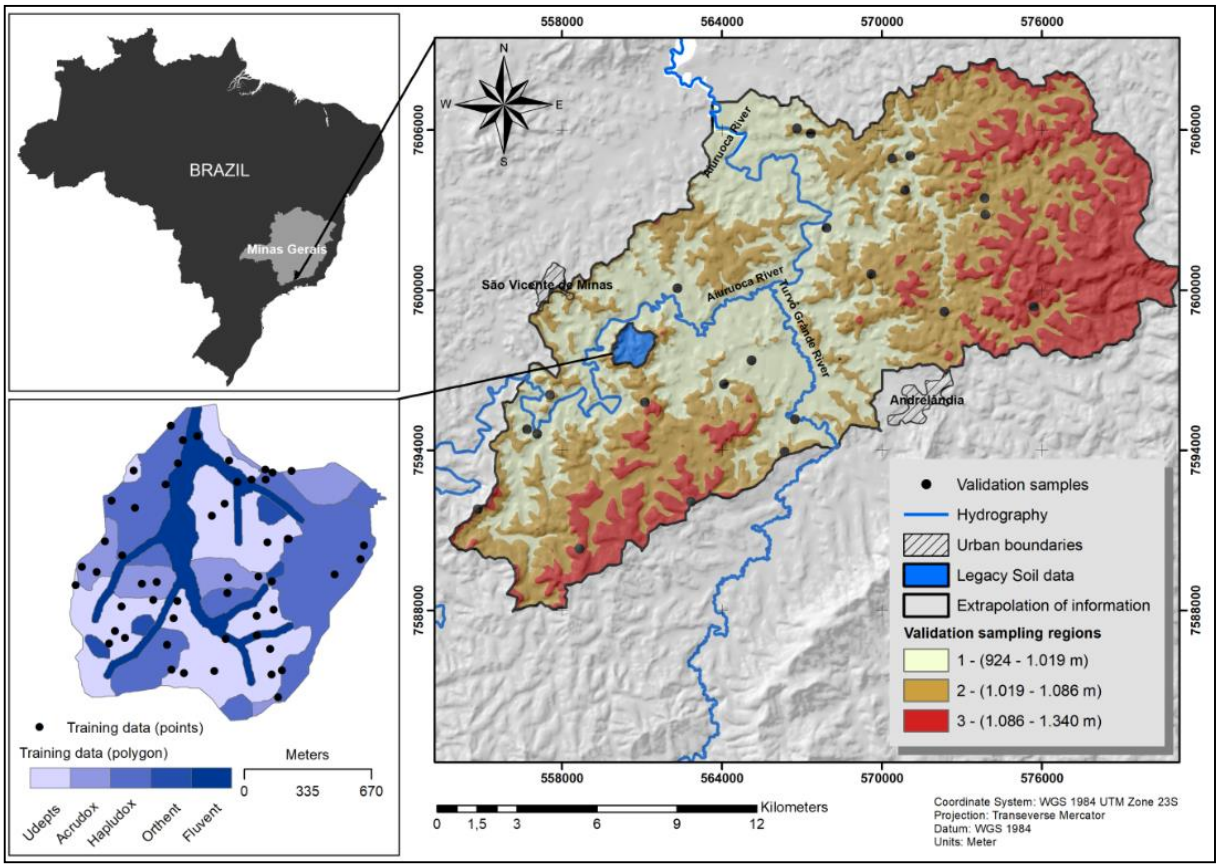


Table 2.1 - Mapping units identified in the study area.

Symbol	Soil classes	Area (ha)	%
Hx	Hapludox	61.2	35
Ax	Acrudox	21.3	12.2
Ut	Udept	61.3	35
Ft	Fluvent	27.2	15.5
Ot	Orthent	4	2.3
Total		175	100

Source: Menezes et al. (2009)

2.2.2 Environmental covariates: relief maps

A digital elevation model (DEM) with 20 m of resolution was generated from contour lines freely available in Brazil from Brazilian Institute of Geography and Statistic (IBGE), at a 1:50,000 scale and 20 m of equidistance. A hydrologic consistent DEM was generated in ArcGIS (version 10.1 of ESRI) through the Topo To Raster tool. From the DEM, 14

topographic indexes were created using the SAGA GIS SOFTWARE (SAGA Development Team version 3.0) and selected due to their capacity to express variations of both morphometrical and hydrological characteristics at local and landscape scales. The following topographic indexes were calculated: catchment slope (CS), convergence index (CI), plan curvature (Plan C) and profile curvature (Prof C) (Zevenbergen and Thorne, 1987), multiresolution index of ridge top flatness (MRRTF) (Gallant and Dowling, 2003), slope, LS-factor, SAGA wetness index (SWI), topographic position index (TPI) (Guisan et al., 1999), terrain surface texture (Texture) (Iwahashi and Pike, 2007), terrain classification index for lowlands (TCI), upslope curvature (USC), valley depth (VD), vertical distance to channel network (VDCN).

2.2.3 Random forest: characteristics and accuracy of models

In order to establish the distribution of soil types according to their relation with topographic indexes, the *randomForest* package (version 4.6-12) in the statistical software R (R Development Core Team, version 1.0.44) was used. The Random Forest algorithm consists of a combination of prediction trees, in which each node is split using the best subset among predictors randomly chosen at that node. The classification procedure consists of growing a predefined number of unpruned classification trees, defined by the parameter *ntree*, from bootstrap samples (2/3) of the entire population, n. Each tree is constructed using a different bootstrap sample from the original data. At each node, instead of choosing the best split predictor considering all variables (p), the predictor is identified from a random subset of predictors, where the number of predictors tried at each split, *mtry*, is defined by the user. For classification trees, the default value for *mtry* is $\lfloor\sqrt{p}\rfloor$ (Hastie et al., 2009; Heung et al., 2014). After a large number of generated trees, the algorithm votes for the most popular class (Breiman, 2001).

Each bootstrap sample leaves out about one-third of the observations. These left-out data are called out-of-bag (OOB) observations. Thus, it is possible to predict the response for the i^{th} observation, using each of the trees in which that observation was OOB. Performing this procedure with all OOB data allows obtaining the OOB estimate of error rate. The resulting OOB error is a valid estimation of the models, since the response for each observation is predicted using only the trees that were not fit employing that observation (James et al., 2013). So, as result of Random Forest proceedings, a single model is established along with classification error estimation (Heung et al., 2014).

Random Forest also uses the OOB samples to measure the importance of predictors (topographic indexes) that can be useful for reducing the number of variables and for interpreting the fitted forest (Cutler et al., 2012). The Random Forest algorithm provides two measures of variable importance:

- mean decrease in accuracy (MDA): evaluates a variable contribution from OOB error estimate, in order to determine changes in prediction accuracy by randomly permuting a single predictor in the OOB data. By measuring the increase in error, this procedure can rank the variables according to their importance in establishing accurate predictions.

- mean decrease in Gini (MDG): indicates the predictors importance based on the quality of each split. A variable that produces less heterogeneity in the descendent nodes scores better at MDG rank (Breiman, 2001).

In summary, a set of soil types information derived from the reference area and their associated topographic indexes were used to train the Random Forest classifier. By feeding the algorithm with different sets of variables, a group of non-spatial Random Forest models was developed and then applied to all unknown points of the study area, resulting in a series of soil type maps in a raster format for the entire study area. The different dataset training used as soil legacy data with their related topographic indexes are presented below.

2.2.4 Training dataset

The complete framework, including the choice of training dataset until the spatial prediction of soil types of the extrapolated area, is presented in the Figure 2.2. The *randomForest* package (version 4.6-12) in the statistical software R (R Development Core Team, version 1.0.44) was used. The choice of *mtry* is often the square root of the number of variables (*p*), in this case it was 4 and the parameter *ntree* was adjusted to 1000. The following way of using legacy data for training Random Forest was applied:

- Point legacy data (Figure 2.2A): a) 53 soil legacy samples; and b) a circular buffer of 30 m radius around each soil sample point, aiming to increase the number of points with soil information to be used by the Random Forest. The buffer increases the size of training dataset, which, in turn, could improve the accuracy of Random Forest prediction (Deng and Wu, 2013).

- Polygons of soil mapping units (Figure 2.2B): a) pixels from the interior of the polygons eliminating 20 m from their boundaries; and b) pixels from the interior of the polygons eliminating 30 m from their boundaries. The use of information closer to the mapping unit boundaries has been frequently avoided, since it is a transitional zone where there is greater

uncertainty (ten Caten et al., 2012; Giasson et al., 2015; Pelegrino et al., 2016). The use of polygon data in predictive models could bring great dataset training gain, and it has been reported as one of the greatest advantages by capturing more details of the landscape variability and the multivariate feature space of a categorical variable (Heung et al., 2016).

- PCA of Polygons and Points training datasets (Figure 2.2C): PCA was applied (FactoMineR package, version 1.36) by means of R software environment (R Development Core Team, version 1.0.44). The PCA extracted the most important information from a multivariate data set and expressed it as a whole new set of information, called principal components. Considering that the contribution of the individuals (pixels) to the principal components of a given dataset can be measured, it was possible to reduce the data to a new ensemble more aligned to the variables, according to the Figure 2.3. The red line in the Figure 2.3A indicates the individuals expected average contribution (EAC). For a given component, an observation with a contribution larger than this cutoff could be considered as important in contributing to the component, reducing the subjectivity in explanatory information reduction. The Figure 2.3B shows the variation of contribution of the dataset. The closer to the center, the lower the contribution of a given observation. Therefore, the contribution of individuals was calculated for each training dataset described above, and the pixels with values below the EAC were excluded (Figure 2.3C). With this procedure, four additional training datasets were created, namely PCA-Point, PCA Buffer-Point, PCA Pol -20 m, PCA Pol -30 m.

Figure 2.2 - Flowchart of training data scheme and their interaction with the variables. A- composition of Point training datasets. B - development of Polygon training datasets. C - training data reduction for development of PCA training datasets. D - summary of the proposed methodology.

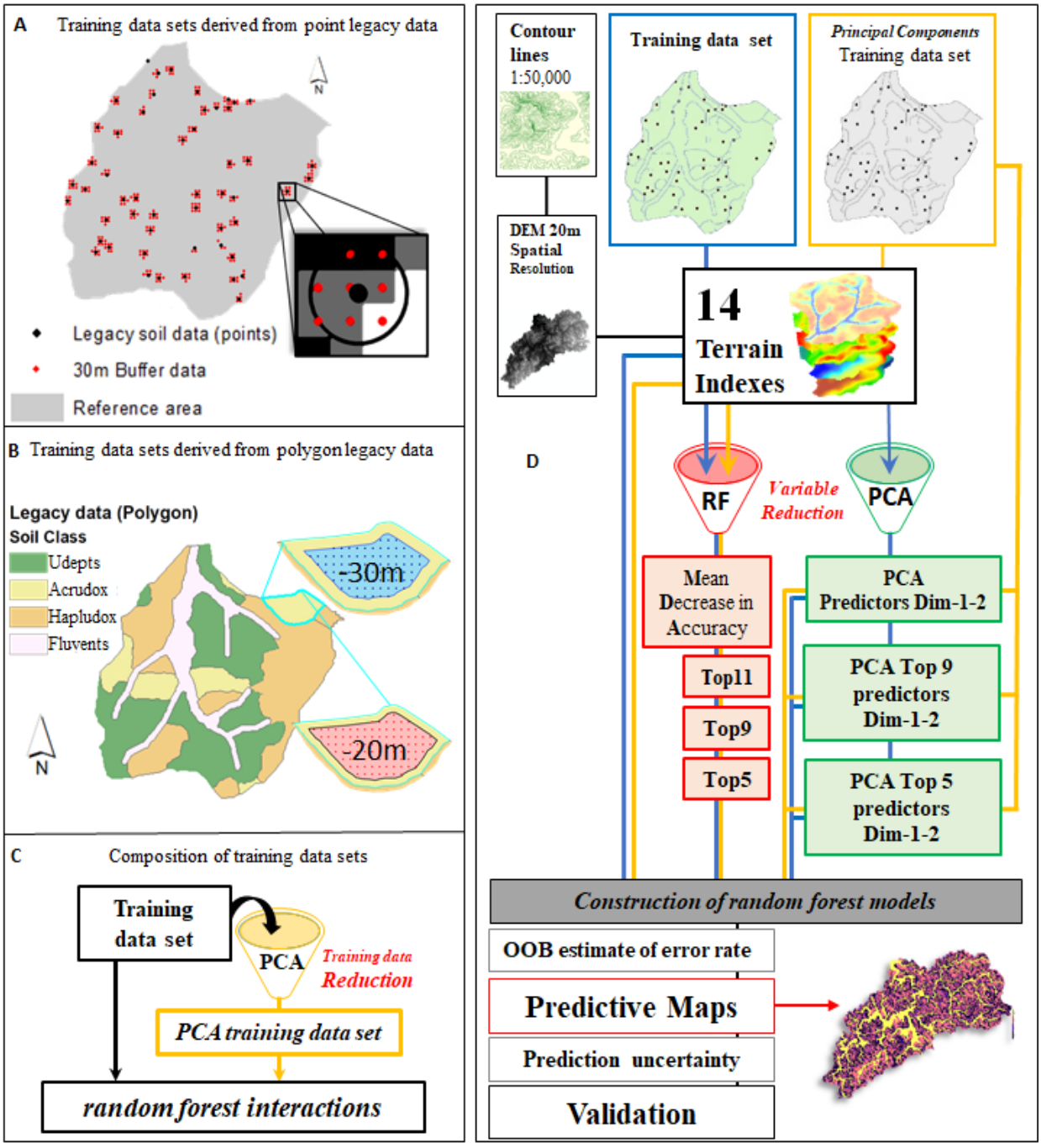
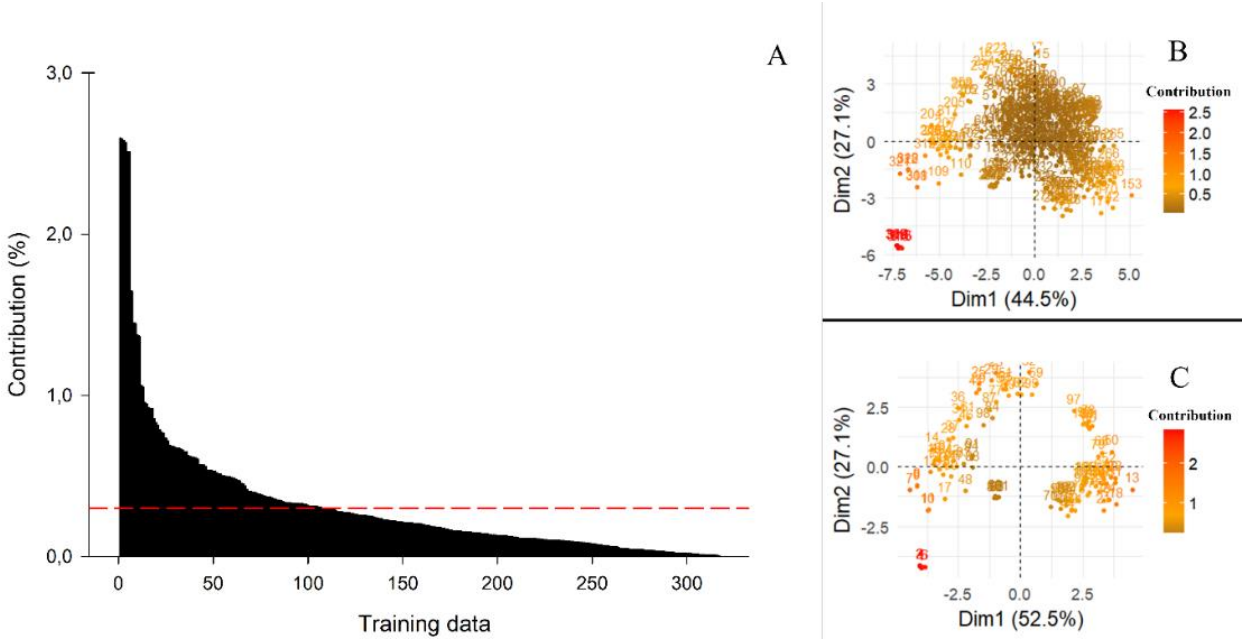


Figure 2.3 - A - Contribution of individuals to dimensions-1-2 of the principal component analysis for the data set. B - The variation of contribution of the data set. C – Reduction of data dimension. Dim - Dimension.



2.2.5 Variables reduction

Possessing a large number of available information does not always contribute to generate accurate models. Even the Random Forest being efficient in applying multiple predictors, the accuracy is not always improved (Svetnik et al., 2003). In this sense, the different kinds of tests, in order to assess the effects of variable reduction in spatial prediction (Figure 2.2D), were developed:

1. The Random Forest classifier was initially loaded with the entire set of predictors (topographical indexes) for each one of the soil information data set (control);

2. Based on Random Forest algorithm, the mean decrease in accuracy was obtained and the variable importance was ranked. For each dataset, the top eleven (MDA11), nine (MDA9) and five (MDA5) variables were selected.

3. The whole set of soil data and its correspondent terrain indexes were submitted to PCA. The reduction of variables was performed from the expected average contribution of the variables for the dimensions 1-2 of PCA. For the given components, the variables with a contribution lower than this cutoff were excluded (PCA1-2).

4. For each dataset from the dimensions 1-2 of the PCA had the top nine attributes selected from a rank (PCA9).

5. For each dataset, from the dimensions 1-2 of the PCA the top five attributes (PCA5) were selected from a rank.

It is important to highlight that PCA was performed for both reduction of predictor variables and training points. Thus, in the procedures 3, 4, and 5 aforementioned, the Random Forest was loaded with the ensemble of variables defined for their original training sets (Figure 2.2D).

2.2.6 Assessment of predictions accuracy within extrapolation of information area

The assessment of the Random Forest accuracy was done with 23 soil profiles (external validation), at which was called here as extrapolation of information area (Figure 2.1). The sampling sites were chosen by means of Regional Random method on ArcSIE (Soil Inference Engine - ArcGIS extension, version 10.3.101). The locations were randomly defined within polygons, representing three altitude levels (sampling regions) as shown in Figure 2.1. Two indexes were calculated: overall accuracy and kappa index. The overall accuracy is the sum of the main diagonal components of the confusion matrix divided by the total of validation samples, as follow:

$$OA = \frac{\sum_{j=1}^{c} xjj}{N}$$

where: xjj is the number of correct samples and N is the total number of samples.

Kappa index is an agreement measure calculated taking into account the total number of samples, the number of soil types and the correctly classified samples (Congalton and Green, 2008). The values may range from -1 (suggesting disagreement) to 1 (suggesting excellent agreement) (Landis and Koch, 1977).

$$KI = \frac{N \sum_{j=1}^{c} xjj - \sum_{j=1}^{c} xjixij}{N^2 - \sum_{j=1}^{c} xjixij}$$

where xij is the value in a row i and in a column j; xji is sum of values in line i; xij is the sum of values in column j; N is the total number of samples (points used for validation); and c is the number of soil types.

User's accuracy and producer's accuracy were also calculated. User's accuracy shows the probability of the predicted class on the map to match the class at the field, while the producer's accuracy expresses the probability of a soil type point being correctly classified in the map (Congalton, 1991). The indexes are presented by the equations below:

$$\text{User's accuracy} = \frac{Xii}{\sum_{i=1}^{r} Xij}$$

$$\text{Producer's accuracy} = \frac{Xjj}{\sum_{j=1}^{r} Xij}$$

where Xii and Xjj are the number of correctly classified samples and Xij the sum of samples of a soil type in a row (user's accuracy) or column (producer's accuracy) of a confusion matrix. An accurate map has indexes values closer to one (100%) (Behrens et al., 2010).

2.2.7 Prediction uncertainty

The prediction uncertainty was evaluated by vote count and entropy maps. The ensemble-modeling, like Random Forest, has as benefits the possibility to estimate the uncertainty by using the vote count surface. In this study, each model corresponds to 1,000 interactions. By the end of the procedure, each pixel receives 1,000 votes. Thus, the range of votes varies from 0% to 100%. Pixel values closer to 0% or 100% indicate less uncertainty. The higher the value, the greater the certainty of that pixel to correspond to a given soil type. The lower the value, the higher the certainty of a given pixel does not correspond to a given soil type. Therefore, the values in between this range carry some uncertainty.

To represent the overall uncertainty, the entropy measure (*H*) was used to describe how the ensemble-model intent their predictions to a particular soil type. It expresses the degree of certainty in a pixel classification which the votes are concentrated in a particular class, rather than spread over a number of classes. *H* is calculated as follows (Zhu, 1997):

$$H(\mathrm{x}) = \frac{1}{1\mathrm{n}n} \sum_{k=1}^{n} Pk(x)\ \ln Pk(x)$$

where Pk is the proportion of instances where pixel x is classified as soil types k and where n is the number of members in the ensemble-model. The entropy values range from 0 to 1, which the higher the entropy value at a location, the higher the uncertainty of classification.

To better understand the uncertainty in predictions, a landforms map was generated. The DEM was selected as input data to the TPI based landform classification module on SAGA GIS (Weiss, 2001) resulting into ten landform classes of the study area (APPENDIX D.) The derived landform classes were intersected with the vote-count and entropy maps for the interpretation of the predictions uncertainty distribution on the landscape.

2.3 RESULTS AND DISCUSSION

2.3.1 Model Evaluation

According to the Table 2.2, the OOB estimate of error varied in a wide range (from 5 to 77%). This index seems to be mainly driven by the number of observations: the error decreases while the number of observations increases. Such difference is clear when comparing the models with training datasets derived from points and from polygons, the last one showing more observations and less errors. For the group of polygons, those that were reduced using PCA presented mean OOB values slightly lower than their respective original sets. However, when analyzing the whole models by means of training data reduction (Figure 2.4), two different groups of OOB estimated of error were found: those with less than 53 and more than 105 training data observations, with or without PCA analysis, as seen in Table 2.2.

Such results indicate that Random Forest models were sensitive to variations in training dataset. Larger training dataset is often necessary to decrease error (Pal et al., 2003), and in this study, such information also brought stability to the errors of the models in training data above 105 observations. However, it is important to highlight that the use of polygons and buffers could bring some uncertainty about the soil type, mainly closer to the boundaries or transition zones (Pelegrino et al., 2016; ten Caten et al. 2012; Giasson et al., 2015). Thus, the key point here is if the more accurate models will deliver accurate soil maps in the extrapolated area.

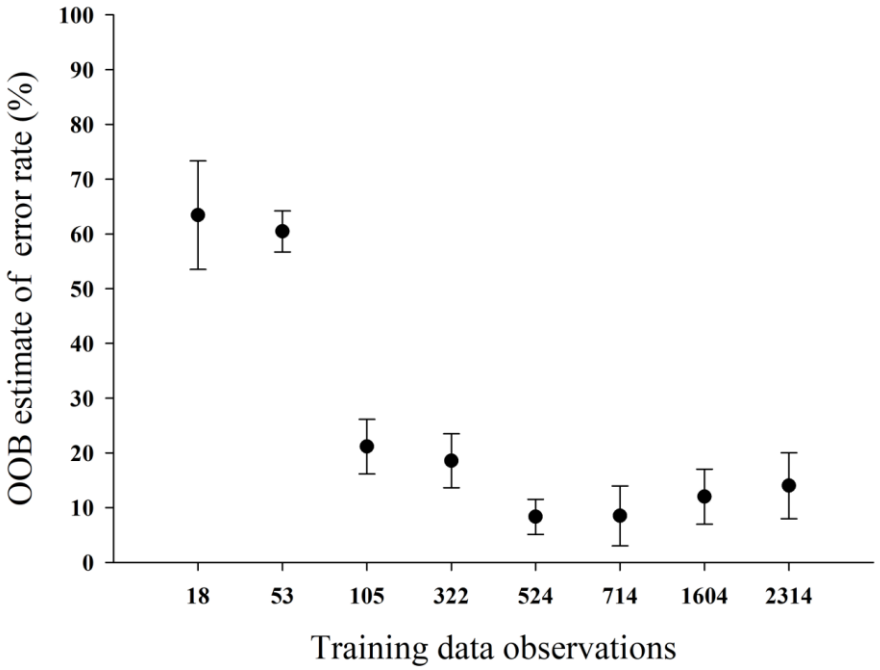
Table 2.2 - Accuracy measurements of the models developed from point and polygon data (continue)

Training dataset	Number of observations	Variable selection	Number of variables	OOB estimate error	Overall accuracy	Kappa Index
				----------%----------		
Point	53	Control	14	62	57	0.358
		MDA(11)	11	57	57	0.345
		MDA(9)	9	54	48	0.209
		MDA(5)	5	64	43	0.143
		PCA1-2	11	60	57	0.345
		PCA (9)	9	62	57	0.345
		PCA (5)	5	64	39	0.069
PCA-Point	18	Control	14	77	49	0.110
		MDA(11)	11	72	40	0.110
		MDA(9)	9	56	40	0.093
		MDA(5)	5	55	49	0.159
		PCA1-2	11	67	49	0.110
		PCA (9)	9	67	39	0.100
		PCA (5)	5	50	35	0.004
Buffer-Point	322	Control	14	15	65	0.476
		MDA(11)	11	15	65	0.476
		MDA(9)	9	16	70	0.540
		MDA(5)	5	17	70	0.550
		PCA1-2	11	20	65	0.476
		PCA (9)	9	18	65	0.476
		PCA (5)	5	29	70	0.546
PCA Buffer-Point	105	Control	14	18	61	0.410
		MDA(11)	11	20	61	0.410
		MDA(9)	9	18	61	0.410
		MDA(5)	5	21	48	0.211
		PCA1-2	11	21	65	0.474
		PCA (9)	9	18	61	0.409
		PCA (5)	5	32	83	0.738

		Control	14	11	57	0.324
		MDA(11)	11	9	52	0.260
		MDA(9)	9	9	52	0.258
Pol -20m	2,314	MDA(5)	5	15	57	0.327
		PCA1-2	*	*	*	*
		PCA (9)	9	15	57	0.337
		PCA (5)	5	25	44	0.158
		Control	14	6	50	0.226
		MDA(11)	11	5	57	0.343
		MDA(9)	9	5	66	0.474
PCA Pol -20m	714	MDA(5)	5	6	61	0.417
		PCA1-2	*	*	*	*
		PCA (9)	9	10	53	0.262
		PCA (5)	5	19	40	0.118
		Control	14	8	52	0.267
		MDA(11)	11	8	52	0.256
		MDA(9)	9	9	48	0.191
POL -30m	1,604	MDA(5)	5	12	48	0.207
		PCA1-2	*	*	*	*
		PCA (9)	9	14	57	0.327
		PCA (5)	5	21	44	0.172
		Control	14	6	52	0.262
		MDA(11)	11	6	57	0.343
		MDA(9)	9	6	57	0.335
PCA Pol -30m	524	MDA(5)	5	8	52	0.254
		PCA1-2	*	*	*	*
		PCA (9)	9	10	53	0.279
		PCA (5)	5	14	44	0.160

PCA – principal component analysis; OOB out-of-bag observations. *For PCA-1-2, there are no values for the polygons group because only 9 variables (Terrain Indexes) reached the expected average contribution; Control – All terrain indexes applied for RF spatial prediction; MDA-Variables reduction by means of Mean Decrease in Accuracy; PCA-Variables reduction by means of Principal Component Analysis.

Figure 2.4 - Variation of OOB estimate of error rate in function of the number of training data observation (pixels).



2.3.2 Assessment of extrapolated information (external validation)

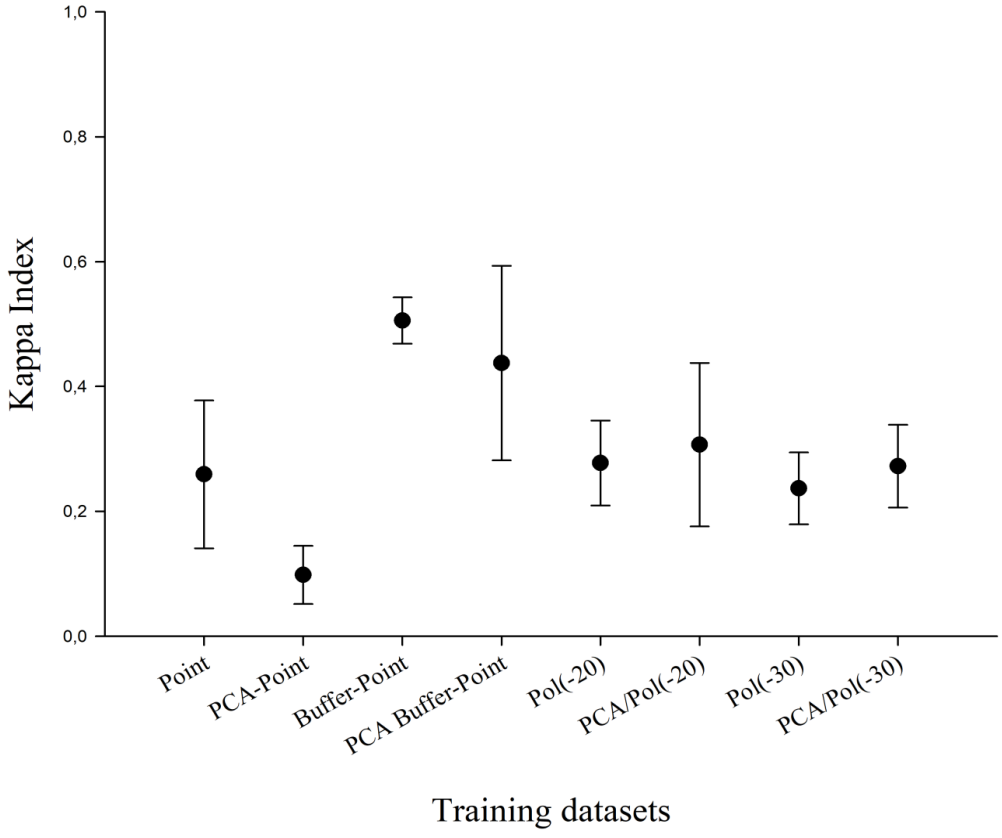
Table 2.2 presents the overall accuracy and Kappa index derived from external validation within the area, to where the information was extrapolated. Figure 2.5 presents the Kappa index organized by training dataset groups. The Point derived models presented the poorest prediction when compared with the Buffer-Point or Polygons, with or without PCA analysis. Also, an increasing in number of data observation does not bring significantly improvement of accuracy, in disagreement with OOB estimate error from the model. Polygon derived models presented intermediate Kappa values, ranging from 0.118 to 0.474, while the Point derived models (original and buffered) gave rise to maps with both lowest and greatest accuracy values (Kappa index from 0.004 to 0.738). The map with the highest absolute accuracy came from PCA Buffer-Point dataset, with 0.738 for the Kappa index and 83% for overall accuracy.

As long as digital soil mapping techniques attempt to take advantage of a large number of explanatory environmental covariates (McBratney et al., 2003), with a relative small proportion of sampling points, the ability of Random Forest to deal with high dimensional datasets should be tested. Thus, the reduction of dimensionality by means of PCA analysis (Behrens et al., 2010) or calibration of data set selection (Kuang and Mouazen, 2011) would improve the accuracy of spatial prediction, since the most important subsets are used (Millard

and Richardson, 2013). In this study, the use of PCA resulted only in a slight improvement of overall accuracy of models. Nevertheless, as already mentioned, the PCA Buffer-Point dataset presented the most accurate map among all 52 prediction models, as evaluated by Kappa index and overall accuracy.

It is possible to notice large variations of accuracy even within the same type of training datasets (Figure 2.5), variations are due to the choice of terrain indexes. In order to better understand the effects of terrain indexes or variables reduction, Table 2.3 presents the difference between the overall accuracy of the control (all terrain indexes as an input on Random Forest) and the reduced ensembles of each training dataset. It is expected that where the most important input data are used, the accuracy would increase (Strobl et al., 2009; Millard and Richardson, 2013). In this study, in general, variable reduction was not related with increasing accuracy. Millard and Richardson (2015) noted high fluctuations regarding the variable importance, even when the same training data was used. Thus, another way to select variable importance from Random Forest output should be tested, seeking for model stability and accuracy improvement.

Figure 2.5 - Accuracy for each dataset based on Kappa Index.



Only two training datasets (PCA Buffer-point; PCA Pol-20m) presented at least one model with relevant increase in overall accuracy (higher than 15% in overall accuracy). No relevant variation or reduction in accuracy was found for the others. Thus, the relationship between models predictive capacity and terrain indexes cannot be explained only by the number of predictor variables used in each model. As an example, from the best results obtained for the Buffer-Point dataset, the variables reduction resulted in a slight or no variation on maps accuracy. Moreover, different sets of terrain indexes presented the same overall accuracy for the same dataset (Buffer-Point MDA(5), 70% and Buffer-Point PCA(5), 70%) as seen in Table 2.2.

Table 2.3 - Difference of overall accuracy of the reduced ensembles of variables in relation to the control, for each training dataset.

Variables		Point	PCA-Point	Buffer-Point	PCA Buffer-Point	Pol (-20m)	PCA-Pol (-20m)	Pol (-30m)	PCA-Pol (-30m)
Ensemble	Number	%							
Control	14	57	49	65	61	57	50	52	52
MDA(11)	11	0	-9	0	0	-5	+7	0	+5
PCA-1-2	11	0	0	0	+4	*	*	*	*
MDA(9)	9	-9	-9	+5	0	-5	+16	-4	+5
PCA(9)	9	0	-10	0	0	0	+3	+5	+1
MDA(5)	5	-14	0	+5	-13	0	+11	-4	0
PCA(5)	5	-18	-14	+5	+22	-13	-10	-8	-8

PCA – principal component analysis. *For PCA-1-2, there are no values for the polygons group because only 9 variables (Terrain Indexes) reached the expected average contribution; MDA-Variables reduction by means of Mean Decrease in Accuracy; PCA-Variables reduction by means of Principal Component Analysis.

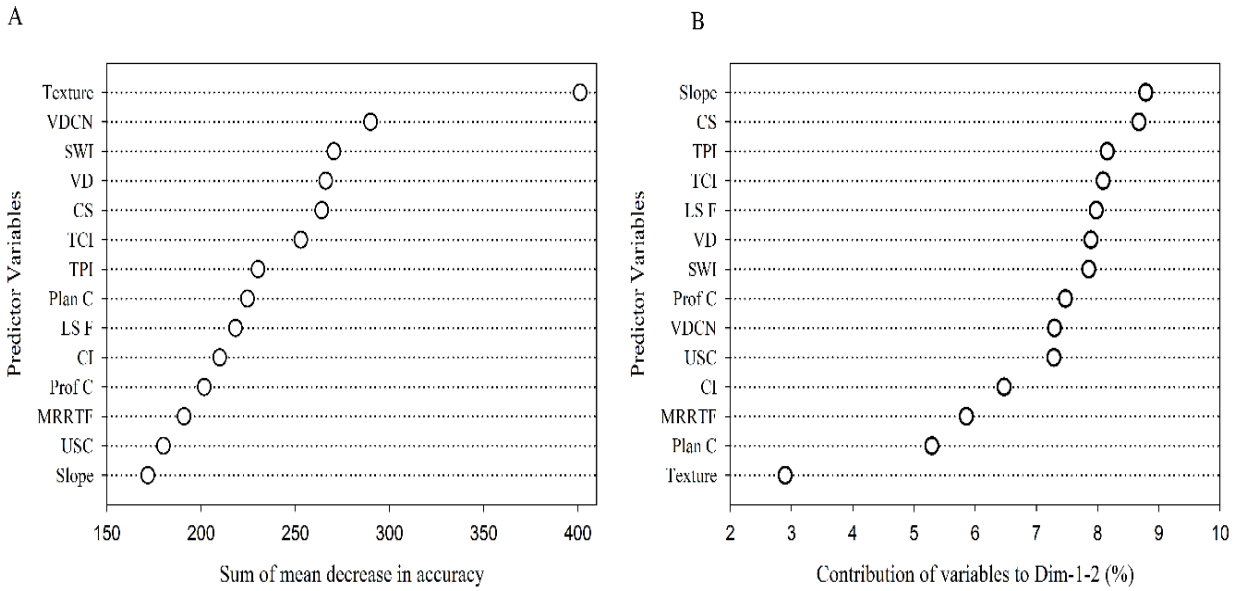
In accordance with the findings of Heung et al. (2014), in this study, the variable reduction did not necessarily result in great accuracy improvements with Random Forest. However, the best result obtained in our study was achieved with the reduction of variables. For the PCA Buffer-Point dataset, by reducing the number of the variables to the 5 most important ones pointed out by the PCA, there was a 22% improvement in the accuracy of the map, in accordance with Table 2.3. In contrast, for the same training data, using the same predictor variables set size, although defined by MDA, the accuracy of the map was 13% lower

when compared to the control model, as seen in Table 2.3. This result is contrary to those obtained by Behrens et al. (2010), which reported that the unsupervised PCA approach turned out to be the worst technique in terms of selecting optimal features for soil classification.

Regarding variable reduction, it is important to note that, there is no single method for best ranking classifiers from distinct datasets (Novakovic et al., 2011). Different ranking methods may result in different classifications, as shown in Figure 2.6. Moreover, a poorly ranked variable that could be considered useless by itself, can afford an expressive performance enhancement when combined with others (Guyon and Elisseeff 2003). At this study, e.g., for PCA Buffer-Point dataset, the best predictor subset was obtained based on PCA-1-2 ranking, composed by the terrain indexes CS, TPI, TCI, LSF and SLOPE, whose MDA order of importance were 11^{th}, 6^{th}, 7^{th}, 13^{th} and 14^{th}, respectively. Once the accuracy is influenced by the choice of features, it is reasonable the use of many rank indices in order to assure that the most accurate subset will be obtained (Novakovic et al., 2011). Another important aspect of identifying the main variables is the time saving in the acquisition and preparation of database and the computational efficiency, if there is interest in applying such model in larger and similar areas (Scarpone et al., 2017; Yu et al., 2016).

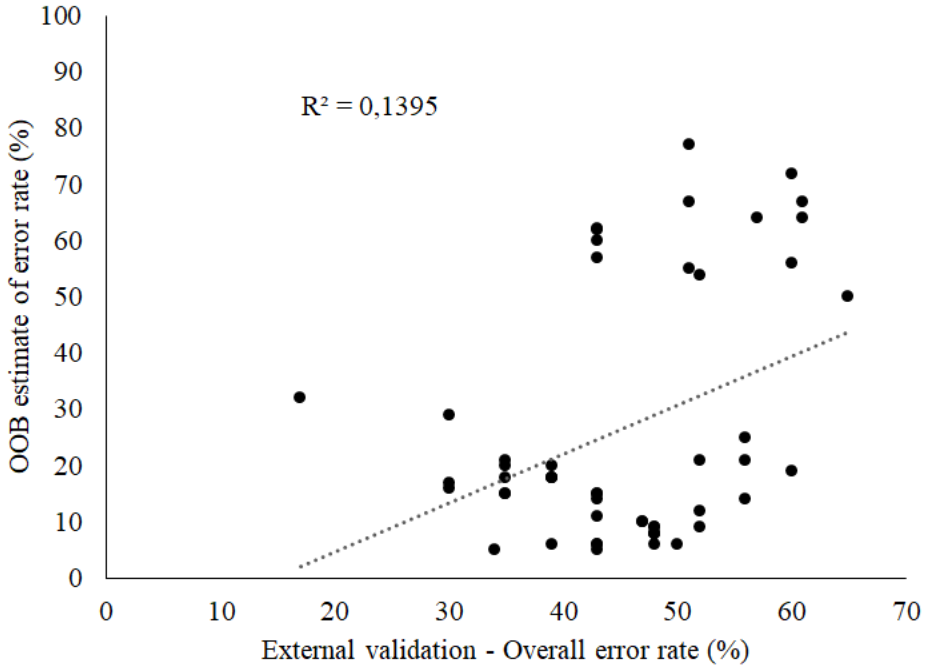
Having a large training dataset and a numerous ensemble of variables does not necessarily result in accurate predictions, despite the low values on OOB error rate. Figure 2.7 shows the relationship between the overall error rate and the OOB estimate of error rate. A weak correlation between model and external validation was found ($R^2 = 0.1395$). Before an extensive sequence of tests in different types of Random Forest training datasets, Millard and Richardson (2015) pointed out that OOB was not a good indicator of error in highly dimensional datasets, and it seems to be driven mainly by dataset training size, as already discussed. Thus, it is recommended to explore different combinations of predictor variables for a single dataset, once, to load the random forest with the whole covariates ensemble, do not necessarily results in the most accurate map, as well as to provide an independent validation data set in order to avoid optimistic bias (Hammond and Verbyla, 1996).

Figure 2.6 - Overall variable importance. A - Variable importance based on mean decrease in accuracy. B - Variable importance based on its contribution for the dimensions-1-2 of principal component analysis.



Dim – dimensions; CS – Catchment slope; CI – Convergence index; PlanC – Planform curvature; ProfC – Profile curvature; LSF – LS-Factor; MRRTF – Multiresolution index of ridge top flatness; SWI – Saga Wetness Index; TPI – Topographic Position Index; TCI – Terrain Classification for Low Lands; USC – Upslope Curvature; VD – Valley Depth; VDCN – Vertical Distance to Channel Network.

Figure 2.7 – Correlation between variations of Out of Bag (OOB) estimate of error rate and the overall rate of external validation.



2.3.2 Prediction uncertainty

Uncertainty analysis was done in the Random Forest predictions for the two best models obtained (point-derived and polygon-derived training dataset). The vote count surfaces of the soil types are presented in the Figure 2.8. Higher values correspond to areas most likely to occur for a given soil type, and the lower values indicate the opposite. Both can be considered areas of low uncertainty. Therefore, intermediate values are indicators of the areas of greatest uncertainty on prediction.

There were quite differences when comparing the vote count surfaces from point-derived training data and the polygon-derived training data. The polygon derived model seems to oversize the area of certainty for the probability of occurrence of Udepts, advancing over areas where Fluvents are expected (APPENDIX F). Despite this, the producer's accuracy for this class was 50%, demonstrating that, in addition to oversizing, spatialization was also impaired. For the Oxisols (Acrudox and Hapludox), the general spatialization pattern was considered similar to that of the point-derived model. However, the dimensionalization may be overestimated since the producer's accuracy (84% and 100%) was greater than user's accuracy (63% and 50%) for the soil type map. The model derived from the polygons was also less efficient in discriminating Fluvents compared to the point-derived one (Figures 2.8G and 2.8H).

Great extensions of uncertainty over areas of Udepts and Hapludox votes surface maps were found for the point-model compared to polygon-model. This effect may come from the least amount of training data for the point dataset, which was respectively 8.4 and 2.4 times greater for polygon derived model. However, both soil types presented satisfactory values of producer's accuracy and user's accuracy (Table 2.4). In this case, the greatest uncertainty was found in Acrudox spatial prediction.

The overall uncertainty prediction was represented by entropy (Figure 2.9A and 9B). Polygon and Buffer-Point presented quite similar results: polygon-derived the entropy ranged from 0 to 0.99, with average of 0.478 and standard deviation of 0.198; for the polygon-derived models the entropy ranged from 0 to 0.99, with average of 0.478 and standard deviation of 0.198; for the Buffer-Point derived models the entropy also ranged from 0 to 0.99, with average of 0.467 and standard deviation of 0.169. Contrary to what was observed by Heung et al. (2017), there was no major difference in the spatial distribution of the overall uncertainty over the study area, considering the different datasets.

Table 2.4 - Producer's and user's accuracy for the highest accuracy spatial prediction from buffer of point and polygon data.

Dataset	Variables	Udept	Acrudox	Hapludox	Fluvent
		Producer's accuracy			
		%			
PCA POL-20	MDA(9)	50	100	84	50
PCA Buffer-Point	PCA(5)	80	0	88	100
		User's accuracy			
		%			
PCA POL-20	MDA(9)	78	50	63	25
PCA Buffer-Point	PCA(5)	89	0	88	100

PCA POL-20 - training data derived from polygon -20 m dataset and reduced with Principal Component Analysis; PCA Buffer-Point - training data derived from Buffer-Point dataset and reduced with Principal Component Analysis; MDA(9) – the best nine variables of the Mean Decrease in Accuracy rank; PCA(5) – the five variables that most contributed for Principal Component Analysis.

Figure 2.10 shows the relative frequency distribution of the uncertainty related to landforms. In general, the uncertainty was low for valley bottom regions, where there is predominance of Fluvents occurring over flatter areas around the drainage network, being formed by the accumulation of sediments from floods depositions. Such values were also found in flat ridge tops and plains, commonly associated to Hapludox.

In sites where the slope is greater than 20%, the entropy values range from low to intermediate, and the steeper the slope, the lower the uncertainty. Such sites are commonly associated with the occurrence of Inceptisols. In the region of the study area, this soil type tends to be located in a wide range of slope gradient (3% to 45%).

Figure 2.8 - Vote count surfaces on 1000 decision trees of the Random Forest using Point and Polygon derived training data. A – Udept from Point data; B – Udept from Polygon data; C – Hapludox from point data; D – Hapludox from polygon data; E – Acrudox from point data; F – Acrudox from polygon data; G – Fluvents from point data; H – Fluvents from polygon data.

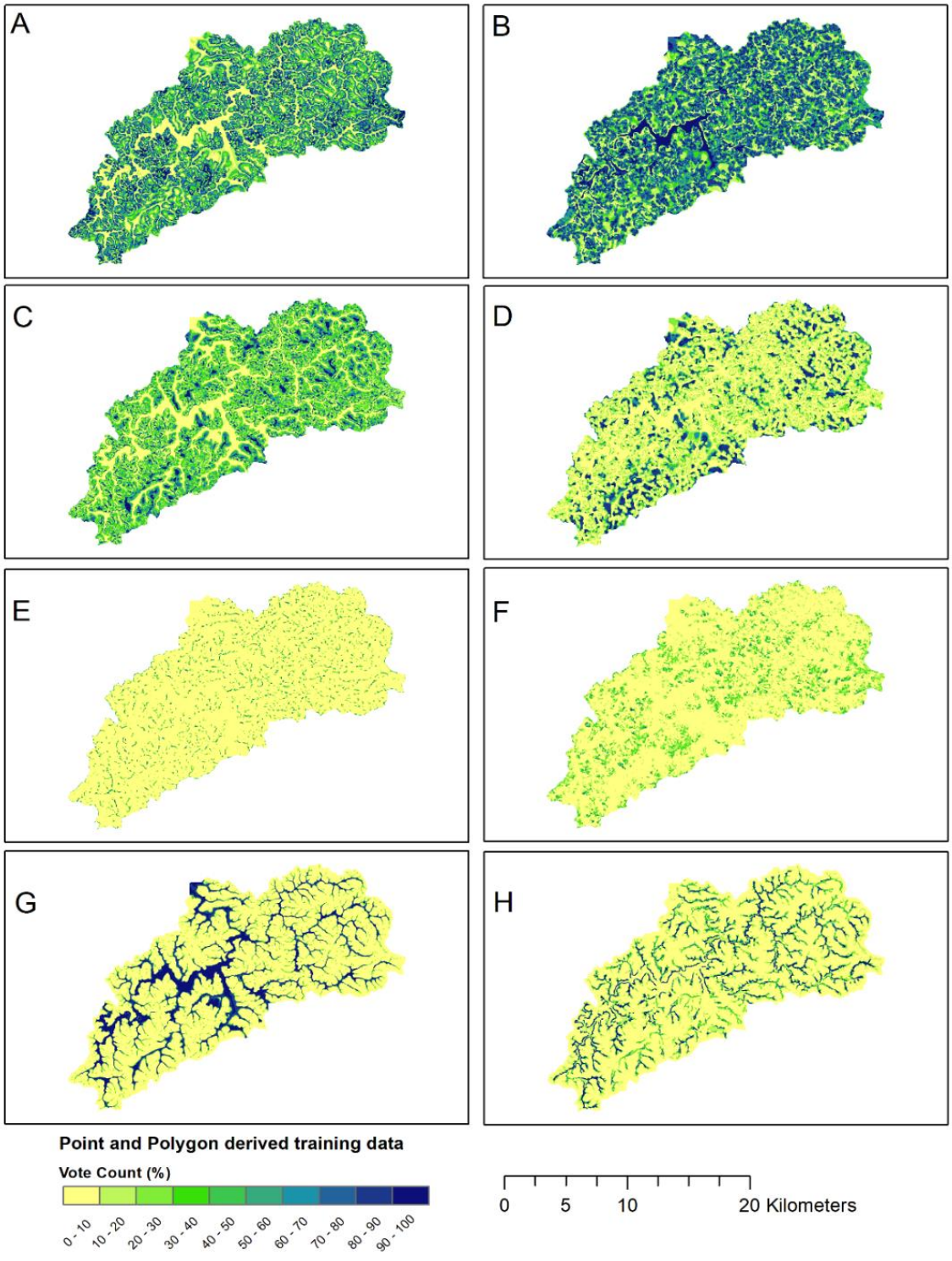
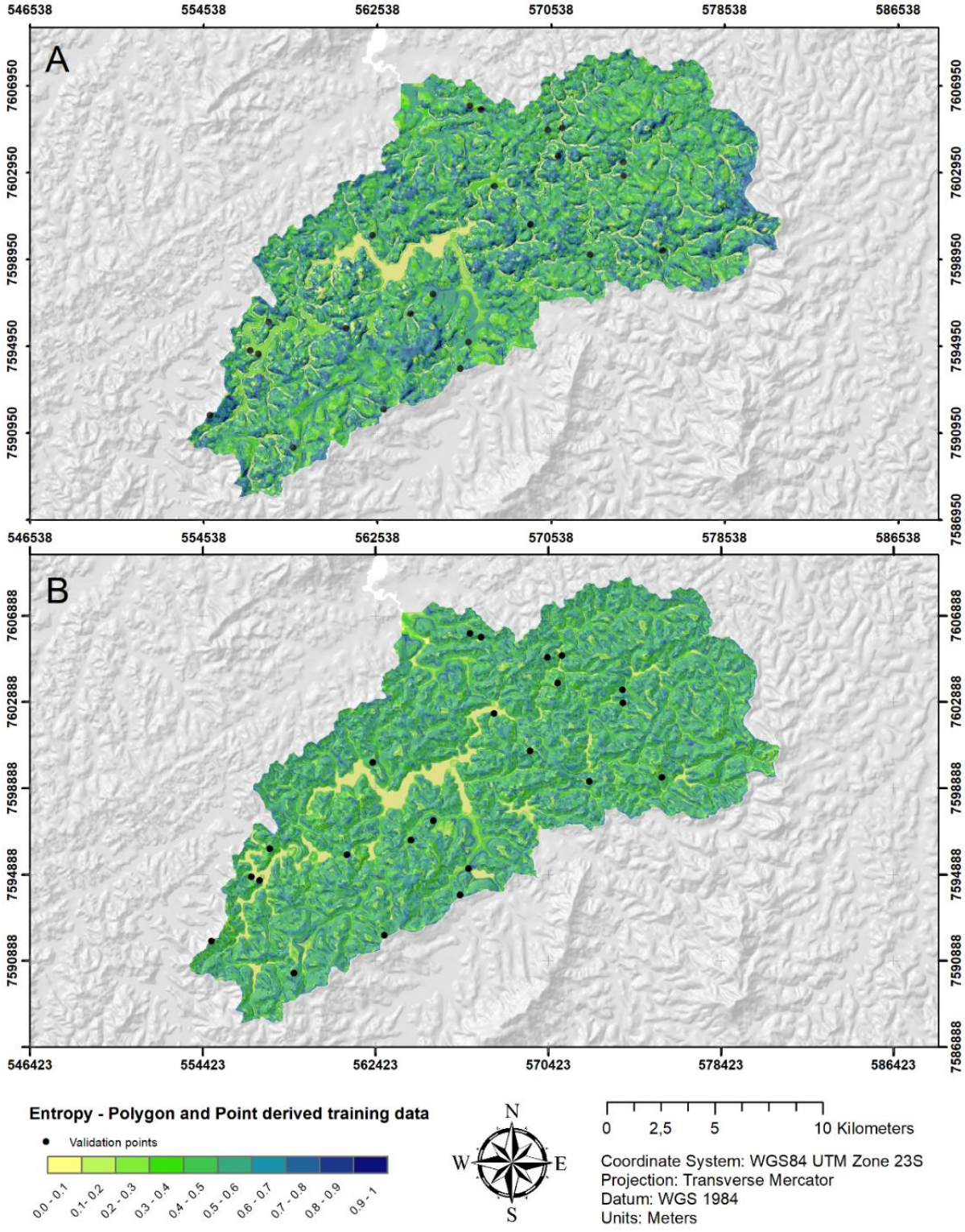


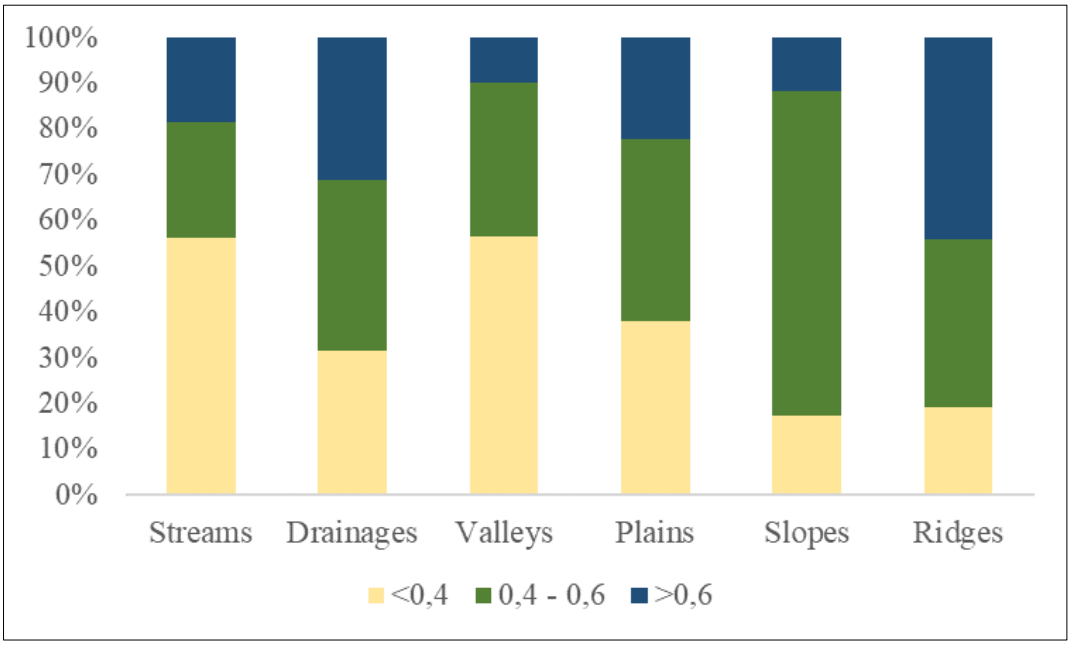
Figure 2.9 - Uncertainty surface based on Random Forest model using Point and Polygon-derived training data produced at a 20m spatial resolution for the study area. (A) Entropy values for Polygon derived training data. (B) Entropy values for Point-derived training data.



Higher uncertainty was found in footslopes and convex ridges. The former is probably related to the common associations between Udepts and Hapludox in this region, which can generate "confusion" when discriminating the domains of each type of soil. Silva et al. (2016) reported an analogous condition studying a nearby area, where on similar landscape positions occurs both, Inceptisols and Oxisols. Inceptisols tends to occur in the upper third and sometimes in the inferior third of backslope in association with Oxisols (Curi et al., 1994). In relation to the convex ridges, higher values of uncertainty may be due to the difficulty in tell apart the domains of Hapludox, Acrudox, and occasionally Udepts. Such pattern of soil distribution is a common situation in the northeastern portion of the study area.

Most of the models presented low accuracy for Acrudox (Table 2.4) along with greater uncertainty, as observed in Figure 2.9A and B. This is mainly related to two factors: a) the low density of the training and validation datasets, since such soil type has low geographical expression in this region when compared with the others. This relative unbalance tends to favor the majority classes within the training dataset (He and Garcia, 2009). In other words, classes over-represented in the training dataset may dominate the classification by the model (Millard and Richardson, 2015). This natural unbalance is common when dealing with soil type distribution. In the reference area (Vista Bela Creek Watershed), Acrudox corresponds to only 12% of the total area. The same was observed during the field work for the extrapolation area, where unlike Hapludox, the Acrudox do not occur in large contiguous areas, but rather, in transitions between Hapludox areas. b) Even though the relief explained most of spatial variability of soil types, it seems that specifically for Acrudox, it is mainly driven by parent material instead of relief solely. Terrain indexes do not efficiently tell apart the Acrudox and Hapludox areas, since both occur in similar portions in the landscape. Since digital soil mapping techniques are updatable (Hengl et al., 2014), the availability of data in the future related to parent material in the same scale of this study, could provide improvements of spatial prediction accuracy.

Figure 2.10 – Overall relative frequency distribution of the entropy values related to the TPI based landforms classification.



Streams – canyons and deeply incised streams; Drainages - midslope drainages, shallow valleys, upland drainages and headwaters; Valleys - U-shape valleys; Slopes - open slopes and upper slopes; Ridges - local ridges/hills in valleys, midslope ridges, small hills in plains and high ridges

2.4 CONCLUSIONS

By executing the Buffer, the point-derived data performed better results compared to Polygon-derived models. Excluding the Buffer and PCA Buffer datasets, there were no great differences between the accuracy of the models. The reduction of variables was able, in a general way, to improve the accuracy of the predicted maps of soil types, the same for training data selection. The best result was obtained by identifying the principal components of the Buffer dataset, and reducing the size of predictors ensemble with the PCA. Although the uncertainty was relatively similar for both Buffer-Point and Polygon derived models, the one derived from Polygons seems to have inserted more noise to the models, as observed by the inconsistences in the spatial prediction of soil types. The natural unbalance in the dataset training related to those soil types with smaller geographical expression could under-represent its spatial prediction from Random Forest and increase the uncertainty over some types, such as Acrudox in the region of the study.

Even though Random Forest has been considered a robust spatial predictor model in Soil Science, it was very clear its sensitivity to different strategies of selecting training dataset.

Effort was necessary to find the best training dataset for achieving suitable accuracy of spatial prediction. To identify a specific dataset in this study seems to be better than a great number of variables or a large size of training data. And so, the efforts here allowed the accurate acquisition (83% for overall accuracy and 0.738 for Kappa index) of a mapped area (2,719 ha) 15.5 times greater than the reference area (175 ha), up to the second hierarchica level according to Soil Taxonomy, at low cost by taking advantage of soil legacy data.

REFERENCES

Arruda, G.P de.; Demattê, J.A.M.; Chagas, C.S.; Fiorio, P.R.; Souza, A.B.; Fongaro, C.T. 2016. Digital soil mapping using reference area and artificial neural networks. Scientia Agricola 73: 266-273.

Behrens, T.; Zhu, A.-X.; Schmidt, K.; Scholten, T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155: 175–185.

Breiman, L. 2001. Random forests. Mach. Learn 45: 5-32.

Chagas, C. da S.; Carvalho Junior, W.; Bhering, S.B.; Calderano Filho, B. 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139: 232-240.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37: 35–46.

Congalton, R.G., Green, K. 2008. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. 2nd ed. CRC Press, Boca Raton, FL, USA.

Curi, N.; Chagas, C. da S.; Giarola,, N.F.B. 1994. Distinction of agricultural environments and soil-pasture relationship in the Mantiqueira fields. = Distinção de ambientes agrícolas e relação solo-pastagens nos campos da Mantiqueira. p. 21-43. In: Carvalho, M.M.; Evangelista, A.R.; Curi, N., eds. Desenvolvimento de Pastagens na Zona Fisiográfica Campos das Vertentes, MG. Embrapa, Brazil (in Portuguese).

Cutler, A.; Cutler, D.R; Stevens, J.R. 2012. Random Forests. p. 157-175. In: Zhang, C; Ma, Y.Q., eds. Ensemble Machine Learning: Methods and Applications. Springer, New York, NY, USA.

Deng, C., Wu, C., 2013. The use of single-date MODIS imagery for estimating largescale urban impervious surface fraction with spectral mixture analysis and machine learning techniques. ISPRS J. Photogramm. Remote Sens. 86: 100–110.

Favrot, J.C. 1989. A strategy for large scale soil mapping: the reference areas method. Science du Sol 27: 351-368 (in French, with abstract in English).

Freeman T.G. 1991. Calculating catchment area with divergent flow based on a regular grid. Computers & Geosciences 17: 413-422.

Gallant, J.C.; Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39: 1347–1359.

Giasson, E.; ten Caten, A.; Bagatini, T.; Bonfatti, B. 2015 Instance selection in digital soil mapping: A study case in Rio Grande do Sul, Brazil. Ciência Rural 45: 1592-1598.

Grinand, C.; Arrouays, D.; Laroche, B.; Martin, M.P. 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143: 180-190.

Guisan A.; Weiss, S.B.; Weiss, A.D. 1999. GLM versus CCA sapatial modeling of plant species distribution. Plant Ecology 143: 107-122.

Guyon, I.; Elisseeff, A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3: 1157-1182.
Hammond, T.; Verbyla, 1996. D. Optimistic bias in classification accuracy assessment. International Journal of Remote Sensing 7: 1261–1266.

Hastie, T.; Tibshirani, R.; Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY, USA.

He, H.; Garcia, E. 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21: 1263–1284.

Hengl, T.; Heuvelink, G.B.M.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; Shepherd, K.; Sila, A.; MacMillan, R.A.; Mendes de Jesus, J; Tamene, L.; Tondoh, J.E. 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. Plos One 10: 1-26.

Hengl, T; Jesus. J.M.; MacMillan, R.A.; Batjes, N.H.; Heuvelink, G.B.M.; Ribeiro, R.; Samuel-Rosa, A.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; Gonzalez, M.R. 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. Plos One: 9, 1-8.

Heung, B.; Bulmer, C.E.; Schmidt, M.G. 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214-215: 141-154.

Heung, B.; Ho, H.C; Zhang, J.; Knudby, A.; Bulmer, C.E.; Schmidt, M.G. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265: 62-77.

Heung, B.; Hodúl, M.; Schmidt, M.G. 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma 290: 51-68.

Iwahashi, J.; Pike, R.J. 2007. Automated Classifications of Topography from DEMs by an Unsupervised Nested Means Algorithm and a Three-Part Geometric Signature. Geomorphology 86: 409-440.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2013. An Introduction to Statistical Learning - with Applications in R. Springer, New York, NY, USA.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill Book Co., Inc., New York.

Kempen, B.; Brus, D.J.; Stoorvogel, J.J.; Heuvelink, G.B.M.; Vries de, F. 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal 76: 2097-211.

Kuang, B.; Mouazen, A.M. 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. European Journal of Soil Science 62: 629–636.

Lagacherie, P.; McBratney, A.B. 2007. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. p. 3-22. In: Lagacherie, P.; McBratney, A.B.; Voltz, M., eds. Developments in Soil Science, volume 31. Elsevier B.V.

Landis, J.R.; Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159–174.
McBratney, A.B.; Santos, M.L.M.; Minasny, B. 2003. On digital soil mapping. Geoderma 117: 3–52.

McBratney, A.B.; Santos, M.L.M.; Minasny, B. 2003. On digital soil mapping. Geoderma 117: 3–52.

McKay J.; Grunwald S.; Shi X.; Long R. 2010. Evaluation of the Transferability of a Knowledge-Based Soil-Landscape Model. In: Boettinger J.L.; Howell D.W.; Moore A.C.; Hartemink A.E.; Kienast-Brown S. (eds) Digital Soil Mapping. Progress in Soil Science, vol 2. Springer, Dordrecht.

Mendonça-Santos M.L.; H.G. dos Santos. 2007. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. p. 39-54. In: Lagacherie, P.; McBratney, A.B.; Voltz, M., eds. Developments in Soil Science. Elsevier, Amsterdam, Netherlands.

Menezes, M.D.; Curi, N.; Marques, J.J.; Mello de, C.R.; Araújo de, A.R. 2009. Pedologic survey and geographic information system for evaluation of land use within a small watershed, Minas Gerais State, Brazil. Ciência e Agrotecnologia 33:1544-1553 (in Portuguese, with abstract in English).

Millard, K.; Richardson, M. 2013. Wetland mapping with LiDAR derivatives, SAR polarimetric decompositions, and LiDAR-SAR fusion using a RF classifier. Canadian Journal of Remote Sensing 39: 290-307.

Millard, K.; Richardson, M. 2015. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. Remote Sensing 7: 8489-8515.

Novakovic, J.; Strbac, P.; Bulatovic, D. 2011. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research 21: 119-135.

Pal, M.; Mather, P. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment 86: 554–565.
Pelegrino M.H.P.; Silva, S.H.G.; Menezes de, M.D.; Silva da, E.; Owens, P.R.; Curi, N. 2016. Mapping soils in two watersheds using legacy data and extrapolation for similar surrounding areas. Ciência e Agrotecnologia 40: 534-546.

Rudiyanto; Minasny, B.; Setiawan, B.I.; Arif, C.; Saptomo, S.K.; Chadirin, Y. 2016. Digital mapping for cost-effective and accurate prediction of the depth and carbono stocks in Indonesian peatlands. Geoderma 272: 20-31.

Samuel-Rosa, A.; Heuvelink, G. B. M.; Vasques, G. M.; Anjos, L. H. C. 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244: 214-227.

Scarpone, C.; Schmidt, M. G.; Bulmer, C. E.; Knudby, A. 2017. Semi-automated classification of exposed bedrock cover in British Columbia's Southern Mountains using a Random Forest approach. Geomorphology 285: 214-224.

Silva, S.H.G.; Menezes, M.D.; Owens, P.R.; Curi, N. 2016. Retrieving pedologist's mental model from existing soil map and comparing datamining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267: 65-77.

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

Souza, E.; Fernandes Filho, E.I.; Schaefer, C.E.G.R.; Batjes, N.H.; Santos, G.R.; Pontes, L.M. 2016. Pedotransefer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Scientia Agricola 73: 525-534.

Strobl, C.; Malley, J.; Tutz, G. 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and RF. Psychol. Method 14: 323–348.

Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R.P.; Feuston, B.P. 2003. Random Forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences 43: 1947-1958.

ten Caten, A.; Dalmolin, R.S.D.; Ruiz, L.F.C. 2012. Digital soil mapping: strategy for data pre-processing. Revista Brasileira de Ciência do Solo 36: 1083-1091.

Voltz, M.P.; Lagacherie, P.; Louchart, X. 1997. Predicting soil properties over a region using sample information from a mapped reference area. European Journal of Soil Science 48: 19-30.

Weiss, A.D. Topographic position and landforms analysis. ESRI Users Conference, San Diego, CA, USA, 2001.

Yu, L.; Fu, H.; Wu, B.; Clinton, N.; Gong, P. 2016. Exploring the potential role of feature selection in global land-cover mapping. International Journal of Remote Sensing 37: 5491-5504.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography Earth Surface Processes and Land 12: 47-56.

Zhu, A. X. 1997. A similarity model for representing soil spatial information. Geoderma, 77: 217-242.

3. ARTICLE 2. Transferability, accuracy, and uncertainty assessment of different Knowledge-based approaches for soil type mapping

* Article prepared according to the rules of Catena

ABSTRACT

Soil legacy data is an important source of soil information, especially when dealing with limited resources. In countries with sparse areas and few financial resources, such as Brazil, it represents an economical alternative to obtaining soil spatial information. By retrieving the soil scientist's knowledge, it can be used as guidance for knowledge based digital soil mapping approaches. In this sense, this work aimed to evaluate Rule-Based Reasoning (RBR) and Case-Based Reasoning (CBR) knowledge-based approaches in order to predict soil types up to the second categorical level (U.S Soil Taxonomy) in a non-sampled area, by retrieving and then extrapolating the information of a detailed soil legacy map, used as a reference area. The study was carried out in Minas Gerais state, Southeastern Brazil. The methodology contains three main processes: i) the knowledge acquisition; ii) the soil inference procedures; iii) accuracy and uncertainty assessment. For the validation, 23 independent samples were chosen by means of the Regional Random method, and the accuracy was assessed by Kappa index, Overall Accuracy, Users', and Producers' Accuracy. The uncertainty was evaluated through ignorance of individuals (entropy) and exaggeration of members. A total of 24 inference models were obtained with the CBR approach, whose the best model presented 61% of overall accuracy and a Kappa index of 0.518. The RBR approach had a greater accuracy than the other models, accounting for 82% of overall accuracy and 0.749 for Kappa index. The efforts made it possible the accurate acquisition of a mapped area 15.5 times larger than the reference area with low cost.

Keywords: Digital soil mapping; soil survey; legacy data.

3.1 INTRODUCTION

Soil mapping is an inference process based on well-structured paradigms, which states that the distribution of soils in the landscape is predictable once the soil-environment relationships are known (Hudson, 1992; Jenny, 1941). Traditionally, soil mapping is related to a mental process, fully depending on the expertise of the soil scientist to decode the soil-environment relationships and delineate the spatial representation of the soil distribution over the landscape. Although very precious, some problems are usually reported. The facts of being time-consuming and its high cost are pointed out as the major factor for the worldwide lack of soil spatial data information in greater detail (McBratney et al., 2003; Kempen et al., 2012). The subjectivity, the inconsistency and the difficulty on representing the continuous variability of soils are pointed out by Shi et al. (2004) as inherent of manual polygon-based mapping process. Another problem is related to the knowledge transmission. "Tacit Knowledge is non-transferable without the exchange of key personnel and all the systems that support them" (Nonaka et al., 2000), therefore, the knowledge transferability often requires collaborative experiences, participation and doing (Foos et al., 2006).

Such problems have motivated many researchers to improve soil mapping techniques, by means of quantitatively descriptions of soil-landscape relationships on digital environments (Skidmore et al., 1991; Zhu et al., 1997a; Dobos et al., 2000; Shi et al., 2004; Qi et al., 2006; Demattê et al., 2015; Silva et al., 2016; Akumo et al., 2017). In general, we can distinguish two groups of techniques being taken in Digital Soil Mapping (DSM); data-driven and knowledge-driven processes. The former is based on statistics, geostatistics, machine learning, and data mining techniques. It is more quantitative and automatic, but often requiring dense sampling schemes, which may be costly, if no data is readily available. The other approach takes advantage of soil scientist's knowledge, combining it with Geographical Information System techniques, trying to reduce the subjectivity problems related to manual processes of conventional soil mapping frameworks, such as inconsistency, and loss of knowledge due to personnel change over time (Shi et al., 2004; Shi et al., 2009).

In Brazil, there is a need for more detailed soil maps contrasting to increasing funding limitations (Giasson, et al., 2006). Other authors have argued that Knowledge systems represents a feasible and economical alternative on making good use of soil scientists' knowledge and soil legacy data (Hudson, 1992; Shi et al., 2004; Silva et al., 2016). Available soil maps synthetize the soil scientist mental model for the soil distribution over the landscape

for a given site (Bui, 2004), serving as reference areas (Favrot, 1989) to soil survey for unmapped areas with similar environmental characteristics (Bui and Moran, 2003).

In order to overcome the necessity of representing the continuity character of the soil distribution, many researchers focused on exploring knowledge systems under fuzzy logic concepts (McBratney et al., 2003). The option for using fuzzy logic attempts to consider the uncertainty in predictions employing a partial membership conception, and the non-linearity of soil-landscape relationships to describe and model it (Zhu et al., 2010; Menezes et al., 2013).

One of the first efforts on developing a routine for soil mapping practice under these concepts was the SoLIM experiments, leaded by Zhu and colleagues (Zhu and Band, 1994; Zhu et al., 1996; 1997a; Zhu et al., 2001). It employs a soil membership vector (Zhu, 1997a) to measure the similarity between a soil at a given location (i,j) to a soil class (k). It is a Rule-Based Reasoning approach (RBR), where the soil scientist employs his/her knowledge to establish rules, which describes the relationship between the soil types and environmental variables (e.g. slope, parent material, elevation, etc). However, Shi et al. (2004) reported that, the need of explicit knowledge and the variable independence assumption limited the experience on using the current SoLIM. In an effort to overcome those limitations, the authors purposed the use of a Case-Based Reasoning (CBR) approach as an alternative to the RBR. It was based on two main assumptions: "cases are capable of representing domain expert's knowledge" and "a new problem can be solved by referring to similar cases" (Shi et al., 2004). Instead of creating a set of rules (parametrical space), which is not always simple, scientists identify locations in geographic space to represent the knowledge of the soil-landscape relationship. It is called tacit point, and corresponds to a case. Similarly, to RBR, the objective is to drive for every location the fuzzy membership values of all soils that occurs in the area.

Shi et al. (2009) introduced a software package named Soil Inference Engine (SIE) to integrate both RBR and CBR. In addition, the SIE has a Knowledge Discovery module, which enables the scientist to explore the knowledge implicit on soil maps. There is also the possibility to explore the uncertainty of predictions, through the entropy and exaggeration of individual indexes. It permits to assess the behavior of the inference process and identify the complexity of the soil-landscape relationships. This is particularly relevant on tropical countries, where it is common the occurrence of polygenetic soils. Given the global increase demand for more detailed maps, the understanding of this aspect regarding landscape is important for generating maps with greater accuracy and lower cost.

This paper presents the use of two knowledge-based approaches founded on fuzzy logic for soil mapping at a non-sampled area. The main objective is to evaluate the efficiency of these approaches on predicting soil types up to the second categorical level (classified according to

U.S. Soil Taxonomy) for a non-sampled area with similar environmental characteristics of the reference area. It uses the knowledge retrieved from a reference area on GIS environment along with field expertise. For this, the ArcSIE (Soil Inference Engine) was applied. The methodology contains three main processes: i) the rule-based and case-based knowledge acquisition; ii) and the soil inference proceedings; iii) validation and uncertainty assessment.

3.2 MATERIAL AND METHODS

3.2.1 Study Area

Two areas were used in this study. The digitally mapped area A1 and the reference area A2. The comparison of their physical characteristics is given in Table 3.1. The reference area A2 (Figure 3.1A) (Vista Bela creek watershed) was mapped by experienced soil scientists, along with intensive fieldwork (total of 53 soil profiles). The map was produced in a detailed scale (1:10,000) composed by simple mapping units (Menezes et al., 2009). Table 3.2, soils are referred to as classified by the Soil Taxonomy (Soil Survey Staff, 2014), and their respective geographical expressions at the reference area (Figure 3.1A).

Table 3.1 - Study area comparison

	Study area A1	Study area A2
Coordinates	553781 and 581138 mW, 7598766 and 7597100 mS fuse 23K, datum WGS 1984	559868 and 561536 mW 7598766 and 7597100 mS, fuse 23K, datum WGS 1984
Size (ha)	2,719 ha	175 ha
Geology	biotite gneiss (banded); biotite schist; gneiss; feldspar schist; phyllites; xystus and quartzite	biotite gneiss (banded); biotite schist and gneiss
Slope (%)	Min: 0 - Max:73 Mean:14 - Std. Dev.: 10	Min: 2.29 - Max: 39.73 Mean: 15,5 - Std. Dev.: 7,8
Elevation (Meters)	Min: 924 - Máx:1342 Mean:1042 - Std. Dev.: 56	Min: 960 - Máx:1068 Mean:1034 - Std. Dev.: 25
Mean annual temperature	20ºC	20ºC
Mean annual precipitation	1,450 mm	1,450 mm
Land use	Pasture	Pasture

Figure 3.1 - Study areas location. A) – Soil types at the reference area (A1) (Menezes et al., 2009); B) The spatial settings of mapping unit polygons. C) An overview of the area A2. D) Digitally mapped area (A1), reference area (A2) and the validation sampling regions.

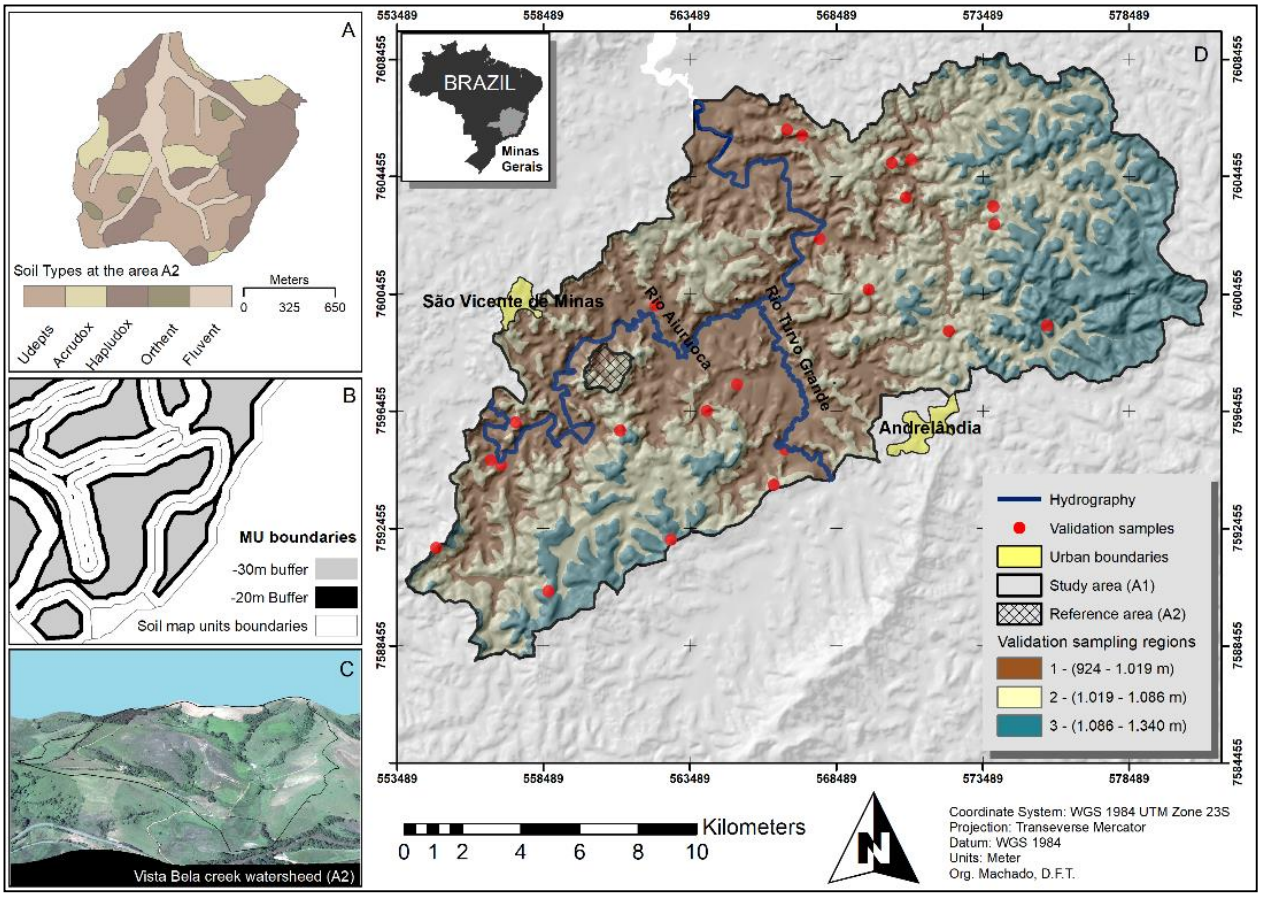


Table 3.2 - Mapping units identified in the study area.

Symbol	Soil classes	Area (ha)	%
Hx	Hapludox	61.2	35
Ax	Acrudox	21.3	12.2
Ut	Udept	61.3	35
Ft	Fluvent	27.2	15.5
Ot	Orthent	4	2.3
Total		175	100

Source: Menezes et al. (2009)

3.2.2 Preparing the Environmental Database

For the present study, a digital cartographic base was developed in GIS environment. Firstly, it was generated a Digital Elevation Model (DEM) with 20 m of spatial resolution, derived from contour lines freely available from IBGE (Brazilian Institute of Geography and Statistic) with a 1:50,000 scale and equidistance of 20 m. A hydrologic consistent DEM was

generated in ArcGIS 10.1 software (ESRI) through the Topo To Raster function. The DEM was used to calculate 14 topographic indexes (TI) using the SAGA GIS software (SAGA Development Team version 3.0), which are: catchment slope, convergence index, plan-curvature, profile curvature, LS-factor, multiresolution index of ridge top flatness (MRRTF), slope, SAGA wetness index (SWI), topographic position index (TPI), texture, terrain classification index for lowlands (TCI LOW), upslope curvature, valley depth and vertical distance to channel network and TPI based landforms. The topographic variables were selected due to their capacity to express variations of morphometric and hydrological characteristics at local and landscape scale, indicating changes in soil-forming factors (Giasson et al., 2015). For the reference area, a soil type map of Vista Bela Creek Watershed on shapefile format was used (Figure 3.1A).

3.2.3 Soil modeling environment - ArcSIE

The ArcSIE (SIE stands for Soil Inference Engine) version 10 is a toolbox extension of ArcGIS 10.1. It is an expert knowledge-based inference tool, supported by fuzzy logic. The soil mapping proceedings is based on the soil-environment model: S= *f* (E), which states that the soil information (S) can be derived from the information about the soil formative environment (E) (Shi et al., 2004), as well known as the soil forming factors (Jenny, 1941), in addition to other information about soils characteristics or attributes (McBratney et al., 2003). There are two inference methods of establishing the relation between the soil and its environment (*f*), namely: RBR, in which the inference is based on rules from direct specifications of soil surveyor and CBR, when the inference is based on cases, the knowledge at a specific location is represented by a point, line, polygon, and pixels defined in a geographic space (Shi et al., 2009). In both cases, the inference procedure consists on derive fuzzy membership functions, which describes the relationship between an environmental feature and the optimality values for a soil type (Shi et al., 2004). A more detailed review about these methods can be found at Menezes et al. (2013) and Zhu et al. (2010).

The ArcSIE also provides a data mining tool named Knowledge Discoverer (KD). In this case, data mining stands for a semi-automated process, and it was used to recognize patterns and thresholds for the purposes of prediction. It works by overlapping vector features over raster layers to generate mathematical functions, represented by curves. These curves are considered a representation of the knowledge about the relationship between the environmental feature (raster layers) and the soil, represented by a vector feature. For a polygon, the cells that are

enclosed by the polygon will be used to calculate statistics and build the curve. The KD is a practical way to convert a case-based into a rule-based, and then apply this "discovered" knowledge to a different mapping area if convenient (Shi et al., 2009).

3.2.4 The soil inference on ArcSIE

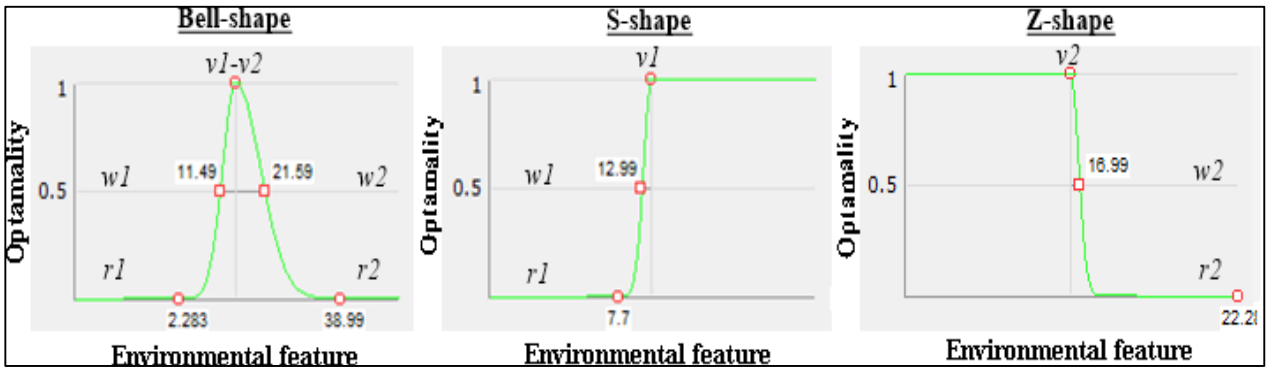
In this study, the inference process was done based on the global knowledge approach (Shi et al., 2009), which are divided on three phases, described by the three functions, *E*, *P*, and *T*, in the following equation (Shi et al., 2004):

$$S_{ij,k} = \prod_{g=1}^{n} T_k \left\{ \prod_{a=1}^{m} P_g \left[E_{g,a}(Z_{ij,a}, Z_{g,a}) \right] \right\}$$

where $S_{ij,k}$ is the fuzzy membership value at a location (*i*,*j*) for a soil *k*. The *m* is the number of environmental features used in the inference. The n is the number of instances for soil type *k*. $Z_{g,a}$ is the most optimal range given by rule or case g, defining the most favorable condition of feature a for soil *k*; $Z_{ij,a}$ is the value of the a^{th} environmental feature at location (*i*,*j*). The *E* function evaluates the optimality value at the environmental features level. The closer $Z_{ij,a}$ is to the $Z_{g,a}$ range, the greater the optimality value assigned by E. In RBR, $Z_{g,a}$ is directly defined by the scientist, whereas in CBR, it is automatically stated based on the case location and the related environmental features.

ArcSIE provides five choices for *E*, namely *nominal*, *ordinal*, *cyclic*, *continuous* and *raw values*. In this study, only the *Continuous* function (Figure 3.2) was applied since it fits with continuous distribution of raster maps used. It is based on Gaussian curves to fine-tune the optimality value. The *Continuous E* function, are defined by the parameters *v*, *w* and *r*, and adjusted in the Inference module on ArcSIE. The *v1* and *v2* are, respectively, the lower and upper limits of the most optimal range of a certain environment feature for a soil type. The *w1* and *w2* defines how optimality will change as environmental conditions deviate from the typical conditions (Zhu et al., 2010). It corresponds to which value of the current attribute feature should present 50% of the optimality value. The *r1* and *r2* are used to control the flatness of the top and the steepness of the side parts of the curve.

Figure 3.2 - The three types of Continuous Function implemented on ArcSIE



Function *P* is responsible for the interaction of the environmental features, resulting in optimality value for the whole instance. ArcSIE provides three methods for the *P* function: *Limiting-Factor*, *Weighted-Average*, and *Multiplication*. Once it is possible that a soil type presents more than one instance or case, *T* function is responsible for integrating their predicted values to return the final predicted values (Shi et al., 2009).

3.2.5 Knowledge discovery and inference models

The prediction of soil classes was performed based on the mentioned environmental variables. In this study, two procedures were tested: i) RBR approach; ii) CBR approach.

3.2.5.1 RBR approach

The RBR is based on attribute rules that can represent and formalize the soil scientist's mental-model for the soil-landscape relationship. Attribute rules are created and fully defined by the user. For the development of the attribute rules and the inference for the non-sampled area, the following procedures were adopted:

1. The first step was to retrieve the soil-landscape model for the study area by revisiting the expertise accumulated in different studies in the reference area and surroundings, which preserves a potential knowledge that could be applied for the study area. In addition, fieldwork was carried out to confirm and update the soil-landscape model.

2. The terrain indexes representing the environmental characteristics were developed and managed in GIS environment. Among them, the TI that better reproduced the relief characteristics associated with the respective soil classes comprised of the mental model were chosen.

3. Once the mental-model and the set of variables were defined, the next procedure was to create the Rulebase. The Rulebase is composed of one or more soil types, which in turn have one or more instances, each one describing a unique environmental configuration, represented by a set of fuzzy membership curves. The adjustments were done by exhaustive evaluations, modeling and comprehending the soil-landscape relationships. For this task, the Knowledge Discoverer was applied. The KD uses curves to expose soil-environment relationships. There are two options of curves: two-side Gaussian optimality curve and kernel-smoothed frequency curve. The curves were generated based on the soil mapping units of the reference area. These proceedings aimed to expose the mental-model for the soil occurrence over the landscape to a formal set of rules.

4. Based on the soil type instances, it was generated a series of fuzzy membership maps in raster format, one for each soil type of the current Rulebase. Since it is a modelling task, the iterative process is done until the membership maps match with the operator expectations (Shi et al., 2004).

5. The hardened map was generated by confronting each fuzzy membership map. For that, ArcSIE assigns to each pixel the soil type with the highest fuzzy membership value. The output is a raster layer that can be then vectorized.

3.2.5.2 CBR approach

In the CBR approach, the knowledge of local soils is represented by a collection of cases, organized into one or more case lists, each one representing a soil type. A case is the same as an instance plus spatial information. Two assumptions are made on CBR. The first is that cases may represent the expert's knowledge. The second is that a new problem can be solved by referring to similar cases already solved (Shi et al., 2004). In this study, the cases were obtained by the mapping unit polygons from the legacy soil survey (reference area – A2). Each polygon represents a case that contains the information of the geographic space, the parametrical space (environmental features), and the solution space (taxonomic space). The construction of the case lists and the inference procedure for the non-sampled area (study area A1) are described below, and summarized in Figure 3.3:

1. The data layers, including the soil map of the reference area and the TI, were developed and stored on GIS environment. For the Knowledge discovery, it was defined two ensembles of topographic indexes. The first (*ei-TI*) corresponds to the same used for previous procedure (RBR). For the second one (*rf-TI*) the Random Forest (RF) algorithm (Breiman,

2001; Liaw and Wiener, 2002) was applied on R software (R Development Core Team 2012, version 1.0.44) using the *randomForest* package to rank and them, select a reduced ensemble of variables. The training data comprehended the pixels enclosed by the mapping units of soil map from the reference area. The RF provides an index for measuring the importance of each variable for the model, called Mean Decrease in Accuracy (MDA). It evaluates a variable contribution for the model by measuring the increasing in error when randomly permuting a single predictor (TI) in the Out-of-Bag data (Breiman, 2001). More details can be found in Hastie et al., (2009) and Liaw and Wiener (2002). The indexes were selected based on the MDA rank, and the top six were selected. Thus, the *rf-TI* ensemble was composed by terrain surface texture, Saga wetness index, valley depth, vertical distance to channel network, longitudinal curvature, and terrain classification index for lowlands.

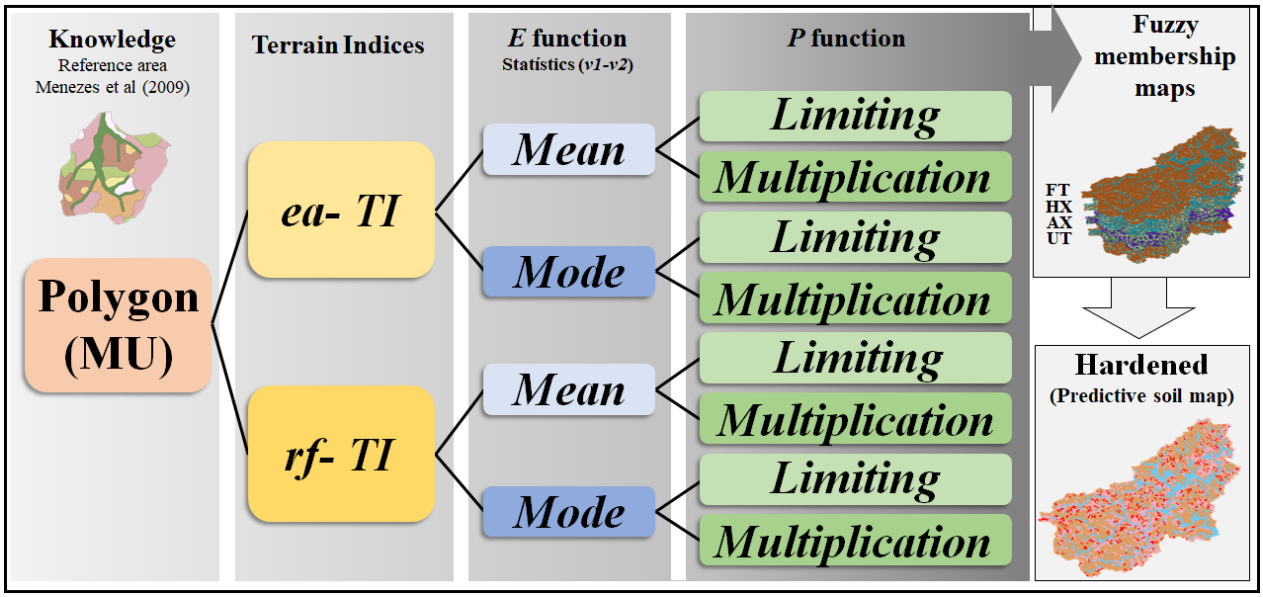
2. The soil type distribution in the landscape (legacy data from the soil map) was related to the TI, and the values were extracted from the maps in three different ways in KD, considering the spatial reference (Figure 3.1B): a) each mapping unit from the entire polygon (MUe); b) each mapping unit polygon excluding 20 m from the boundaries (MUP-20); c) each mapping unit polygon excluding 30 m from the boundaries (MUP-30). For b) and c), the transition zones (closer to the borders) were avoided since such areas might bring uncertainty for soil information (Ten Caten et al. (2012), Giasson et al. 2015).

3. For adjusting the models, the sampling scheme on KD was the No subsampling/averaging, where the original raster was used for statistics. For *E* function, it was applied the Continuous Bell-shaped (the optimality value decreases as the difference between the environmental feature value and the central values *v1* and *v2* increases). For the optimality values *v1* and *v2*, two statistics were tested: the mean and mode values of the TI extracted for each MU. For *w* and *r*, the default values were applied. For *P* function, multiplication and limiting factor were also tested.

4. Once extracted the values and developed the cases, the soil class fuzzy membership maps were generated for each interaction.

5. Finally, for each set of adjustments a hardened map was developed.

Figure 3.3 - Knowledge acquisition, membership functions adjustments and inference processes for the Case-based approach. MU – Mapping Unit; ea-TI – conjunct of terrain idexes used on RBR approach; rf-TI – conjunct of terrain idexes selected with the Random Forrest.



3.2.6 Assessment of spatial predictions

The assessment of spatial predictions was done by 23 soil profiles. The sampling sites were chosen by means of Regional Random method implemented in ArcSIE (version 10.3.101). The locations were randomly specified within three sampling regions, representing different altitude levels as shown in Figure 3.1. Aiming to assess the quality of the predictive maps, two indexes were calculated: Overall Accuracy (OA) and Kappa Index (KI). The overall accuracy is the sum of the main diagonal components of the confusion matrix divided by the total number of validation samples.

$$Overall\ Accuracy = \frac{\sum_{j=1}^{c} xjj}{N}$$

where: *xjj* is the number of correct samples and *N* is the total number of samples.

Kappa index is an agreement measure calculated taking into account the total number of samples, the number of soil classes and the correctly classified samples (Congalton and Green, 1999). The values may range from -1 (suggesting disagreement) to 1 (excellent agreement) (Landis and Koch, 1977).

$$Kappa = \frac{N\sum_{j=1}^{c} xjj - \sum_{j=1}^{c} xjixij}{N^2 - \sum_{j=1}^{c} xjixij}$$

where *xij* is the value in a row *i* and in a column *j*; *xji* is sum of values in line *i*; *xij* is the sum of values in column *j*; *N* is the total number of samples (points used for validation); and *c* is the number of soil classes.

The prediction uncertainty caused by the hardening process was evaluated by the ignorance uncertainty (entropy) and exaggeration of members, calculated as follows:

$$Entropy = \frac{-\sum_{h=1}^{n} \frac{S_k}{\sum_{k=1}^{n} S_k} \ln \frac{\left(\frac{S_k}{n}\right)}{\sum_{h=1}^{n} S_k}}{\ln(n)}$$

In the above equation, S_k is the fuzzy membership value of soil type *k* at a given location, and *n* is the total number of soil types. The entropy values range from 0 to 1, and the higher the entropy value at a location, the higher the uncertainty caused by the hardening process (Zhu, 1997b). The exaggeration was calculated as follows:

$$E_{ij} = 1 - S_{ij}^{a}$$

where E_{ij} is the estimated exaggeration uncertainty and S_{ij}^{a} is the similarity value of the instance for at each pixel to its correlate predicted category (a) (Zhu, 1997b).

3.3 RESULTS AND DISCUSSION

3.3.1 RBR approach: soil-landscape relationships

The study area is situated at the Andrelândia Plateau. The relief presents, in general, a homogeneous dissection pattern, with predominantly medium to coarse drainage densities. This pattern results in hills with convex to tabular tops and also convex slopes, interspersed by elongated crests. Below the 1,200 meters, the hill tops lose their sharp appearance, getting more softened (Neto, 2014; RADAM, 1983).

The most common soil types are Oxisols, Inceptisols and Entisols. Regarding the Oxisols, it is mainly related to hills with elongated and flat top, and slopes gentler than 20%. Inceptsols have a relevant geographical expression in the study area, and, as well as Orthents, are usually associated with steep slopes, and fine-textured terrain surfaces. The FT occur in flat areas around the drainage network, formed by sediments from floods deposition.

In the region comprising the study area, it is common the occurrence of different soil types in similar portions of the landscape. As observed by Curi et al. (1994) and Menezes et al. (2009), UT can occour associated to OT, and also in similar relief conditions to Oxisols.

Figure 3.4 illustrates a case by means of optimality curves. The curves represent values of slope and cross-sectional curvature extracted from the mapping units (UM) of the reference area, being Figure 3.4A and 4B, UT and Figure 3.4C and 4D, HX. The optimality values correspond to the modal value of the TI, for each MU. The curves in light green represent the TIs values, enclosed by their respective mapping units, highlighted in light blue in the maps. The more individualized is the optimality value of the TI for a certain soil type, the better that TI is in discriminating that soil type from others. In this case, it is noteworthy that both the slope and curvature values present similar optimal values for HX and UT. This is not a typical configuration for HX, which means that, formulating a Global Rule-based considering just the main soil-landscape configurations for each soil type could exclude scenarios like this one, affecting the prediction accuracy.

Since the soil-landscape relationships for a given soil type is not homogeneous, it was not possible to identify a single threshold-value based on the terrain indexes. Thus, the use of a single instance for each soil type in the Rule-base would be ineffective. In this sense, it was decided to identify specific soil-landscape configurations that could be translated into instances. This artifice was also used by Shi et al. (2009), which, for handling with local exceptions, to identify local cases was the most effective way to represent the knowledge.

Figure 3.4 - Two-side Gaussian optimality curves: A – Slope (Udept); B – Cross-sectional curvature (Udept); C – Slope (Hapludox); D – Cross-sectional curvature (Hapludox). Red-dashed lines depict the optimality value of the marked curves.

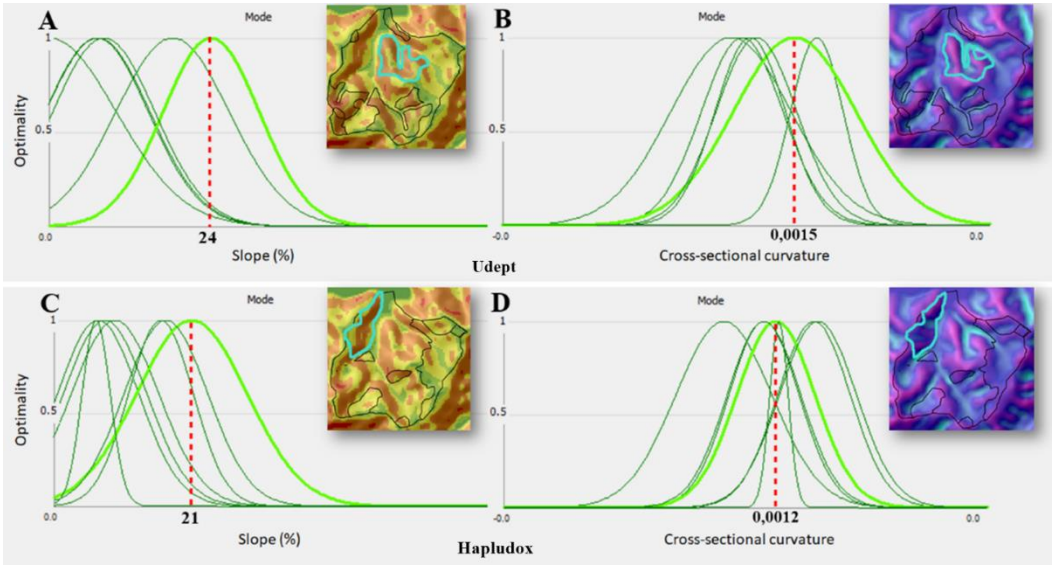
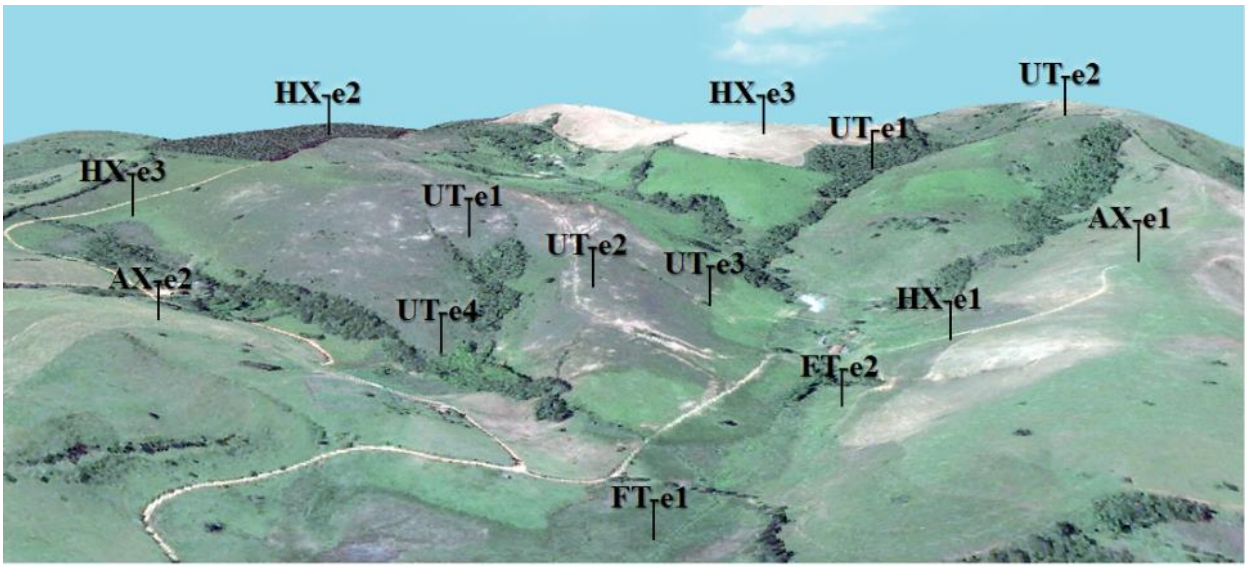


Figure 3.5, represents the places or situations taken as reference for the formulation of the Rule-base described in Table 3.3, which presents the instances for each soil type. The FT occur in the lowest portions of the landscape (Figure 3.5 FT-e1), presenting high SWI values (>6.4), and closer to the channel network (VDCN ≤ 1.5 m). The SWI indicates the places with higher tendency of water accumulation (Beven and Kirkby, 1979), while the VDCN calculates the vertical distance from each pixel to the channel network (Conrad et al., 2015). Once some upper-plains could present SWI values similar to the lower limit for FT, the VDCN was used as a limiting factor to avoid misleading classification (Figure 3.5 FT-e2). For UT, four instances were assigned. The first is related with concave and hilly slopes, usually associated with drainage headwaters (Figure 3.5 UT-e1) characteristics of fine-textured surfaces, which tend to correlate with erosional topography. The second are the narrow top areas, transitioning to steep slopes on fine-textured surfaces (Figure 3.5 UT-e2). The SWI (<3,6) was applied, once it gives information not only of flow accumulation, but also indirectly of curvatures and relative position (McKay et al., 2010). The third instance for UT describes locations where slope gradient and length leads to erosive surface conditions (Figure 3.5 UT-e3). The erosion tends to be intensified with the increasing of the slope and the length of the slopes, influencing the volume and speed of the water flow on the surface (Bertoni, Lonbardi Neto, 2012; Morgan, 2005), indicating erosive surfaces, and probably shallower soils. The last instance for UT describes the transition between floodplain and Inceptsols area related to convex slopes (Figure 3.5 UT-e4) and fine-texture surfaces.

Figure 3.5 - Schematic distribution of the soils in the reference area. Each label corresponds to the approximate location of a soil-landscape configuration described by the instances at the Rule-base.



According to Table 3.3, the Terrain Surface Texture was the main distinguishing parameter, mainly for UT and HX. The Texture is directed related with drainage density and changes in sign of slope aspect or curvature per unit area (Iwahashi and Pike, 2007). While the Inceptsols are associated to fine-texture surfaces, the Oxisols tend to occur on coarse-texture surfaces, related to areas of few dissection and planar slopes. The Texture allowed to distinguish HX sites with slope gradient higher than 20% (Figure 3.5 HX-e1) from Inceptisols domains, commonly associated with this slope class. The other instances for HX describe planar surfaces, gentle slope convex areas (Figure 3.5 HX-e2) and open concave hills with SWI between 4.4 and 6.4, if the VDCN is higher than 2 m (Figure 3.5 HX-e3).

Both HX and AX presented similar environmental conditions. However, AX commonly occupies the highest altitudes of the landscape, and presents a relatively better drainage than HX, favoring the formation of hematite, which is reflected by their color, redder than 2,5YR (Curi and Franzmeier, 1984). Thus, for the AX first instance (Figure 3.5 AXe-1), the MRRTF was applied. This TI was designed to identify higher flat areas, based on slope and position in landscape (Gallant and Dowling, 2003). In addition, AXe-2 describes a slope pattern between the flatter slopes of HX and the hilly slopes, characteristic of UT.

The OT was manually delineated. Its low geographical expression in the reference area, along with its intricate pattern of occurrence with Inceptsols may have hindered its individualization, and consequently, the knowledge transferability. To delineate the OT areas, it was identified some sites during the field work for establishing an identification key, by relating its occurrence on the field, as depicted in Figure 3.6.

Figure 3.6 - Identification Key for OT based on field verification.

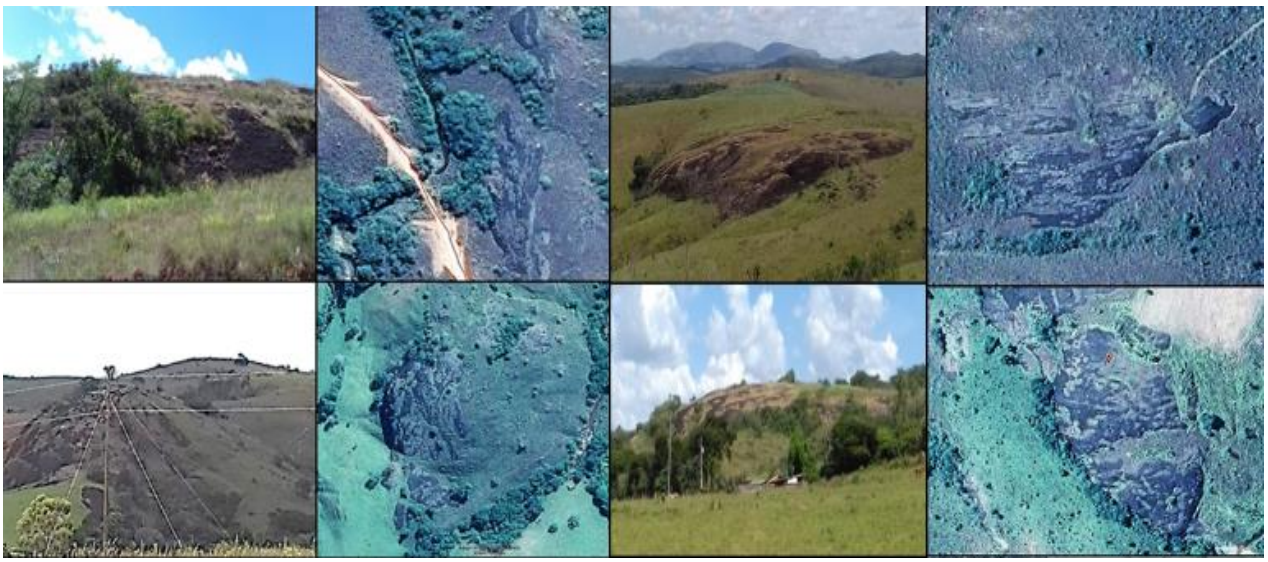


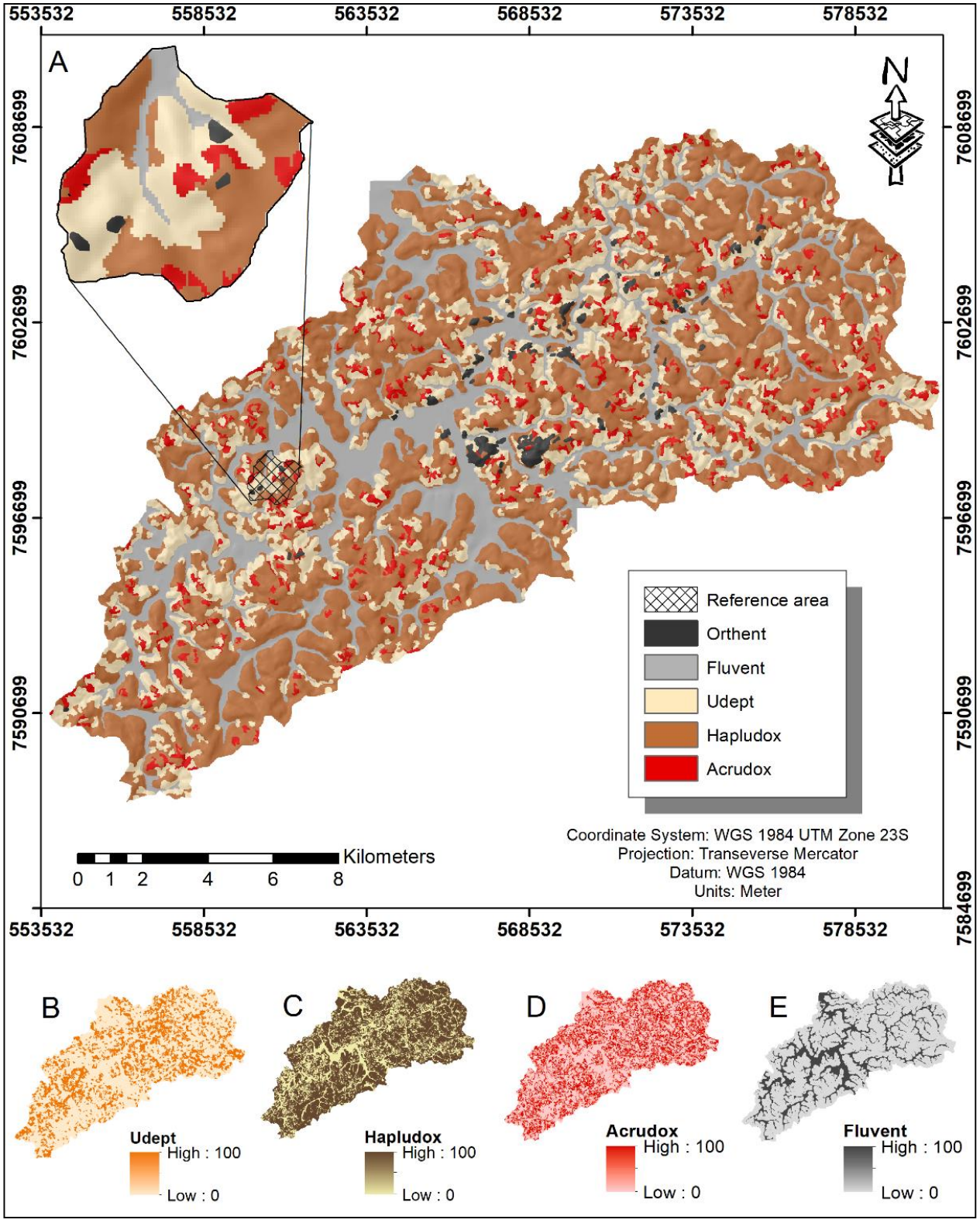
Table 3.3 - Rules for Udept (UT), Hapludox (HX), Acrudox (AX) and Fluvent (FT)

Soil Class	Instance	Terrain Index	*E function*					*P function*
			v1	w1	v2	w2	Curve shape	
UT	e1	Catch slope	0.2	0.014	-	-	S-shape	Limiting
		Texture	1.2	0.2	-	-	S-shape	
	e2	SWI	-	-	3.6	0.2	Z-shape	Limiting
		Texture	2.6	0.5	-	-	S-shape	
	e3	Texture	6.7	0.59	-	-	S-shape	Limiting
		LS-factor	11	0.2	-	-	S-shape	
	e4	Texture	2.4	0.59	-	-	S-shape	Limiting
		VDCN	1.8	0.5	20	2	Bell-shape	
HX	e1	Slope %	16	2	30	5	Bell-shape	Limiting
		Texture	-	-	1.5	0.5	Z-shape	
	e2	Catch slope	-	-	0.06	0.014	Z-shape	Limiting
		VDCN	2	0.2	-	-	S-shape	
	e3	SWI	4.46	0,35	6.5	0,35	Bell-shape	Limiting
		VDCN	2	0.2	-	-	S-shape	
AX	e1	MRRTF	0.38	0.16	-	-	S-shape	Limiting
		Texture	2.34	0.59	6.6	0.59	Bell-shape	
		VDCN	50	5.2	-	-	S-shape	
	e2	LS-factor	4	3	8	3	Bell-shape	Limiting
		Catch slope	0.136	0.014	0.18	0.014	Bell-shape	
FT	e1	SWI	6.5	0.3	-	-	S-shape	Limiting
	e2	VDCN	-	-	1.5	0.3	Z-shape	Limiting

Catch slope – Catchment slopes; Texture – Terrain surface texture; SWI – Saga wetness index; VDCN – Vertical distance to channel network; MRRTF – multiresolution index of ridge top flatness.

After the model process detailed above, a hardened soil type map (Figure 3.7A) was generated, based on fuzzy-membership maps of each soil type (Figure 3.7B, 3.7C, 3.7D and 3.7E). Darker colors represent higher memberships for that soil type.

Figure 3.7 - A – RBR predicted soil class map for the study areas and the fuzzy membership maps of Udept (B), Hapludox (C), Acrudox (D) and Fluvent (E).



3.3.2 RBR Accuracy Assessment

For the accuracy assessment, 23 soil profiles were described for validation purposes only. The inference procedure on ArcSIE resulted in an accurate soil type map for the extrapolation area. Of the 23 validation samples, 19 (83%) were correctly predicted, resulting in a Kappa index of 0.75, considered as a substantial classification, according to Landis and Koch (1977). The confusion matrix is presented in Table 3.4, as well the Producer´s and User´s accuracies.

Table 3.4 - Confusion matrix and uncertainty of the RBR predictions

Predictions	Observations				User's Accuracy	Mean Entropy	Mean Exaggeration
	Fluvent	Hapludox	Udept	Acrudox			
Fluvent	4	0	0	0	100	0.34	0.00
Hapludox	0	8	0	0	100	0.14	0.12
Udept	0	2	6	1	66	0.53	0.06
Acrudox	0	1	0	1	50	0.40	0.20
Producer's Accuracy	100	72	100	50			
Total number of samples = 23							
Overall accuracy = 82%							
Kappa index = 0,75							

Based on the mapping units from the legacy soil type map, the overall agreement with the predicted soil map was about 57%. Even though the modeling process could be time-consuming, RBR has become a promising technique where there is suitable knowledge about soil-landscape relationships. Silva et al. (2016) reported better results for a knowledge-driven approach based on fuzzy-logic using the ArcSIE compared with data-driven approach (decision trees) for soil type prediction and knowledge transferability. In addition, the results here demonstrate an adequate capacity of knowledge transfer for the expansion of mapped areas based on reference area. McKay et al. (2010) tested how well a soil prediction model developed for a relatively small area would work to a different one, with similar environmental conditions. They found that a knowledge-based model such as SIE can be developed and effectively transferred, once the environmental factor constraints are considered.

3.2.3 RBR Prediction Uncertainty

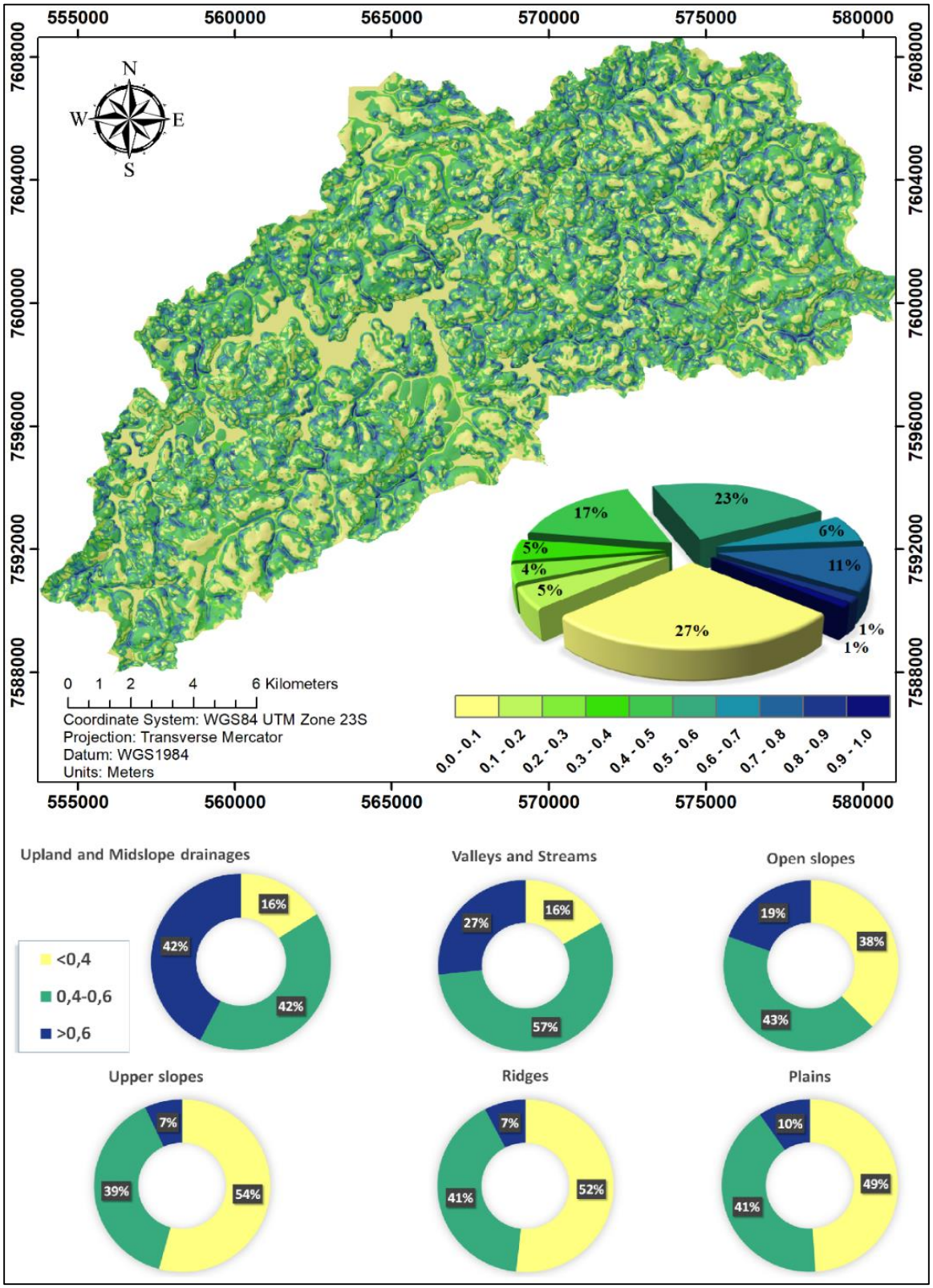
The ignorance of individuals can be estimated based on the entropy measure (Zhu,1997b; Qi and Zhu, 2011), which is related to the certainty on the inference results. For a location (*i, j*), the higher the membership for a particular class and lower the similarities to others, the smaller the uncertainty. Figure 3.8 shows the spatial distribution of the entropy values for RBR approach and its relative frequency distribution according to a landform classification. The entropy ranged from 0.0 to 1.0, with a mean of 0.38 and standard deviation of 0.27. The highest entropy values are associated with transitional zones between floodplains and footslopes. The soils occupying such transitional zones preserves similarities to their neighboring (based on their instances), enhancing the continuous aspect of soil distribution, which makes it difficult to fully apply a crisp limit for a certain soil type. Another observation for these lowland sites is the high percentage of medium entropy values (0.4-0.6). This is related to some similar values of SWI for both upper-concave areas and flood plains. Despite this, the inference results were quite accurate for lowland soil types. For open and upper slopes, the uncertainty is mostly related with two different soil-environment associations. The first is for coarse, gently to moderately slope surfaces (3-12%). Once these landforms are strongly related to Oxisols domains, the difficulty in discriminating the HX from AX, based only on morphometric variables would be the reason for the high uncertainty. The second is for slopes from 13% to 25%. These are common conditions for both Oxisols and Inceptisols. The uncertainty related to the inferred soil type were also investigated. The results are depicted in Table 3.4. HX and FT presented lowest entropy values than UT and AX. These results are strongly related to the overall accuracy, found with the field samples.

For exaggeration, the mean value was 0.04 and the standard deviation 0.13. Only 18% of the study area presented values greater than zero, of which 71% were lower than 0.3. The higher exaggeration was related to ridges. The higher the exaggeration, the lower is the similarities to the assigned soil type. In general, high exaggeration values are related to random variations that were not well captured by the membership functions, or the similarity vector is not an accurate representation of the local soil types (Zhu, 1997b). In this study, the latter would be more appropriated due to the difficulty on defining limiars for the membership functions due to heterogeneity in the soil-landscape relationship in these sites.

As shown, the entropy and exaggeration measures are important sources of information, with regards to the quality of the inference process and also for the quality of the knowledge obtained. It also can be useful for the development of the Rulebase. When the inference process results in high entropy levels, it indicates that the instances are not well captured by the soil-environment configurations, or the database is not adequate to capture the soil variations. With this information, the user can revisit the knowledge and make a fine tuning of the instances. Additionally, as highlighted by Qi and Zhu (2011) this source of data would be useful for proper

management of soil resources, by considering the transitional zones as different of their respective optimal characteristics, and so, properly suggest their potential use.

Figure 3.8 - Inference uncertainty based on entropy values and its relative frequency distribution according to a landforms classification.



3.2.4 CBR approach

A total of 24 inferences were performed. The configuration of each Casebase and the results of external validation are shown in the Table 3.5. The mean OA was 46% and KI of 0,279. The best performance (61% (14/23) OA and 0,518 KI) was obtained by the dataset with the full polygon with the *ea*-TI ensemble of variables, using the mean as optimality value and the multiplication factor for *P* function.

Regarding the variables selection, while other inference approaches as Random Forest and Decision Trees well handle a great number of covariates, ArcSIE seems to perform better with a reduced subset of variables. The multidimensionality has been tested and discussed for data-driven techniques in different areas of knowledge (Millard and Richardson, 2015; Heung et al., 2014; Scarpone et al., 2017; Yu et al., 2016), however, there are no such discussion for knowledge-based tools. In general, for data mining tools, by reducing the data sets dimensionality, the expectation is that the processing results in equally, if not better, inferences as for the original data set (Liu and Motoda, 1998).

There are an extensive number of algorithms to figure out the importance of environmental covariables or subset selection. John et al. (1994) divide them into filter (e.g., Principal component analysis) and wrapper (e.g., Mean Decrease in Accuracy) approaches. However, the variables selection can also be Knowledge-driven (Shi et al., 2004; Shi et al., 2009; Zhu et al., 2010; Zhu et al., 2014) and, if desired, statistical tools (as Box Plots) can be used to define a subset (Silva et al., 2016; Brown et al., 2012). In this study, based on the Table 3.5, the Knowledge-driven approach for subset selection (*ea-TI*) presented better performance for 58% of the cases in direct comparison to the wrapper approach (*rf-TI*), with an average difference of 11% for OA and 0.152 for KI. The exceptions were mainly for the MUP-30 polygons.

Different sampling areas were also tested. In general, by using the original polygons and the ones with 20 m of exclusion zones, the inference accuracy was relatively similar, on average, 46.4% of OA and 0.296 of KI and 50% OA and 0.358 KI, respectively. The best prediction was obtained using the complete polygon (Casebase - AEMei - 61% of OA) (APPENDIX G). The cases developed from MUP-30 polygons resulted in the worst average accuracy, 39.9% OA and 0.183 KI. In respect to the buffer zones, these results are in agreement with Giasson et al. (2015), by failing to provide relevant advantages for the inference process. Furthermore, as stated by Pelegrino et al. (2016), in this same area of study, but using a Decision

Tree approach, there was a reduction in prediction accuracy when the exclusion zones were larger than 20 m from boundaries.

Table 3.5 - Settings, accuracy and uncertainty of the CBR models.

MUP	Optimality	P function	Variables	OA%	K I	EntOA	Exag	Casebase
e	mean	limiting	ei-TI	39	0.229	0.68	0.61	AELei
e	mean	multiplication	ei-TI	61	0.518	0.52	0.84	AEMei
e	mode	limiting	ei-TI	47	0.290	0.6	0.74	AOLei
e	mode	multiplication	ei-TI	52	0.343	0.43	0.9	AOMei
-20	mean	limiting	ei-TI	56	0.423	0.6	0.63	BELei
-20	mean	multiplication	ei-TI	56	0.423	0.44	0.82	BEMei
-20	mode	limiting	ei-TI	52	0.391	0.64	0.62	BOLei
-20	mode	multiplication	ei-TI	56	0.470	0.47	0.85	BOMei
-30	mean	limiting	ei-TI	43	0.232	0.37	0.75	CELei
-30	mean	multiplication	ei-TI	43	0.196	0.36	0.78	CEMei
-30	mode	limiting	ei-TI	34	0.128	0.36	0.88	COLei
-30	mode	multiplication	ei-TI	21	0.004	0.44	0.97	COMei
e	mean	limiting	rf-TI	52	0.353	0.67	0.57	AELrf
e	mean	multiplication	rf-TI	43	0.225	0.47	0.8	AEMrf
e	mode	limiting	rf-TI	39	0.197	0.57	0.67	AOLrf
e	mode	multiplication	rf-TI	39	0.211	0.38	0.85	AOMrf
-20	mean	limiting	rf-TI	47	0.306	0.66	0.58	BELrf
-20	mean	multiplication	rf-TI	34	0.174	0.46	0.81	BEMrf
-20	mode	limiting	rf-TI	47	0.305	0.55	0.72	BOLrf
-20	mode	multiplication	rf-TI	52	0.374	0.4	0.88	BOMrf
-30	mean	limiting	rf-TI	52	0.345	0.38	0.82	CELrf
-30	mean	multiplication	rf-TI	48	0.283	0.36	0.94	CEMrf
-30	mode	limiting	rf-TI	39	0.101	0.35	0.88	COLrf
-30	mode	multiplication	rf-TI	39	0.172	0.42	0.97	COMrf

MUP – spatial reference (e=full polygon; -20= 20 m of exclusion zones; -30= 30 m of exclusion zones); Optimality – statistics applied for the *V1-2* parameter of *E* function; Variables – ensemble of variables (ei-TI=knowledge driven selection; rf-TI=wrapped approach); OA% - Overall Accuracy; KI – Kappa Index; Ent – Entropy; Exag – Exaggeration.

This study also tested different statistics as the central concept for the fuzzy membership functions. At first, guided by the idea of major (most common) components, it was expected that, by applying the modal value of a given feature as its optimality value, it would result in more accurate predictions. However, the findings pointed out this assumption to be wrong. In direct comparisons, the models using the mean as the optimality value resulted in the most accurate predictions in 83% of the cases, with an average difference of 8.2% and 9.8% for OA and KI, respectively. It was not clear what may be caused this bias. By analyzing the effect of

alternating the integration parameter (*P* function), it was noted that, although the individual analysis indicates a better setting for a given function, the multivariate inference character turns it more complex to define an "overall better configuration", analyzing an automated method. For the cases developed with the full polygons, it was noticed that, independently of the statistic applied to the optimal values, using the *ei-TI* ensemble, the *Multiplication* factor performed better than the *Limiting* factor, and for the *rf-TI,* using *Limiting* for *P* function resulted in equal or better accurate inferences compared to *Multiplication*. For the polygons with 20 m of exclusion zones, those developed based on mean as optimality values performed better with the *Limiting* factor, and those using the mode were more accurate when integrated by *Multiplication* as *P* function, independently of the ensemble of variables. For the polygons with 30 m of exclusion zones, the *Limiting* factor presented better results as *P* function, independently of the others parameters.

Based on the exposed, it is reasonable to affirm that different data sets will demand different configurations. The best setting resulted from a trial and error procedure. However, the great issue of developing the Casebase as a semi-automatic procedure is the adjustments for *v1-v2* (optimality) and *w1-w2* (deviations). In this study, while the values for *v1-v2* corresponded to the mean or mode of a specific ensemble of environmental features (TI), the *w1-w2* values were set as their standard deviation. The curves were bell-shape by default. These settings result in a symmetric curve, which, in some (not rare) cases, did not well represent the knowledge implicitly represented in an existing mapping unit.

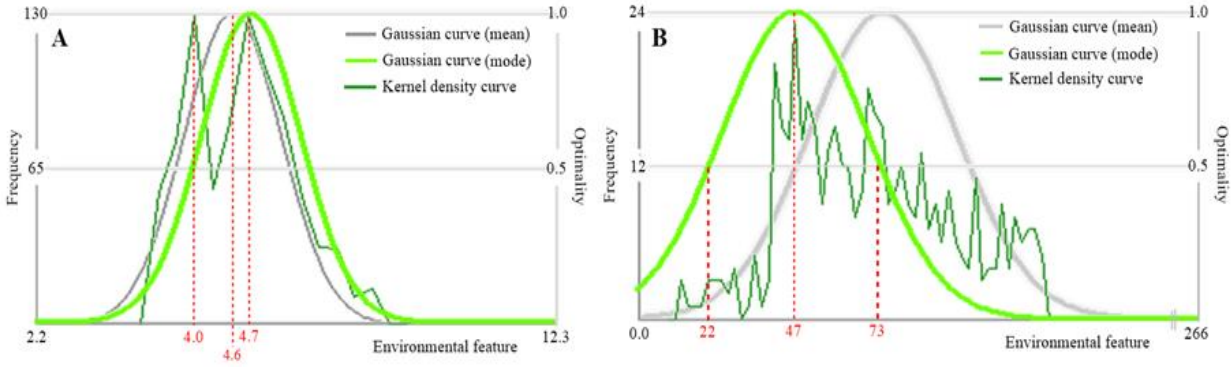
Figure 3.9 shows the fuzzy membership functions (Gaussian curves) and the kernel-smoothed frequency curve of a given environmental feature related to a specific mapping unit, which, in turn, corresponds to a case. Figure 3.9A, observing the kernel curve, it is possible to note two peaks, at 4.0 and 4.7 (environmental feature values). Comparing with the Gaussian curves, both values, despite its similar frequency, are not equivalent on optimality, so the curve fitted with the modal value, peak 4.0 corresponds to half of the optimization value of 4.7. In such cases, it would be feasible to keep the bell-shaped curve, but adjustment of values *v1-v2* to comprehend the range from 4.0 to 4.7, or multiple instances should be considered.

Figure 3.9B represents a common problem for relatively large polygons, wrapping a wide range of the environmental feature values. By setting the *w1-w2* as the standard deviation, the Gaussian curve attributed high optimality values for low-frequency feature values on the left side of the curve and lower optimality for relatively most frequent ones on the right side. Also, in spite of the mean curve seems to be more realistic, the most common value (47), which in turn should have a high (if not full) optimality value, corresponds to only a half. For this

case, it should be considerate that a S-shaped curve might best characterize the relationship between the soil and this environmental factor.

In general, the models presented poor accuracy rates, even the best model (AEMei) and was not satisfactory based on the criteria of the Brazilian Pedology Technical Manual (IBGE, 2015), which establishes that soil maps must present up to 30% of inclusions to be acceptable. Nevertheless, once these models were created on KD module and stored, the Casebase can be revisited whenever necessary. Thus, this kind of semi-automated approach would be useful for a first evaluation on the different sets of data and settings, and if desired, a given Casebase would be improved.

Figure 3.9 - Optimality and Kernel density curves of two different cases. The Gaussian curves were adjusted with the mean and mode of a given environmental feature spatially constricted by a mapping unit (polygon).



3.2.5 CBR prediction uncertainty

The CBR uncertainty indexes (entropy and exaggeration) are shown in Table 3.6. Comparing with RBR approach, the 24 predicted maps presented higher uncertainty. In part, these values can be explained by the already discussed soil-landscape associations, which naturally make it difficult to discriminate the soil types. However, the development of models based on symmetric Gaussian curves also contributed as discussed above.

A major difference observed from the comparison of the six groups of maps in Table 3.6 is that using the MUP with 30 m of exclusion zone (MUP -30m) as spatial reference, in which the overall level of entropy was lower than both the complete MUP (MUPe) and the one with 20 m of exclusion zone (MUP -20m), but resulted in an overall higher level of exaggeration uncertainty. By examining the optimality curves derived from the different sets of spatial references (polygons), it was noted that the MUP-30m group tended to generate narrower

curves than others. This is directly related to the reduction of the polygons dimensionality. Smaller polygons are more constrained in terms of capturing the environmental feature range. As a result, there is an overlapping reduction among other soil types. Thus, lower entropy values are found, but also resulted in higher exaggeration uncertainty. This effect of narrower curves resulting in lower entropy values and higher exaggeration uncertainty was also observed by Qi and Zhu (2011).

Table 3.6 - Overall entropy and exaggeration measures of the predicted soil type maps.

MUP	Entropy	Exaggeration	Entropy	Exaggeration
	ei-TI		*rf-TI*	
e	0.558	0.773	0.523	0.723
-20	0.538	0.730	0.518	0.748
-30	0.383	0.845	0.378	0.903

MUP – Casebase spatial reference (e=full polygon; -20= 20 m of exclusion zones; -30= 30 m of exclusion zones); ei-TI – knowledge driven ensemble of variables; rf-TI wrapped approach for variables selection.

The entropy is related to the membership diffusion in the similarity vector (Zhu, 1997b) in the inference result. For a pixel (*i*,*j*), the higher the membership for a single class (and lower for others), the lower the entropy, and hence the uncertainty. This explains how, by reducing the overlapping, the entropy of MUP -30m groups was lower than others. On the other hand, the uncertainty related to the exaggeration deals with the membership saturation to the assigned class (Zhu, 1997b). Models based on narrower curves tend to result in a relatively lower frequency of high membership values leading to higher exaggeration uncertainty and lower accuracy of the predicted maps.

3.3 CONCLUSIONS

The RBR approach led to the development of a formal framework, structuring and formalizing the knowledge retrieved from the soil legacy data. This mapping technique allowed the creation of an accurate soil map, covering an area 15.5 times greater than the reference area, up to the second level according to Soil Taxonomy.

The CBR approach proposed in this study, as a semi-automatic modeling technique did not result in a satisfactory soil spatial prediction. The inference process resulted in relatively great uncertainty over all of the 24 tested sets. However, it could be useful for analyzing the

performance of different data sets under different configurations for variables selection or function adjustments. Regarding the CBR configurations, the use of the mean or mode as optimal value did not result in great differences in accuracy, as well as for the Limiting or Multiplication settings for P function.

For the spatial settings, by excluding 30 m from the boundaries of the polygons, the membership curves became narrower, increasing the uncertainty in the inference process, being the original polygons those that promoted the best predictions.

REFERENCES

Akumu, C.E., Johnson, J.A., Etheridge, D., Uhlig, P., Woods, M., Pitt, D.G.; McMurray, S., 2015. GIS-fuzzy logic based approach in modeling soil texture: Using parts of the Clay Belt and Hornepayne region in Ontario Canada as a case study. Geoderma. 239, 13-24.

Behrens, T., Schmidt, K., Ramirez-Lopez L., Gallant, J., Zhu, A., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma. 213, 578-588.

Beven, K.J., Kirkby, M.J., 1979. A physically based variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43-69.

BRASIL. Levantamento de Recursos Naturais – Rio de Janeiro/Vitória, Folhas SF.23/24. Rio de Janeiro: RADAMBRASIL, 1983. 779 p.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5-32.

Brow, R.A., McDaniel, P., Gessler, P.E., 2012. Terrain attribute modeling of volcanic ash distributions in northern Idaho. Soil Sci. Soc. Am. J. 76, 179-187.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17–26.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma. 111, 21–44.

COMPANHIA MINERADORA DE MINAS GERAIS. Nota explicativa dos mapas geológicos, metalogenético e de ocorrências minerais do estado de Minas Gerais. Escala 1:1,000,000. Belo Horizonte, 1994.

Congalton, R.G., Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, New York, NY, USA.

Conrad, O., Bechtel, B., Bock,M., Dietrich, H., Fischer, E., Gerlitz, L.,Wehberg, J.,Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991–2007.

Curi, N., Franzmeier, D.P., 1984. Toposequence of Oxisols from the Central Plateau of Brazil. Soil Sci. Soc. Am. J. 48, 341–346.

Curi, N., Chagas, C. da S., Giarola, N.F.B. 1994. Distinção de ambientes agrícolas e relação solo-pastagens nos campos da Mantiqueira. In: Carvalho, M.M., Evangelista, A.R., Curi, N., (Eds.) Desenvolvimento de Pastagens na Zona Fisiográfica Campos das Vertentes, MG. Embrapa, pp. 21-43.

Demattê, J.A.M., Rizzo, R., Boteon, V.W., 2015. Pedological mapping through integration of digital terrain models spectral sensing and photopedology. Ciência Agronômica. 46(4), 669-678. http://dx.doi.org/10.5935/1806-6690.20150053.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma. 97, 367–391.

Favrot, J.C., 1989. A strategy for large scale soil mapping: the reference areas method. Science du Sol. 27, 351-368.

Foos, T., Schum, G., Rothenberg, S. 2006. Tacit knowledge transfer and the knowledge disconnect. Journal of Knowledge Management. 10(1), 6-18.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347–1359.

Giasson, E., Clarke, R.T., Inda Junior, A.V., Merten, G.H., Tornquist, C.G., 2006. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Sci. Agric. 63(3), 262-268.

Giasson, E., Ten Caten, A., Bagatini, T., Bonfatti, B., 2015 Instance selection in digital soil mapping: A study case in Rio Grande do Sul, Brazil. Ciência Rural. 45, 1592-1598.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY (734 pp.).

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma. 214-215, 141-154.

Heung, B., Hodúl, M., Schmidt, M.G., 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma. 290, 51-68.

Hudson, B.D. 1992. The soil survey as a paradigm-based science. Soil Science Society of America Journal. 56, 836-841.

IBGE, 2015. Manual Técnico de Pedologia. third ed. IBGE, Rio de Janeiro.

Iwahashi, J. and Pike, R.J., 2007. Automated Classifications of Topography from DEMs by an Unsupervised Nested- Means Algorithm and a Three-Part Geometric Signature. Geomorphology. 86, 409-440.

Jenny, H., 1941. Factors of Soil Formation.ASystemofQuantitative Pedology,McGraw-Hill, New York.

John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. Proceedings of the Eleventh International Machine Learning Conference. Morgan Kaufmann, New Brunswick, pp. 121–129.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., Vries de, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal. 76 (6), 2097-211.

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

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News. 2, 18-22.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluver.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma.117, 3–52.

McKay J., Grunwald S., Shi X., Long R. 2010. Evaluation of the Transferability of a Knowledge-Based Soil-Landscape Model. In: Boettinger J.L., Howell, D.W., Moore A.C., Hartemink A.E., Kienast-Brown S. (Eds) Digital Soil Mapping. Progress in Soil Science, vol 2. Springer, Dordrecht

Menezes de, M.D., Curi, N., Marques, J.J., Mello de, C.R., Araújo de, A.R. 2009. Pedologic survey and geographic information system for evaluation of land use within a small watershed, Minas Gerais State, Brazil. Ciência e Agrotecnologia. 33,1544-1553.
Menezes de. M.D., Silva, S.H.G., Owens, P.R., Curi, N., 2013. Digital soil mapping approach based on fuzzy logic and field expert knowledge. Ciência e Agrotecnologia. 37(4), 287-298.

Menezes, M.D., Silva, S.H.G., Mello, C.R., Owens, P.R., Curi, N., 2014. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. Sci. Agric. 71, 316–323.

Millard, K., Richardson, M., 2015. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. Remote Sensing. 7, 8489-8515.

Neto R.M. 2014. Geomorphological Surfaces and the Evolution of Brazilian Relief: Passing ideas and connections in Southern Minas Gerais, Southeast Brazil. Ra'Ega. 32, 267-295.

Nonaka, I., Totama, R. and Nagata, A., 2000. A firm as a knowledge-creating entity: a new perspective on the theory of the firm. Industrial and Corporate Change. 9(1), 1-20.

Pelegrino M.H.P., Silva, S.H.G., Menezes de, M.D., Silva da, E., Owens, P.R.,

Curi, N., 2016. Mapping soils in two watersheds using legacy data and extrapolation for similar surrounding areas. Ciência e Agrotecnologia. 40(5), 534-546.

Qi, F., Zhu, A.-X., 2011. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. Computers and Geosciences. 37, 1425-1436.

Qi, F., Zhu, A-X., Harrower, M., Burt, J. E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma. 136, 774-787.

Scarpone, C., Schmidt, M. G., Bulmer, C. E., Knudby, A., 2017. Semi-automated classification of exposed bedrock cover in British Columbia's Southern Mountains using a Random Forest approach. Geomorphology. 285, 214-224.

Shi, X., Long, R., Dekett, R., Philippe, J., 2009. Integrating Different Types of Knowledge for Digital Soil Mapping Soil Science Society of America Journal. 73, 1682-1692.

Shi, X., Zhu, A.-X., Burt, J.E., Qi, F., Simonson, D., 2004. A case-based reasoning approach to fuzzy soil mapping. Soil Science Society of America Journal. 68, 885-894.

Silva, S.H.G., Menezes de, M.D., OWENS, P.R., CURI, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma. 267, 65-77.

Skidmore, A.K., Ryan, P.J., Dawes, W., Short, D., Emmett, O.L., 1991. Use of an expert system to map forest soils from a geographical information system. Int. J. Geogr. Inf. Syst. 5, 431–445.

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

Ten Caten, A., Dalmolin, R.S.D., Ruiz, L.F.C., 2012. Digital soil mapping: strategy for data pre-processing. Revista Brasileira de Ciência do Solo. 36, 1083-1091.

Yu, L., Fu, H.,Wu, B., Clinton, N., Gong, P., 2016. Exploring the potential role of feature selection in global land-cover mapping. International Journal of Remote Sensing. 37 (23), 5491-5504.

Zhu, A.X. 1997a. A similarity model for representing soil spatial information. Geoderma. 77, 217–242.

Zhu, A.-X., 1997b. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Photogrammetric Engineering and Remote Sensing. 63(10), 1195-1202.

Zhu, A.-X., Wang, R., Qiao, J., Qin, C.Z., Chen, Z., Liu, J., Du, F., Lin, Y., Zhu, T., 2014. An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic. Geomorphology. 214, 128-138.

Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Can. J. Remote Sens. 20, 408–418.

Zhu, A.-X., Band, L.E., Dutton, B., Nimlos, T.J., 1996. Automated soil inference under fuzzy logic. Ecological Modeling. 90,123-145.

Zhu, A.-X., Yang, L., Li, B., Qin, C., Pei, T., Liu, B., 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma. 155, 164-174.

APPENDIX

APPENDIX A – Soil profiles descriptions and physicochemical analysis

APPENDIX B - Distribution of soil profiles used for external validation of prediction models

APPENDIX C - Geology at the study area

APPENDIX D – TPI based landforms classification map of the study area

APPENDIX E - Hypsometric map of the study area

APPENDIX F – Predicted soil maps based on models: A) PCA Buffer-Point (PCA5); B) PCA POL-20 (MDA9)

APPENDIX G – CBR predicted soil map by the model AEMei

APPENDIX A – Soil profiles descriptions and physicochemical analysis

HAPLUDOX (PT2)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	08/feb/2017
Coordinates UTM	-	556687 x 7594778 m, fuse 23, *datum* WGS 1984
Land use	-	native pasture
Parent Material	-	biotite schist or gneiss
Landform	-	high Ridge
Relief [2]	-	steeply sloping to very steeply sloping
Slope (%)	-	7.2
Elevation (m)	-	1107
Aspect (º)	-	141 (Southeast)
Erosion	-	severe laminar
Permeability	-	well drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.1 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	28	14	58
B	35	16	49

Table 4.2 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.1	40.86	0.39	0.31	0.14	0	3.95
B	5.8	8.22	0.06	0.16	0.1	0	1.2

Table 4.3 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-------cmol/dm³-------			------ % ------		dag/kg	mg/L
A	0.55	0.55	4.5	12.33	0	2.96	19
B	0.28	0.28	1.48	18.99	0	0.75	3.75

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (PT3)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	08/feb/2017
Coordinates UTM	-	561114 x 7595805 m, fuse 23, *datum* WGS 1984
Land use	-	pasture
Parent Material	-	biotite schist or gneiss
Landform	-	High Ridges
Relief [2]	-	steeply sloping
Slope (%)	-	7.2
Elevation (m)	-	1106
Aspect (º)	-	275 (West)
Permeability	-	well drained
A horizon	-	prominent
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.4 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	54	17	29
B	52	16	32

Table 4.5 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5	34.33	0.17	0.21	0.12	0	8.05
B	5.3	10.39	0.01	0.15	0.1	0	2.65

Table 4.6 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.42	0.42	8.47	4.94	0	3.09	8.57
B	0.28	0.28	2.93	9.44	0	1.16	6.27

[1] SiBCS - Brazilian Soil Classification System

[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

UDEPT (PT4)

Classification (SiBCS[1])	-	Cambissolo Háplico
Date	-	08/feb/2017
Coordinates UTM	-	557545 x 7596076 m, fuso 23, *datum* WGS 1984
Landuse	-	native forest (litter accumulation)
Parent Material	-	biotite schist or gneiss
Landform	-	open slope (Upper third)
Relief [2]	-	steeply to very steeply sloping
Slope (%)	-	27.4
Elevation (m)	-	1020
Aspect (º)	-	256 (West)
Permeability	-	well drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.7 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	29	28	43
B1	28	21	51
B2	24	35	41

Table 4.8 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	6.2	82.22	2.08	4.25	1.5	0	3.2
B1	5.2	32.16	0.48	0.43	0.17	0	3.5
B2	5.7	29.98	0.23	0.63	0.33	0	1.22

Table 4.9 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	5.96	5.96	9.16	65.07	0	5.31	22.6
B1	0.68	0.68	4.18	16.33	0	1.55	18.66
B2	1.04	1.04	2.26	45.88	0	0.53	25.73

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

ACRUDOX (PT5)

Classification (SiBCS[1])	-	Latossolo Vermelho
Date	-	08/feb/2017
Coordinates UTM	-	562840 x 7592060 m, fuso 23, *datum* WGS 1984
Landuse	-	pasture
Parent Material	-	biotite gneiss (banded) or quartzite
Landform	-	high ridges
Relief [2]	-	strongly to steeply sloping
Slope (%)	-	15.2
Elevation (m)	-	1140
Aspect (º)	-	1 (North)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/6; B - 2.5YR 4/8
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.10 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	24	23	53
B	32	24	44

Table 4.11 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.8	147.51	0.78	1.7	1.6	0	2.37
B	5.7	56.1	0.31	0.22	0.59	0	2.03

Table 4.12 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	3.68	3.68	6.05	60.8	0	3.29	36.91
B	0.95	0.95	2.98	32.01	0	0.18	9.24

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (PT6)

Classification (SiBCS [1])	-	Latossolo Vermelho-Amarelo
Date	-	08/feb/2017
Coordinates UTM	-	566356 x 7593936 m, fuso 23, datum WGS 1984
Landuse	-	pasture
Parent Material	-	biotite gneiss (banded) or quartzite
Landform	-	plain
Relief [2]	-	moderately to strongly sloping
Slope (%)	-	4.9
Elevation (m)	-	1040
Aspect (º)	-	299 (Northwest)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 3/4; B - 5YR 5/8
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.13 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	21	10	69
B	35	6	59

Table 4.14 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.4	43.04	0.34	0.64	0.29	0	2.74
B	5.8	6.04	0.01	0.11	0.1	0	1.04

Table 4.15 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	1.04	1.04	3.78	27.52	0	2.96	29.02
B	0.23	0.23	1.27	17.75	0	0.35	3.49

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (PT8)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	08/feb/2017
Coordinates UTM	-	565105 x 7597367 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	biotite gneiss (banded)
Landform	-	plain
Relief [2]	-	very gently to moderate sloping
Slope (%)	-	6.9
Elevation (m)	-	991
Aspect (º)	-	59 (Northeast)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/6; B - 5YR 5/6
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.16 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	30	13	57
B	42	12	46

Table 4.17 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.5	58.27	0.56	0.8	0.29	0	2.4
B	6.1	6.04	0.09	0.14	0.1	0	1.3

Table 4.18 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	1.24	1.24	3.64	34.05	0	2.22	22.99
B	0.26	0.26	1.56	16.38	0	0.67	4.66

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (´PT9)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	08/feb/2017
Coordinates UTM	-	558661 x 7590305 m, fuso 23, *datum*, WGS 1984
Landuse	-	remnant of native forest
Parent Material	-	biotite gneiss (banded)
Landform	-	upper slopes
Relief [2]	-	strongly sloping
Slope (%)	-	5
Elevation (m)	-	1129
Aspect (º)	-	337 (North)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)		horizons: A - 5YR 3/4; B - 5YR 5/6
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.19 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	44	10	46
B	55	12	33

Table 4.20 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5	53.92	0.59	0.38	0.22	0	5.96
B	5.7	14.74	0.12	0.14	0.1	0	2.4

Table 4.21 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.74	0.74	6.7	11.02	0	3.25	15.45
B	0.28	0.28	2.68	10.37	0	1.23	8.22

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

FLUVENT (PT10)

Classification (SiBCS[1])	-	Neossolo Flúvico
Date	-	08/feb/2017
Coordinates UTM	-	566742 x 7595145 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	alluvial sediments of the quaternary
Landform	-	floodplain
Relief [2]	-	level to gently sloping
Slope (%)	-	3.9
Elevation (m)	-	975
Aspect (º)	-	284 (West)
Erosion	-	not apparent
Permeability	-	imperfectly drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.22 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	17	15	68
C	13	7	80

Table 4.23 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.6	56.1	2.89	0.94	0.47	0	3.42
C	5.4	12.57	1.79	0.31	0.1	0	2.12

Table 4.24 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	1.55	1.55	4.97	31.26	0	2.45	27.79
C	0.44	0.44	2.56	17.27	0	0.41	27.64

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

FLUVENT (P11)

Classification (SiBCS[1])	-	Neossolo Flúvico
Date	-	07/feb/2017
Coordinates UTM	-	567936 x 7602331 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	alluvial sediments of the quaternary
Landform	-	floodplain
Relief [2]	-	depressional to level
Slope (%)	-	0
Elevation (m)	-	958
Aspect (º)	-	-1 (flat)
Erosion	-	not apparent
Permeability	-	imperfectly drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.25 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	37	45	18
C	48	41	11

Table 4.26 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.7	88.74	3.16	1.67	1.06	0	5.34
C	5.7	51.74	1.41	0.74	0.31	0	5.28

Table 4.27 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	2.96	2.96	8.3	35.63	0	3.54	13.97
C	1.18	1.18	6.46	18.31	0	1.82	5.58

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

FLUVENT (PT12)

Classification (SiBCS[1])	-	Neossolo Flúvico
Date	-	07/feb/2017
Coordinates UTM	-	573864 x 7603428 m, fuso 23, *datum*, WGS 1984
Landuse	-	permanent protection area
Parent Material	-	alluvial sediments of the quaternary
Landform	-	floodplain
Relief [2]	-	strongly to steeply sloping
Slope (%)	-	11.3
Elevation (m)	-	1032
Aspect (º)	-	327 (Northwest)
Permeability	-	imperfectly drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado
Observations	-	narrow floodplain

Table 4.28 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	13	1	86
C	11	7	82

Table 4.29 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.8	14.74	0.87	0.16	0.1	0	1.24
C	5.4	34.33	0.78	0.26	0.21	0	3.35

Table 4.30 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.3	0.3	1.54	19.34	0	0.36	28.02
C	0.56	0.56	3.91	14.27	0	1.74	34.69

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

UDEPT (PT13)

Classification (SiBCS[1])	-	Cambissolo Háplico
Date	-	07/feb/2017
Coordinates UTM	-	572339 x 7599193 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	biotite schist or gneiss
Landform	-	Plain
Relief [2]	-	moderately to steeply sloping
Slope (%)	-	2.3
Elevation (m)	-	1061
Aspect (º)	-	166 (South)
Permeability	-	well drained
A horizon	-	weak (underdeveloped)
Described by	-	M. D. de Menezes; D. F. T. Machado
Observations	-	"Epipedregoso" (in portuguese)

Table 4.31 - Granulometric analysis

Horizon	Particle Size Analysis		
	Clay	Silt	Sand
	---dag/kg---		
A	13	20	67
B	11	16	73

Table 4.32 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.4	32.16	1.67	0.44	0.19	0	2.24
B	5.3	27.8	1.24	0.32	0.12	0	2.42

Table 4.33 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.71	0.71	2.95	24.15	0	2.09	37.32
B	0.51	0.51	2.93	17.45	0	1.8	39.24

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

ACRUDOX (PT15)

Classification (SiBCS[1])	-	Latossolo Vermelho
Date	-	07/feb/2017
Coordinates UTM	-	566813 x 7606055 m, fuso 23, *datum*, WGS 1984
Landuse	-	Soybean crop
Parent Material	-	phyllites; xystus or quartzite
Landform	-	upper slope
Relief[2]	-	strongly to steeply sloping
Slope (%)	-	14.2
Elevation (m)	-	1030
Aspect (º)	-	21 (North)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/3; B - 2.5YR 4/6
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.34 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	48	28	24
B	57	20	23

Table 4.35 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	6.3	95.27	0.73	2.64	1.06	0	3.27
B	4.9	43.04	0.01	0.32	0.12	0	2.62

Table 4.36 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	3.94	3.94	7.21	54.71	0	4.34	10.92
B	0.55	0.55	3.17	17.36	0	1.84	4.41

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (PT16)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	07/feb/2017
Coordinates UTM	-	562323 x 7600072 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	biotite gneiss (banded)
Landform	-	Plain
Relief [2]	-	strongly sloping
Slope (%)	-	4.3
Elevation (m)	-	1005
Aspect (º)	-	85 (East)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/4; B - 5YR 5/6
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.37 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	28	16	56
B	37	14	49

Table 4.38 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.2	40.86	0.12	0.41	0.16	0	4.72
B	5.7	14.74	0.01	0.21	0.12	0	1.78

Table 4.39 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.67	0.67	5.39	12.52	0	2.2	12.47
B	0.37	0.37	2.15	17.11	0	0.93	6.44

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (PT17)

Classification (SiBCS[1])	-	Latossolo Vermelh-Amarelo
Date	-	07/feb/2017
Coordinates UTM	-	567333 x 7605862 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	phyllites; xystus or quartzite
Landform	-	open slope (upper third)
Relief [2]	-	strongly to steeply sloping
Slope (%)	-	44.3
Elevation (m)	-	999
Aspect (º)	-	179 (South)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/4; B - 5YR 5/6
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.40 - Granulometric analysis

Horizon	Particle Size Analysis		
	Clay	Silt	Sand
	---dag/kg---		
A	23	14	63
B	28	13	59
2B	19	33	48

Table 4.41 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	6	167.1	0.31	0.37	0.15	0	2.42
B	5.6	19.1	0.26	0.18	0.1	0	1.92
2B	5.7	10.39	0.2	0.14	0.1	0	1.16

Table 4.42 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.95	0.95	3.37	28.14	0	1.92	17.8
B	0.33	0.33	2.25	14.62	0	1.13	16.14
2B	0.27	0.27	1.43	18.65	0	0.39	17.18

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

UDEPT (PT18)

Classification (SiBCS[1])	-	Cambissolo Háplico
Date	-	07/feb/2017
Coordinates UTM	-	569600 x 7600598 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	gneiss or feldspar schist
Landform	-	upper slope (near the ridge)
Relief [2]	-	steeply sloping
Slope (%)	-	26.5
Elevation (m)	-	1071
Aspect (º)	-	190 (South)
Erosion	-	frequent linear erosion
Permeability	-	well drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado
Observations	-	degraded pasture and poorly maintained roads

Table 4.43 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	23	15	62
Bi	24	15	61

Table 4.44 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	4.9	19.1	0.5	0.2	0.1	0	3.91
Bi	5.1	10.39	0.26	0.15	0.1	0	4.04

Table 4.45 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.35	0.35	4.26	8.19	0	2.47	27.64
Bi	0.28	0.28	4.32	6.4	0	1.44	23.99

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

HAPLUDOX (EX2)

Classification (SiBCS[1])	-	Latossolo Vermelho-Amarelo
Date	-	07/feb/2017
Coordinates UTM	-	570391 x 7604930 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	biotite schist or gneiss
Landform	-	open slope (upper third)
Relief [2]	-	strongly to steeply sloping
Slope (%)	-	6.8
Elevation (m)	-	1038
Aspect (º)	-	11 (North)
Permeability	-	well drained
A horizon	-	moderate
Soil color (wet)	-	horizons: A - 5YR 4/6; B - 5YR 5/8
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.46 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	23	15	62
B	31	14	55

Table 4.47 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.2	23.45	0.34	0.16	0.1	0	2.77
B	5.6	8.22	0.17	0.23	0.1	0	1.35

Table 4.48 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	0.32	0.32	3.09	10.36	0	1.81	21.27
B	0.35	0.35	1.7	20.65	0	0.46	6.01

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

UDEPT (EX5)

Classification (SiBCS[1])	-	Cambissolo Háplico
Date	-	07/feb/2017
Coordinates UTM	-	575696 x 7599388 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	biotite schist or gneiss
Landform	-	open slope (upper third)
Relief [2]	-	steeply sloping
Slope (%)	-	19.6
Elevation (m)	-	1132
Aspect (º)	-	138
Permeability	-	well drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado

Table 4.49 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	33	14	53
B	35	11	54
C	28	26	46

Table 4.50 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.4	51.74	0.59	0.72	0.45	0	2.93
B	5	16.92	0.31	0.29	0.14	0	3.91
C	5.5	6.04	0.26	0.25	0.13	0	1

Table 4.51 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	1.3	1.3	4.23	30.8	0	2.33	22.73
B	0.47	0.47	4.38	10.81	0	1.64	16.54
C	0.4	0.4	1.4	28.25	0	0.2	14.6

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

FLUVENT (EX7)

Classification (SiBCS[1])	-	Neossolo flúvico
Date	-	07/feb/2017
Coordinates UTM	-	556687 x 7594778 m, fuso 23, *datum*, WGS 1984
Landuse	-	pasture
Parent Material	-	alluvial sediments of the quaternary
Landform	-	valley
Relief [2]	-	floodplain
Slope (%)	-	0.4
Elevation (m)	-	978
Aspect (º)	-	31 (Northeast)
Erosion	-	not apparent
Permeability	-	imperfectly drained
A horizon	-	moderate
Described by	-	M. D. de Menezes; D. F. T. Machado
Observations	-	profile near an artificial-channel

Table 4.52 - Granulometric analysis

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---dag/kg---		
A	8	7	85
C	6	6	88

Table 4.53 – Soil pH, sorption complex, assimilable phosphorus and extractable acidity.

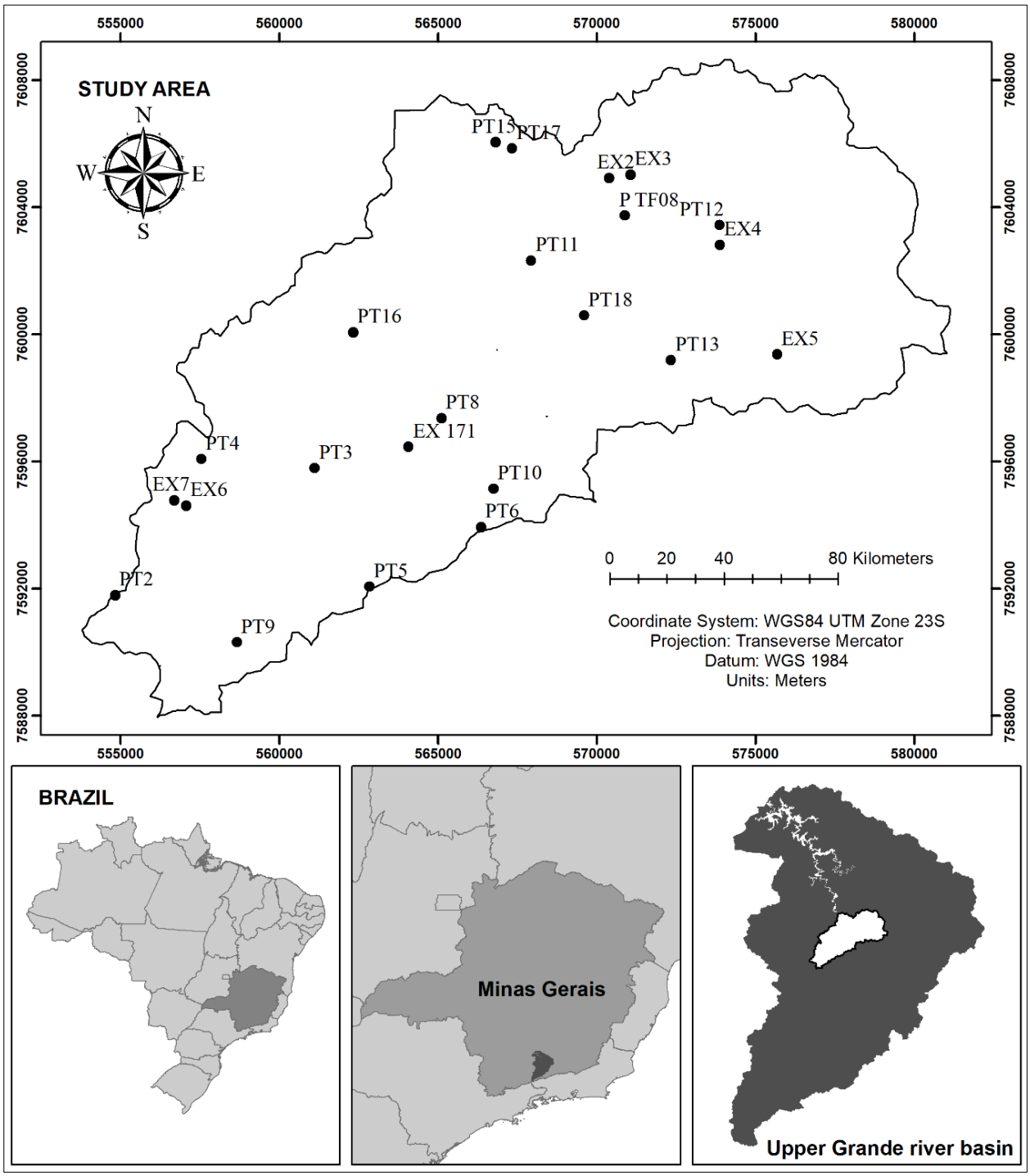
Horizon	pH	K	P	Ca	Mg	Al	H+Al
		----mg/dm³----		-----cmol/dm³-----			
A	5.9	43.04	3.13	0.9	0.32	0	1.74
C	6.6	47.39	2.92	0.72	0.25	0	1.33

Table 4.54 – Sum of bases (SB), Cation exchange capacity (CEC and CEC at pH 7,0), base saturation (V), aluminum saturation (m), organic matter and remaining phosphorus (P-Rem)

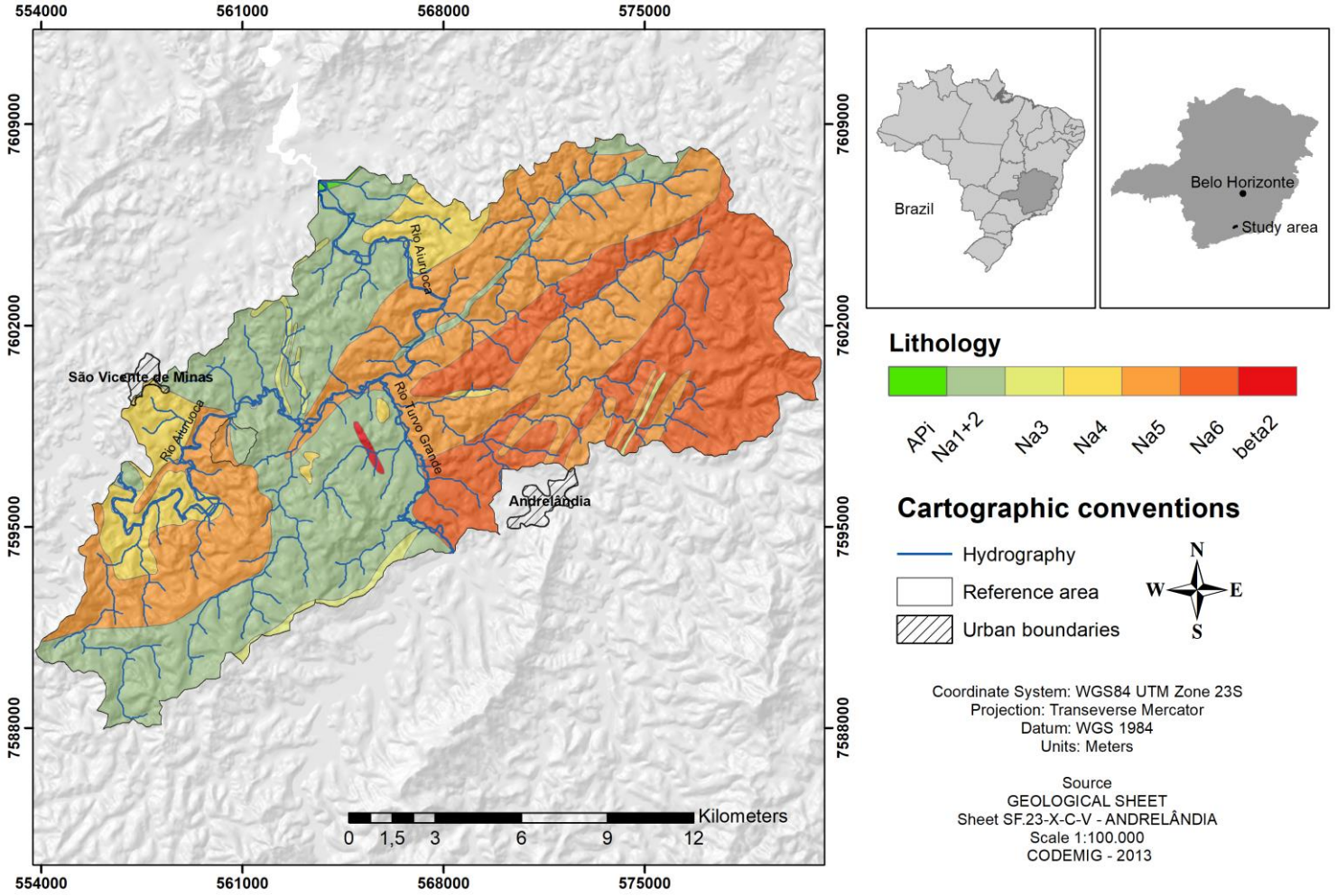
Horizon	SB	t	T	V	m	O.M.	P-Rem
	-----cmol/dm³-----			---- % ----		dag/kg	mg/L
A	1.33	1.33	3.07	43.33	0	1.97	39.78
C	1.09	1.09	2.42	45.1	0	0.77	36.19

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)

APPENDIX B - Distribution of soil observations used for external validation of Random Forest, CBR and RBR prediction models

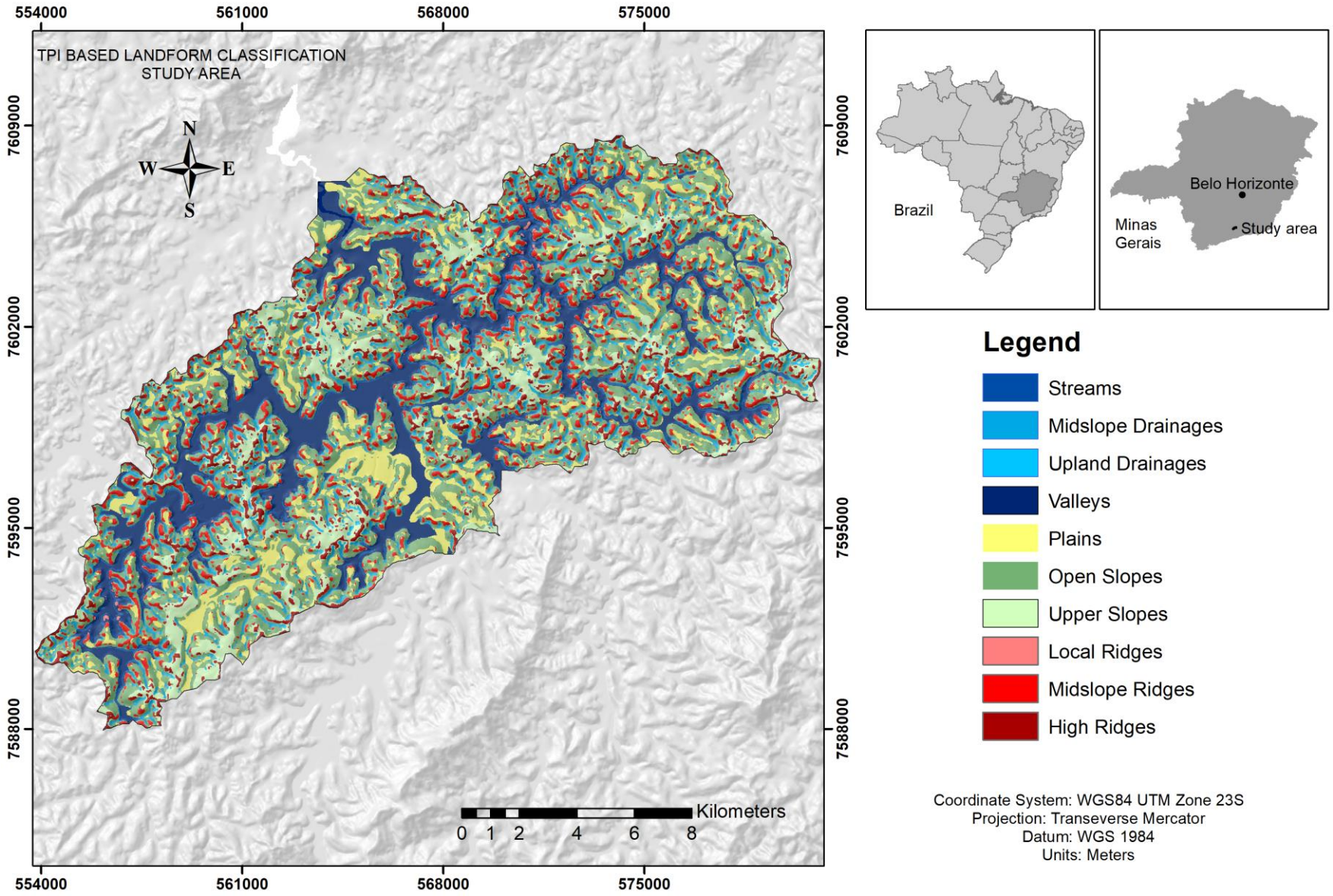


APPENDIX C - Geology at the study area

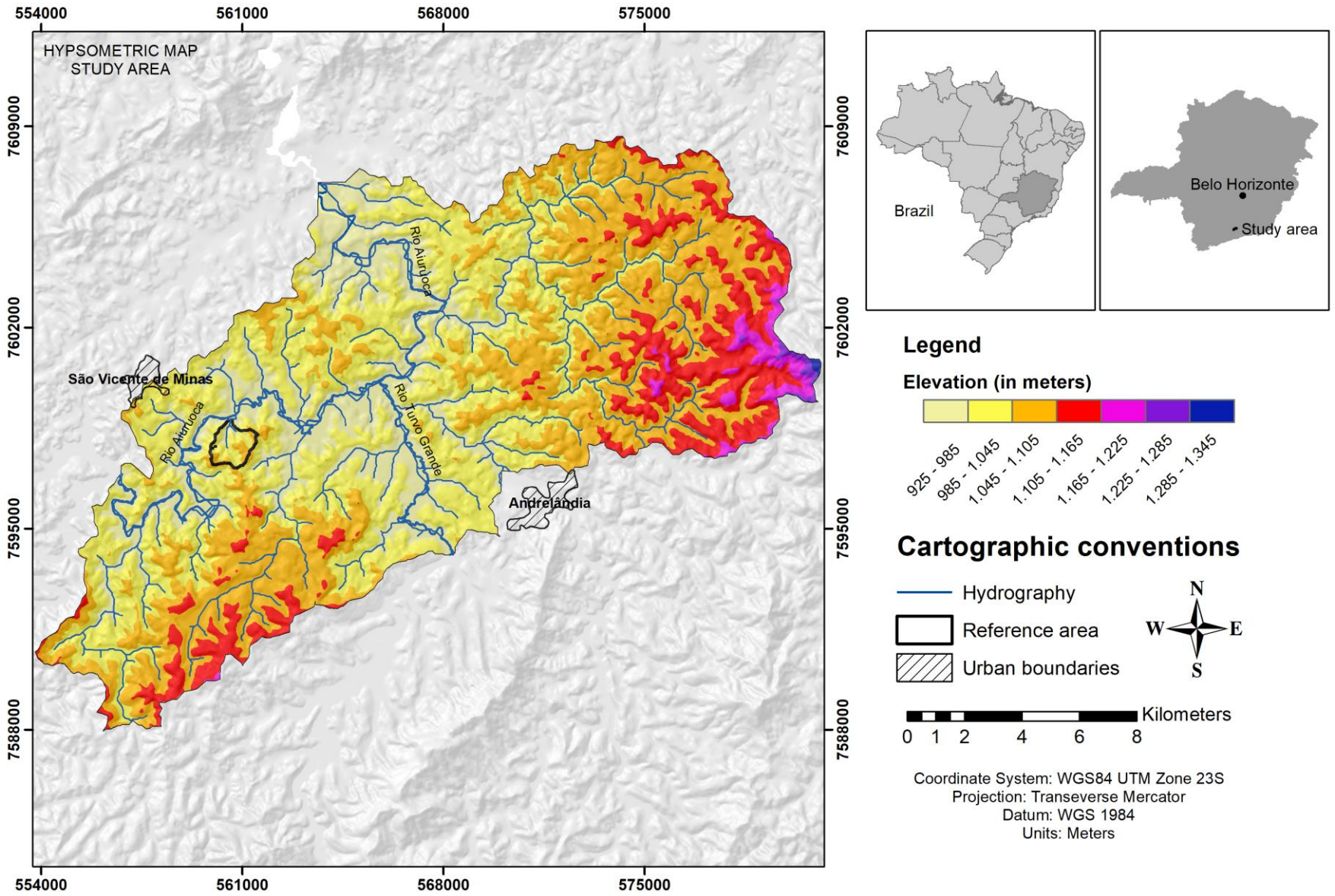


Na1+2 - Biotita gnaisses bandados com intercalações de filitos / xistos cinzentos, muscovita xistos, quartzitos, anfibolitos e rochas ultramáficas; **Na3** - Quartzito e quartzo xisto, em geral com muscovita esverdeada; **Na4** - Filitos ou xistos cinzentos e quartzitos; **Na5** - Biotita xisto ou gnaisse; **Na6** - Gnaisse e xisto feldspático; com intercalações de muscovita xisto, quartzito, quartzito micáceo, quartzito manganesífero, rochas alcissilicáticas e anfibolitos, com veios de turmalinito e pegmatito; **APi** - Ortognaisses migmatíticos indivisos; **beta2** - Anfibolitos

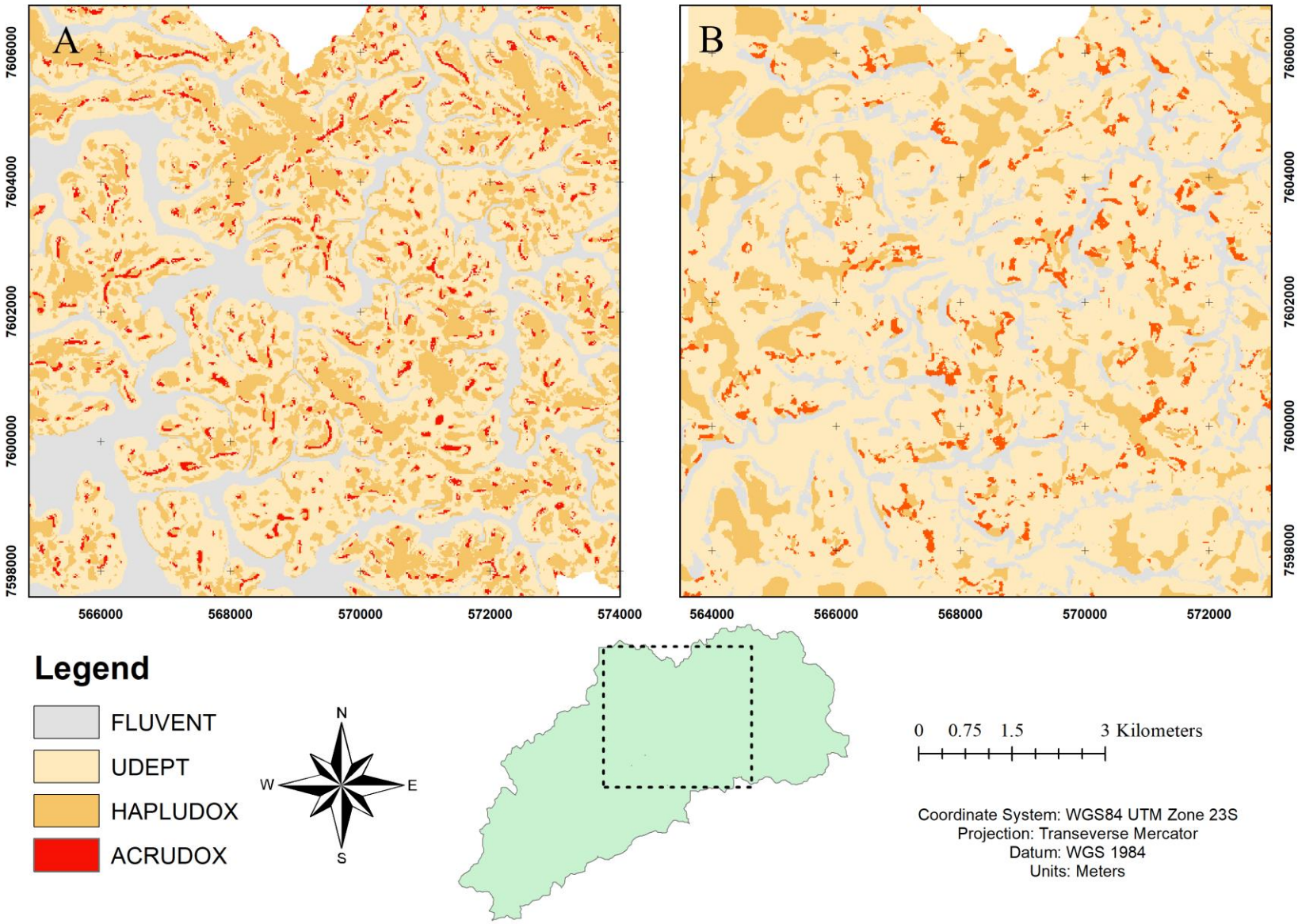
APPENDIX D – TPI based landforms classification map of the study area



APPENDIX E - Hypsometric map of the study area



APPENDIX F – Predicted soil maps based on models: A) PCA Buffer-Point (PCA5); B) PCA POL-20 (MDA9)



APPENDIX G – CBR predicted soil map using the model AEMei