SÉRGIO HENRIQUE GODINHO SILVA

DIGITAL SOIL MAPPING: EVALUATION OF SAMPLING SYSTEMS FOR SOIL SURVEYS AND REFINEMENT OF SOIL MAPS AT LOWER COST USING LEGACY DATA

LAVRAS – MG

2016

SÉRGIO HENRIQUE GODINHO SILVA

DIGITAL SOIL MAPPING: EVALUATION OF SAMPLING SYSTEMS FOR SOIL SURVEYS AND REFINEMENT OF SOIL MAPS AT LOWER COST USING LEGACY DATA

Tese 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 a obtenção do título de Doutor.

Orientador

Dr. Nilton Curi

LAVRAS – MG

2016

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

Silva, Sérgio Henrique Godinho.
Digital soil mapping : Evaluation of sampling systems for soil surveys and refinement of soil maps at lower cost using legacy data / Sérgio Henrique Godinho Silva. – Lavras : UFLA, 2016.
104 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2016.
Orientador(a): Nilton Curi.
Bibliografia.

1. Pedology. 2. Digital soil maps. 3. Soil surveys. I. Universidade Federal de Lavras. II. Título.

SÉRGIO HENRIQUE GODINHO SILVA

DIGITAL SOIL MAPPING: EVALUATION OF SAMPLING SYSTEMS FOR SOIL SURVEYS AND REFINEMENT OF SOIL MAPS AT LOWER COST USING LEGACY DATA

(**MAPEAMENTO DIGITAL DE SOLOS:** AVALIAÇÃO DE SISTEMAS DE AMOSTRAGEM EM LEVANTAMENTOS DE SOLOS E REFINAMENTO DE MAPAS DE SOLOS EXISTENTES USANDO DADOS LEGADOS A BAIXO CUSTO**)**

Tese 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 a obtenção do título de Doutor.

APROVADA em 20 de maio de 2016.

Dr. Nilton Curi	UFLA
Dr. Carlos Rogério de Mello	UFLA
Dr. Gilberto Coelho	UFLA
Dr. Fausto Weimar Acerbi Junior	UFLA
Dr. João Bosco Vasconcellos Gomes	Embrapa Florestas

Dr. Nilton Curi

Orientador

LAVRAS – MG

2016

Aos meus pais, Taísa e Walter, avós, Maria José e Danilo, e Tia, Tânia.
DEDICO.

AGRADECIMENTOS

Agradeço a Deus pela saúde, proteção, paz e suporte em toda a minha vida.

À minha família, pelo amor e apoio incondicionais, por sempre me desejar o melhor, pela compreensão e por serem para mim um exemplo e minha maior motivação.

Ao Prof. Nilton Curi, pelo apoio, amizade, confiança, ensinamentos, incentivo e por ser um exemplo durante esses 8 anos no DCS/UFLA.

A Prof[a]. Michele Duarte de Menezes, pela amizade, confiança, grande ajuda, e ensinamentos.

Ao CNPq pela concessão da bolsa de Doutorado, incentivando os estudos, e à CAPES e FAPEMIG por outros auxílios financeiros.

Aos Professores do DCS/UFLA, pela amizade e ensinamentos.

A Dirce e Damiany, pelos grandes auxílios, esclarecimentos e amizade.

A Anita, pela paciência, companheirismo, amor, amizade e motivação.

Aos meus amigos, Berardo, Totonho, Oswaldo, Otávio, Fabio, Hudson, Walbert, Bruno, Brasil, Caík, e tantos outros que não caberiam nesta página.

Aos meus amigos da turma 2008/01 de Eng. Florestal, em especial Aliny, Rafael e Rosado, pela amizade mesmo após o fim do curso.

Aos meus amigos do DCS, Giovana, Leandro, Jeani, Eliete, Lili, Luiz, Thaís, Marcelo, e aos que já passaram pela nossa sala de estudos, pela amizade e ensinamentos ao longo desses anos. E ao Zélio, Samara, Bombinha, Eduardo, Érika, Diego, Pedro, Dani, Fábio, Soraya, Ferreira, Lu, Flávia, Aline e Pezão.

Obrigado a todos que me ajudaram de alguma forma nesta jornada.

ABSTRACT

In soil surveys, several sampling systems can be used to define the most representative sites for sample collection and description of soil profiles. In recent years, the conditioned Latin hypercube sampling system has gained prominence for soil surveys. In Brazil, most of the soil maps are at small scales and in paper format, which hinders their refinement. The objectives of this work include: (i) to compare two sampling systems by conditioned Latin hypercube to map soil classes and soil properties; (II) to retrieve information from a detailed scale soil map of a pilot watershed for its refinement, comparing two data mining tools, and validation of the new soil map; and (III) to create and validate a soil map of a much larger and similar area from the extrapolation of information extracted from the existing soil map. Two sampling systems were created by conditioned Latin hypercube and by the cost-constrained conditioned Latin hypercube. At each prospection place, soil classification and measurement of the A horizon thickness were performed. Maps were generated and validated for each sampling system, comparing the efficiency of these methods. The conditioned Latin hypercube captured greater variability of soils and properties than the cost-constrained conditioned Latin hypercube, despite the former provided greater difficulty in field work. The conditioned Latin hypercube can capture greater soil variability and the cost-constrained conditioned Latin hypercube presents great potential for use in soil surveys, especially in areas of difficult access. From an existing detailed scale soil map of a pilot watershed, topographical information for each soil class was extracted from a Digital Elevation Model and its derivatives, by two data mining tools. Maps were generated using each tool. The more accurate of these tools was used for extrapolation of soil information for a much larger and similar area and the generated map was validated. It was possible to retrieve the existing soil map information and apply it on a larger area containing similar soil forming factors, at much low financial cost. The KnowledgeMiner tool for data mining, and ArcSIE, used to create the soil map, presented better results and enabled the use of existing soil map to extract soil information and its application in similar larger areas at reduced costs, which is especially important in development countries with limited financial resources for such activities, such as Brazil.

Keywords: Pedology. Digital soil maps. Soil surveys.

RESUMO

Em levantamentos de solos, diversos sistemas de amostragem podem ser empregados para a definição dos locais mais representativos para coleta de amostras e descrição de perfis. Nos últimos anos, o hipercubo latino condicionado tem ganhado destaque como sistema de amostragem em levantamentos de solos. No Brasil, a maioria dos mapas contém escalas menores e está em formato impresso, o que dificulta o seu refinamento. Os objetivos deste trabalho contemplam: (I) comparar dois sistemas de amostragem pelo hipercubo latino condicionado para mapear classes e atributos de solos; (II) resgatar informações de mapa de solos detalhado de sub-bacia hidrográfica piloto para o seu refinamento, comparando-se duas ferramentas de mineração de dados, e validação do novo mapa de solos; e (III) criar e validar um mapa de solos de área maior e similar a partir da extrapolação das informações extraídas do mapa de solos existente. Foram criados dois sistemas de amostragem pelo hipercubo latino condicionado e pelo hipercubo latino condicionado restrito pelo custo. Em cada local de prospecção, foram realizadas a classificação do solo e mensurada a espessura do horizonte A. Mapas foram gerados e validados para cada sistema de amostragem, comparando-se a eficiência desses métodos. O sistema do hipercubo latino condicionado capturou maior variabilidade de solos e atributos que o sistema restrito pelo custo, apesar daquele ter proporcionado maior dificuldade nos trabalhos de campo. O hipercubo latino condicionado padrão consegue capturar maior variabilidade dos solos da área de interesse e o hipercubo latino condicionado restrito pelo custo apresenta grande potencial para uso em levantamentos de solos, principalmente em áreas de difícil acesso. A partir de um mapa de solos em escala detalhada existente para uma sub-bacia hidrográfica piloto, foram extraídas as informações topográficas de cada classe de solo, a partir de um Modelo Digital de Elevação e seus derivados, por ferramentas de mineração de dados. Mapas foram gerados utilizando-se cada metodologia. A melhor delas foi utilizada para extrapolação de informações de solos para a área muito maior e similar e o mapa gerado foi validado. Foi possível recuperar informações do mapa de solos existente e aplicá-las em área maior, que apresenta fatores de formação do solo semelhantes, a mais baixo custo financeiro. A ferramenta KnowledgeMiner, para mineração de dados, e o ArcSIE, para criar o mapa de solos, apresentaram melhores resultados e possibilitaram o uso de mapa de solos existente para extrair informações de solos e aplicá-las em áreas maiores, com baixo custo financeiro, o que é importante principalmente em países em desenvolvimento com escassez de recursos para tais atividades, como o Brasil.

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

SUMMARY

FIRST PART

SAMPLING SYSTEMS AND LEGACY DATA FOR DIGITAL SOIL MAPPING

1 INTRODUCTION

1.1 General introduction

Soil surveys, activities that aim to identify and classify soils of an area of interest (RESENDE et al., 2014), require field work that includes description of soil profiles and collection of samples at different portions of the landscape. However, it is common for the pedologists to face some difficulties regarding the collection of samples, since many areas are difficult to be reached. This fact constrains the sampling sites to places of easy access.

Soil surveys generate several products, such as reports containing the description of the study area, including the native vegetation, climate, geographical expression, the methodology employed on the field works, the soil classes found, soil-landscape relationships, parent materials, and the final soil map at a determined scale (MOTTA et al., 2001).

Soil maps provide information on the spatial distribution of a soil class or property, allowing for planning soil management and defining the most appropriate land use for each location. This information is related to the scale of the final map: the greater the scale, the more details are provided. In Brazil, most of the existing soil maps are at small scales, which hinders a more detailed planning of activities (COELHO; GIASSON, 2010), in addition to being in press (paper-based format), making their refinement more difficult.

However, due to the advent of digital soil mapping, some alternatives have emerged to get through those refinement limitations. Digital soil mapping refers to the creation of spatial information systems, using numerical models to infer spatial and temporal variations about soil types and properties, from field

observations and expert knowledge, and also correlated environmental variables, such as satellite data and the so-called terrain attributes (LAGACHERIE; MCBRATNEY, 2007), which are derivatives of digital elevation models, a raster-based representation of topography composed of pixels that inform the local elevation value. Digital elevation models and terrain attributes are widely used for prediction of soil classes and properties, besides being of great help to define sampling places in areas of interest (ADHIKARI et al., 2013). Furthermore, this information can be used in addition to existing soil maps in order to retrieve mental models embedded on the map to be used in soil mapping models (BUI, 2004). In this sense, those technological advances are contributing to obtaining soil information at greater scales.

1.2 Objectives

The objectives of this work include: (i) to compare two sampling systems by conditioned Latin hypercube to map soil classes and soil properties; (II) to retrieve information from a detailed scale soil map of a pilot watershed for its refinement, comparing two data mining tools, and validation of the new soil map; and (III) to create and validate a soil map of a much larger and similar the area from the extrapolation of information extracted from the existing soil map.

2 REVIEW

2.1 Conditioned Latin hypercube

Various sampling systems can be employed to assist in defining the most suitable places for soil observations in the field, besides collections of samples and other measurements. Among the various sampling systems, the conditioned Latin hypercube (CLH) has gained prominence in recent years (MINASNY; MCBRATNEY, 2006; BRUNGARD; BOETTINGER, 2010; MULDER et al., 2012).

The CLH is derived from Monte Carlo sampling system, combining powerful stratification, randomness and efficient allocation of samples (MINASNY; MCBRATNEY, 2006). This system defines the sampling sites based on information related to the property to be mapped, for example, Digital Elevation Model (DEM) and its derivatives, satellite images and other available maps of both continuous and categorical variables.

However, CLH usually chooses sample sites throughout the study area. In some regions, especially those containing dense vegetation cover, lack of roads, steep relief, outcrops, swamps, and rivers, it is not possible to reach every place of the study area and, thus, sampling is constrained to places that can be reached.

Roudier et al. (2012) improved CLH adding to this algorithm a factor that favors the choice of samples in more easily accessible locations, which saves time and financial resources for field work, resulting in the cost-constrained conditioned Latin hypercube (CCLH). In addition to facilitating the field work, the CCLH considers the information related to the property to be mapped in the process of allocation of samples, which tends to increase the sampling representativeness. However, it is known that soil maps generated from a field survey are dependent on the sampling locations and their distribution over the area, which means that different sampling schemes may

result in different soil maps. Thus, a careful analysis have to be performed in order to define the most adequate sampling system to be used in a soil survey, keeping in mind that some extra samples might be necessary to be collected at places besides those determined by the sampling system in order to better capture soils variability across the area.

2.2 Legacy soil data

Legacy soil data are considered important sources of information about soils. They represent available information about soils of an area of interest and their major sources are soil maps and reports of soil surveys. Diverse works have used legacy soil data as source of information to refine soil class and property maps worldwide (BUI; MORAN, 2001; SUN et al., 2011; MALONE et al., 2014).

In Brazil, where most existing soil maps for large areas were created prior to the advent of digital soil mapping and there is a current lack of resources for soil surveys, the use of available information becomes a low cost alternative to obtain data to help to create more detailed soil maps.

Since most available maps were published in a paper-based format, the use of such information is dependent on its transformation into digital data. Thus, using geoprocessing techniques, one may convert those maps to be used in a digital environment. Furthermore, through digital soil mapping techniques, one can retrieve information embedded on the map, which actually reflects the pedologist's mental model used to define the boundaries among soil classes (BUI, 2004). With these techniques, for example, it can be used a digital elevation model and other terrain maps derived from it (e.g., slope, topographic

wetness index, and curvature) to study the relation between soil classes and topographical attributes (MOORE et al., 1993; GESSLER et al., 1995; MCBRATNEY et al., 2003; IWASHITA et al., 2012;. BROWN et al., 2012; MENEZES et al., 2013; HENGL et al., 2015). Thus, it is possible to retrieve the pedologist's mental model and use it for refinement of both the existing soil map and the maps of larger areas under similar soil forming factors (climate, organisms, parent material, relief and time) (JENNY, 1941) to the mapped area, requiring much lesser costs to perform such activities.

REFERENCES

ADHIKARI, K.; KHEIR, R.B.; GREVE, M.B.; BØCHER, P.K.; MALONE, B.P.; MINASNY, B.; MCBRATNEY, A.B.; GREVE, M.H. High-Resolution 3-D Mapping of Soil Texture in Denmark. **Soil Science Society of America Journal, Madison**, v.77, p.860-876, maio 2013.

BROWN, R. A.; MCDANIEL, P.; GESSLER, P.E. Terrain Attribute Modeling of Volcanic Ash Distributions in Northern Idaho. **Soil Science Socienty of America Journal**, Madison, v.76, p.179-. jan. 2012.

BRUNGARD, C.W.; BOETTINGER, J.L. Conditioned Latin hypercube sampling: Optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: BOETTINGER, J.L. et al. (Ed.). **Digital Soil Mapping**: Bridging Research, Environmental Application, and Operation. Springer: Nova Iorque, 2010. p. 67–75.

BUI, E.N. Soil survey as a knowledge system. **Geoderma,** Amsterdã, v.120, p.17–26, maio 2004.

BUI, E.; MORAN, C. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma, Amsterdã, v.103, p.79-94, set. 2001.

COELHO, F.B.; GIASSON, E. Métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. **Ciência Rural**, Santa Maria, v.40, n.10, p.2099-2106, out. 2010.

GESSLER, P.E.; MOORE, I.D.; MCKENZIE, N.J.; RYAN, P.J. Soil-landscape modelling and spatial prediction of soil attributes. **International Journal of Geographical Information Systems**, Londres, v.9, p.421–432, jul. 1995.

HENGL, T.; HEUVELINK, G.B.M.; KEMPEN, B.; LEENAARS, J.G.B.; TAMENE, L.; TONDOH, J.E. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **PLoS ONE**, São Francisco, v.10, p.1–26. jun. 2015.

IWASHITA, F.; FRIEDEL, M. J.; RIBEIRO, G. F.; FRASER, S. J. Intelligent estimation of spatially distributed soil physical properties. **Geoderma**, Amsterdã, v.170, p.1-10, jan. 2012.

JENNY, H. **Factors of soil formation**: A System of Quantitative Pedology. Nova Iorque: McGraw-Hill Book Co. Inc., 1941. 281 p.

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

MALONE, B.P.; MINASNY, B.; ODGERS, N.P.; MCBRATNEY, A.B. Using model averaging to combine soil property rasters from legacy soil maps and from point data. **Geoderma**, Amsterdã, v.232-234, p.34–44, nov. 2014.

MCBRATNEY, A.B.; SANTOS, M.L.M.; MINASNY, B. On digital soil mapping. **Geoderma**, Amsterdã, v.117, n.1/2, p.3-52, nov. 2003.

MENEZES, M.D.; SILVA, S.H.G.; OWENS, P.R.; CURI, N. Digital soil mapping approach based on fuzzy logic and field expert knowledge. **Ciência e Agrotecnologia**, Lavras, v.37, p.287–298. jul. 2013.

MINASNY, B.; MCBRATNEY, A.B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, Oxford, v.32, p.1378–1388. nov. 2006.

MOORE, I.D.; GESSLER, P.E.; NIELSEN, G.A.; PETERSON, G.A. Soil Attribute Prediction Using Terrain Analysis. **Soil Science Socienty of America Journal**, Madison, v.57, n.2, p.443-452. mar. 1993.

MOTTA, P. E. F. et al. **Levantamento pedológico detalhado, erosão dos solos, uso atual e aptidão agrícola das terras de microbacia piloto na região sob influência do reservatório de Itutinga/Camargos, MG**. Belo Horizonte: CEMIG, 2001. 51 p.

MULDER, V.L.; DE BRUIN, S.; SCHAEPMAN, M.E. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. **International Journal of Applied Earth Observation and Geoinformation**, Amsterdã, v.21, p.301–310. abr. 2012.

RESENDE, M. et al. **Pedologia:** base para distinção de ambientes. 6. ed. Lavras: UFLA, 2014. 404 p.

ROUDIER, P.; HEWITT, A.E.; BEAUDETTE, D.E. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: Minasny, B. et al. (Ed.). **5th Global Workshop on Digital Soil Mapping**: Digital Soil Assessments and Beyond (Sydney, Australia). CRC Press: Boca Raton, 2012. p. 227–231.

SUN, X.L.; ZHAO, Y.G.; ZHANG, G.L.; WU, S.C.; MAN, Y.B.; WONG, M.H. Application of a digital soil mapping method in producing soil orders on mountain areas of Hong Kong based on legacy soil data. **Pedosphere**, Nanjing, v.21, p.339-350, jun. 2011.

SECOND PART - ARTICLES

2. ARTICLE 1. Evaluation of conditioned Latin hypercube sampling as a support for soil mapping and spatial variability of soil properties

***Article prepared according to the rules of Soil Science Society of America Journal.**

ABSTRACT

In soil surveys, the number of collected samples is commonly reduced by factors that hamper the field activities, such as rugged terrain and lack of roads. Conditioned Latin hypercube (CLH) sampling scheme has been used to choose the most representative places to be sampled and properly capture soils variability across the landscape, whereas cost-constrained conditioned Latin hypercube (CCLH) limits the sampling to areas of easy access. The objectives of this work were: (a) to compare the efficiency of CLH and CCLH sampling systems to create soil maps, considering the number of soil classes covered per system; (b) to compare both systems to map soil A horizon thickness; and (c) to generate a detailed soil map of the study area to assist in decision makings. The study was carried out in Minas Gerais State, Brazil. A digital elevation model (DEM) and its terrain derivatives were the basis for CLH and CCLH to determine the sampling points. CCLH also required a cost map that represents the difficulty of reaching every place in the area. At the sampling places, soil information was observed, allowing for the creation of those maps that were further validated in the field. Kappa index, global index, RMSE, 1:1 ratio graphic and R^2 were the comparison parameters. CLH presented higher accuracy

than CCLH to represent both soil classes and soil attributes, although the samples were spread out in the area. CCLH was less representative than CLH, but it may contribute to soil sampling in areas of difficult access, requiring less time and investments to accomplish the field works, situations that are common in developing countries, such as Brazil.

Index terms: digital soil mapping, soil prediction, soil sampling, soil survey.

2.1 INTRODUCTION

Soil surveys provide the baseline information for planning and properly utilizing the soil resource. Soil maps allow for the spatial representation, identification and classification of soils across the landscape in order to organize and represent soils into more homogeneous units. Understanding the soil resources provides the basic infrastructure for nations to provide planning for conserving the soil resource, accounting for water storage and transmission.

Soil survey technology has advanced rapidly in the past two decades going from a paper-based mapping process to a digital soil mapping process. In soil surveys, the number of collected samples is commonly constrained by time and cost restrictions to thoroughly visit the area, especially in places where there is a deficiency of roads, dense vegetation and rugged terrain. This situation provides the impetus for more efficient sampling methods that are able to capture the spatial variability of soils and their properties to reduce the number of samples, time and investments needed for fieldwork, however, ensuring a good quality of the final maps.

In the new era of digital soil mapping, there is an increasing number of widely available digital information that aids field activities. Digital Elevation Models (DEMs) are important tools because they provide means for representing the terrain features that have commonly been linked with soil catenas, which are mappable units at varying scales. Some of the terrain attributes derived from a DEM have demonstrated correlations between soil mapping units and slope gradient, curvatures and topographic wetness index (Zhu et al., 1997; Lagacherie and Voltz, 2000; Mendonça-Santos et al., 2007; Ashtekar and Owens, 2013).

Due to the ease access to such kind of information, Digital Soil Mapping (DSM) has emerged as an important set of tools for the production of more detailed soil maps, based on quantitative relationships between soils and environments (McBratney et al., 2003). In addition, to ensure the final quality of the maps, sampling systems that represent the variability of soils across the landscape in combination with the expertise of soil scientists familiar with soil-landscape relationships in the study area can be employed.

One of the sampling methods that has been increasingly used in soil surveys is the so-called Conditioned Latin Hypercube (CLH) (Minasny and McBratney, 2006). This tool is derived from Monte Carlo sampling system, combining the power of stratification, randomness and efficient allocation of samples from multivariate distributions (McKay et al., 1979; Minasny and McBratney, 2006). Thus, from a set of information that is related to the soil property to be mapped, points are chosen at the sampling locations possibly more representative of the soil variability over the area.

The CLH assumes that the locations to be sampled must actually exist on the landscape (Brungard and Boettinger, 2010) and the tool produces a representative distribution of points according to the number of possible points

for sampling. Minasny and McBratney (2006) compared the efficiency of simple random sampling, stratified sampling and spatial CLH and concluded that the latter showed the best results. Xu et al. (2005), by combining the stochastic simulation with the CLH, found that this latter captured greater variability of the area than the simple random sampling, particularly when the sampling density (samples / unit area) was low.

Although the CLH has proven to be very efficient in selecting sampling locations, some study areas are difficult to access, especially in tropical countries, such as Brazil, due to lack of roads and dense vegetation. These factors culminate in a difficulty to use the CLH in certain areas. In this context, Roudier et al. (2012) have proposed an alternative to improve this sampling scheme by conditioning the locations chosen for sampling according to the difficulty (cost) that one may face to reach that place, being named Cost-constrained Conditioned Latin Hypercube (CCLH). This system takes into account factors that hinder or even make impossible sampling in certain places, such as distance from roads, slope gradient, vegetation, water courses, and so forth, characterizing them as being of "high cost". Thus, the sampling scheme prioritizes sites that are easy to access still taking into account the variability of attributes that may influence soil properties to make a representative sampling of the study area. Besides, a comparison between these two sampling systems applied in a soil survey for tropical conditions is needed.

With the advent of these new sampling technologies for digital soil mapping, the objectives of this work were: (a) to compare the efficiency of Conditioned Latin Hypercube and Cost-constrained Conditioned Latin Hypercube sampling systems to create soil maps, considering the number of different soil classes covered by each system, to validate both of them in the

field and to assess their effectiveness; (b) to compare the sampling systems to map soil A horizon thickness; and (c) to generate a detailed soil map of the study area to assist in finding out the most appropriate use and management for each segment of the landscape.

2.2 MATERIALS AND METHODS

2.2.1 Study Area

The study area is located in Bom Sucesso county, Minas Gerais State, Brazil, at coordinates 21°06'50"S and 44°49'22"W, with altitudes ranging from 858 to 890 m, and average slope of 25% (Figure 1). The climate is Cwb, according to the Köppen classification system, represented by cold and dry winters and warm and rainy summers, with average annual rainfall of 1,500 mm. The mapped location covers an area of 1.6 ha which has been used with Australian cedar (*Toona ciliata*) plantation for 2 years. Prior to Australian cedar, the area had been characterized as a degraded pasture of *Urochloa decumbens*.

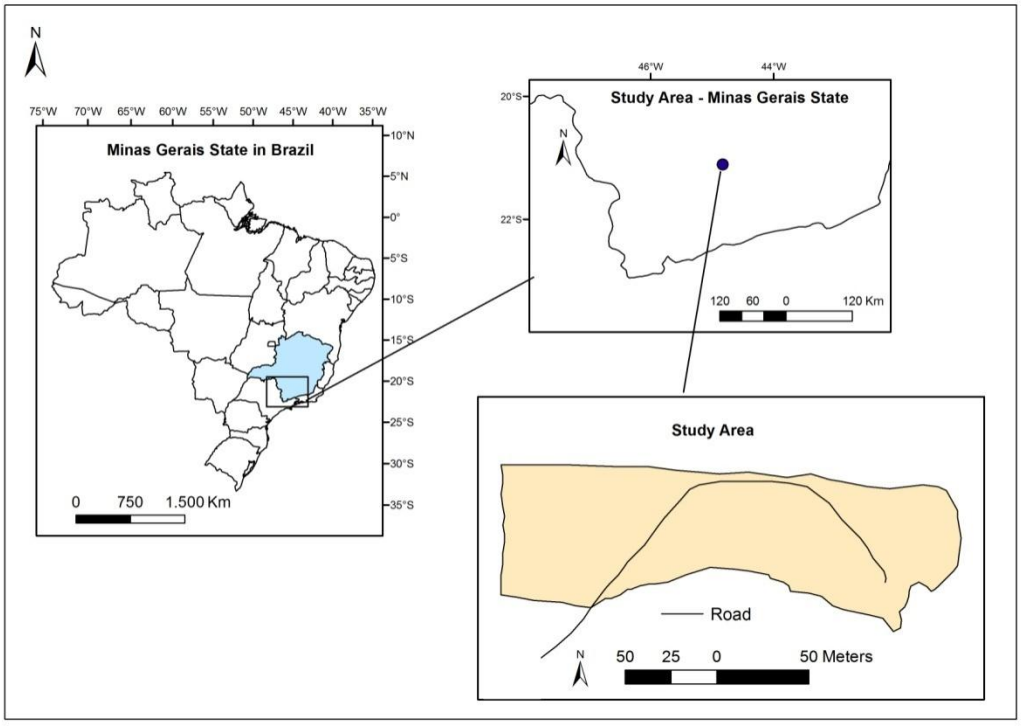

Figure 1 - Local of the 1.6 ha study area in Minas Gerais, Brazil.

2.2.2 Sampling Systems

The overall goal of the research was to compare the efficiency of Conditioned Latin Hypercube (CLH) and Cost-constrained Conditioned Latin Hypercube (CCLH) systems evaluating the gain of information for mapping both soil classes (detailed soil survey) and a soil property (A horizon thickness), and the reduction of the time required for the fieldwork. Through the detailed soil survey, it was possible to compare the number of soil classes contemplated

by each sampling system and to retrieve information about A horizon thickness, in order to also compare both systems to map a soil property. The thickness of the A horizon was selected for comparison because this soil property can serve as initial benchmark considering it can be a highly variable soil property over an area. Furthermore, it generally reflects the influence of both management and terrain attributes (Zhu et al., 1997). For that, two sampling schemes based on the aforementioned sampling systems were carried out.

Prospections were performed in the field to compare the CLH with the CCLH to map soils at a detailed scale. Their location were defined by using the software R (R Development Core Team, 2009) and clhs package (Roudier, 2012). The CLH locations were determined based on the method proposed by Minasny and McBratney (2006), derived from Latin Hypercube proposed by McKay et al. (1979). While the CLH provides sampling point locations throughout the study area, based on the variability of the terrain attributes employed to predict the soil properties to be mapped, CCLH conditions the samples for easy-to-access sites, identified by a "cost" raster, but still taking into account the variability of the terrain attributes in this task.

The terrain attributes used by both sampling systems to choose the locations to be sampled were elevation, Digital Elevation Model (DEM) created from contour lines from planialtimetric survey of the study area, slope and SAGA wetness index (Figure 2). The last two terrain attributes were created in SAGA GIS (Böhner et al., 2006) and were derived from the aforementioned digital elevation model.

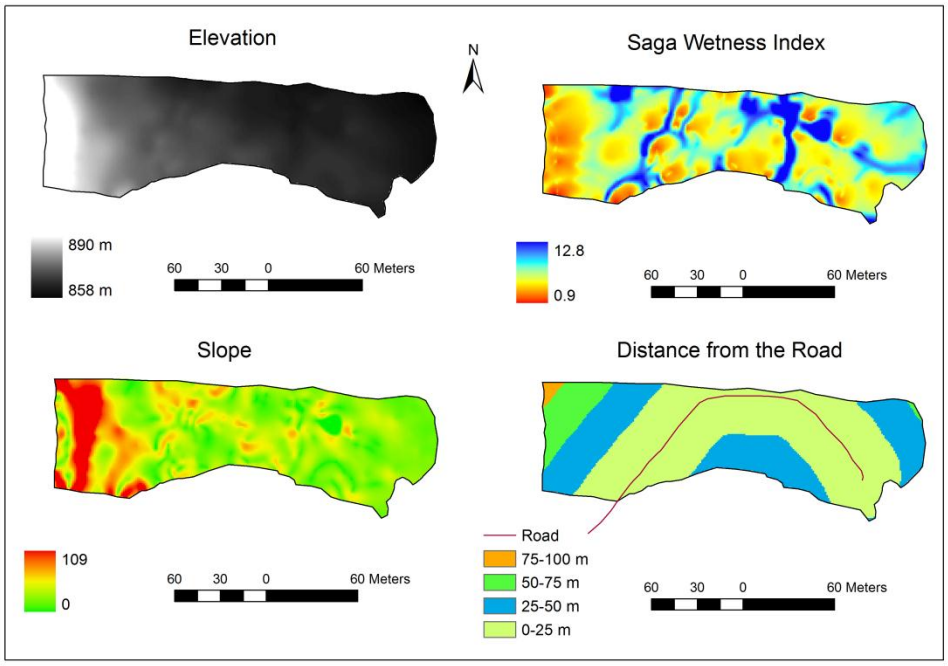

Figure 2 - Terrain attributes used to select the points to be sampled by CLH and CCLH, and distance from the road, which, in association with the slope, will originate the cost raster that constrains the sampling sites in the CCLH to easy-to-reach places.

The cost raster necessary to discriminate sites of easy and difficult access was generated from the slope and distance from the road rasters (Fig. 2), and the latter was calculated using the Euclidean Distance function in ArcGIS software (ESRI). Values (weights) were assigned to these two rasters, reclassifying their original values in order to represent the difficulty of accessing each local on the landscape, as shown in Table 1. Accordingly, to each pixel of these two raster layers it was assigned a value (weight) and a pixel-by-pixel

addition of these two rasters was performed to create the cost raster, representing the difficulty of reaching the location of each pixel on the landscape (Figure 2). For example, a location that is 60 m away from the road (weight 5) and on a hill of 30% slope (weight 7) will have a final cost of 12 (5 +7), more difficult to access than a weight 2 place (flat and close to roads).

Table 1 - Weights assigned to each class of raster values employed on the creation of the cost raster.

Distance from roads (m)	Weight	Slope gradient (%)	Weight
0-25	1	0-3	1
25-50	3	3-8	3
50-75	5	8-20	5
75-100	7	20-45	7
> 100	9	>45	9

To quantify the representativeness of the two sampling systems, descriptive statistics analyses were performed to describe the variability of the terrain attributes for the entire study area and these parameters were compared with those values assessed by each sampling system.

2.2.3 Soil Mapping

The soil survey was conducted through investigations within the whole area, with 12 trenches whose places were chosen by each of the sampling

systems, CLH or CCLH, and description of three modal profiles and collection of soil samples, as proposed by Santos et al. (2013). Twelve trenches for a soil survey of a 1.6 ha area are considered to be sufficient to generate a detailed soil map according to Normative Procedures for Pedological Survey (EMBRAPA, 1995) and Brazilian Pedology Technical Manual (IBGE, 2007), which are the guidance books for tropical conditions for such activity. According to those books, a detailed soil survey must contain from 0.2 to 4 observations per hectare. However, this study performed 7.5 observations per hectare, making up a total of 12 observations, in order to improve the comparisons of soils variability captured by each sampling system. Soils were classified according to the Brazilian Soil Classification System (Embrapa, 2013) and US Soil Taxonomy (Soil Survey Staff, 1999).

Two spatially explicit detailed soil maps were created using the field information obtained by each sampling system. In this procedure, pedological mapping units (PMUs) were created, according to Santos et al. (2014) standards, using, in addition to soil order and suborder levels, the soil fertility (dystrophic or eutrophic, base saturation <50% and ≥50%, respectively), A horizon type, soil texture, presence or absence of gravels, native vegetation, and relief phase because these are some factors that can influence crop development and aid decision makers in regard to soil management and installation of future scientific experiments.

The soil maps created from each sampling system were validated in the field, with 13 validation points (additional observations) chosen at places where the maps differed most. To find out the sampling system that best accessed the soils variability in the area, it were calculated both the global index (ratio between the number of corrected classified and the total number of samples) and

Kappa index, calculated through a confusion matrix taking into account the number of classes and the proportion between the correctly classified samples and the total number of samples (Congalton and Green, 2009). With the information from the two sampling systems, a final soil map of the study area was created using all the sampled points to support the choices of adequate sites for installing future scientific experiments.

2.2.4 Mapping of soil A horizon thickness

From the soil survey information (CLH and CCLH sampling schemes), A horizon thickness data was retrieved in order to create spatial maps of this soil property, according to each sampling system, by the inverse distance weighting (IDW) interpolation. Kriging and splines were tested as interpolation methods, but kriging did not have an adequate adjust and splines presented less accuracy than IDW. The assumption to predict a value for any non-sampled location is that a measured point has a local influence that decreases with distance. The values at non-sampled points are estimated using linear combination of values at the sampled points, weighted by an inverse function of the distance from the point of interest to the sample points. The weights (λ_i) can be expressed as:

$$\lambda_i = \frac{\frac{1}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

where d_i is the distance between x_0 and x_i, p is a power parameter and n represents the number of sampled points used for the estimation (Li and Heap,

2008). This interpolation was carried out in ArcGIS 10.0 (Esri), where the power parameter equals to 2 (default) was chosen.

To define the sampling method that best represented the A horizon thickness spatial variability along the area, those maps were validated in the field at 13 sites (additional observations). Statistical analysis was performed based on the root mean square error (RMSE) (formula presented below), graphic 1:1 between real (field) and estimated values, and R^2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ei - mi)^2}$$

where: *n* is the number of observations (12), *ei* is the estimated value of the A horizon thickness and the *mi* is the real value of the A horizon thickness.

2.3 RESULTS AND DISCUSSION

2.3.1 Differences Between the Sampling Systems

Samples were allocated using the CLH and the CCLH methods. Analyzing the allocation of points for both sampling systems, the CCLH samples were placed close to roads and in easy-to-access sites, which means that the low-cost sampling locations were achieved. Silva et al. (2014), using CCLH in a watershed of difficult access in Minas Gerais State, Brazil, found that the samples had been allocated in places relatively easier to access and the representativeness of soil properties was considerably contemplated. On the other hand, the CLH samples were well distributed over the whole area (Figure 3). Minasny and McBratney (2006) noted that the CLH, as well as the spatial

stratified sampling, showed adequate sample point coverage of the study area demonstrating that this is a promising system to access the sampling variability using ancillary data (terrain attributes).

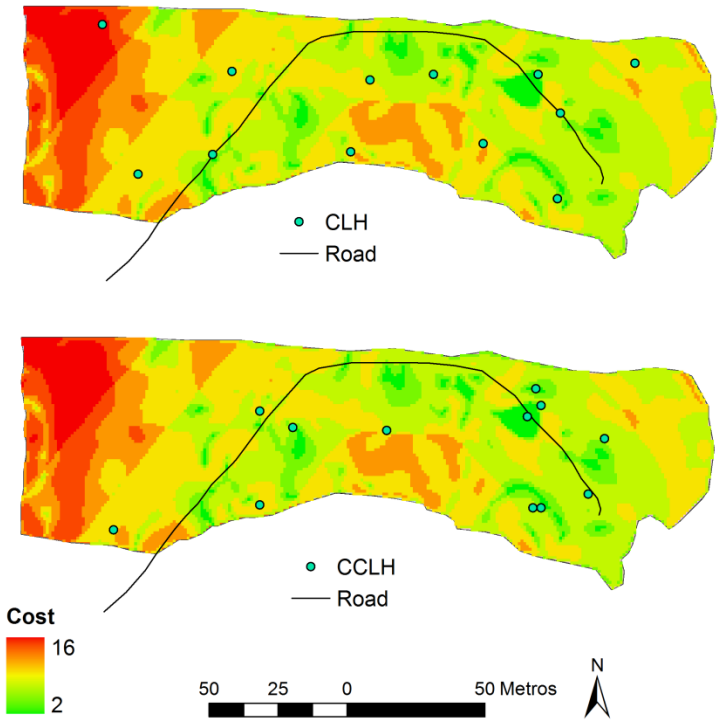

Figure 3 - Points allocation through both CLH and CCLH sampling systems over the cost raster that shows the difficulty (cost) to reach every place on the landscape.

By analyzing the variability of the terrain attributes employed in the sampling systems, it was verified that, in general, the CLH compared with CCLH, both with only 12 sampling points, showed values of mean, standard deviation and median closer to the ones calculated considering all the pixels of

the study area (Table 2). These results indicate that the CLH, with a reduced number of samples, was able to better represent the variability of the terrain attributes, and, hence, the soil properties. Minasny and McBratney (2006), comparing the CLH with simple random sampling and with stratified spatial sampling also found out that the CLH, with the same number of samples, better accessed the variability of the attributes used for defining the sampling places.

Table 2 - Comparison between CLH and CCLH sampling systems through terrain attributes data for the entire study area.

System	Sampled Pixels	25% Quartil	Median	75% Quartil	Mean	Standard Deviation
----------Slope gradient----------						
All pixels	15919	10.68	16.89	26.96	21.65	16.707
CCLH[1]	12	5.74	11.94	19.11	13.56	11.495
CLH[2]	12	11.01	16.71	26.58	20.85	15.563
----------Elevation----------						
All pixels	15919	862.5	865	870.9	868.1	7.682
CCLH	12	862.6	864.1	867.9	865.8	4.815
CLH	12	862.9	864.2	868.8	866.1	4.926
----------SWI----------						
All pixels	15919	3.434	4.304	5.615	4.73	1.876
CCLH	12	3.421	4.621	5.172	6.531	2.54
CLH	12	3.77	4.82	5.667	4.938	1.868

[1]Cost-constrained Conditioned Latin Hypercube; [2]Conditioned Latin Hypercube. SWI - Saga Wetness Index.

2.3.2 Mapping Soils Through CLH and CCLH

The spatially explicit detailed soil maps were created in order to compare the efficiency of both sampling systems involving various soil properties, since soil maps provide further basis to support planning and decisions than maps of just one property. These maps are shown in Figure 4.

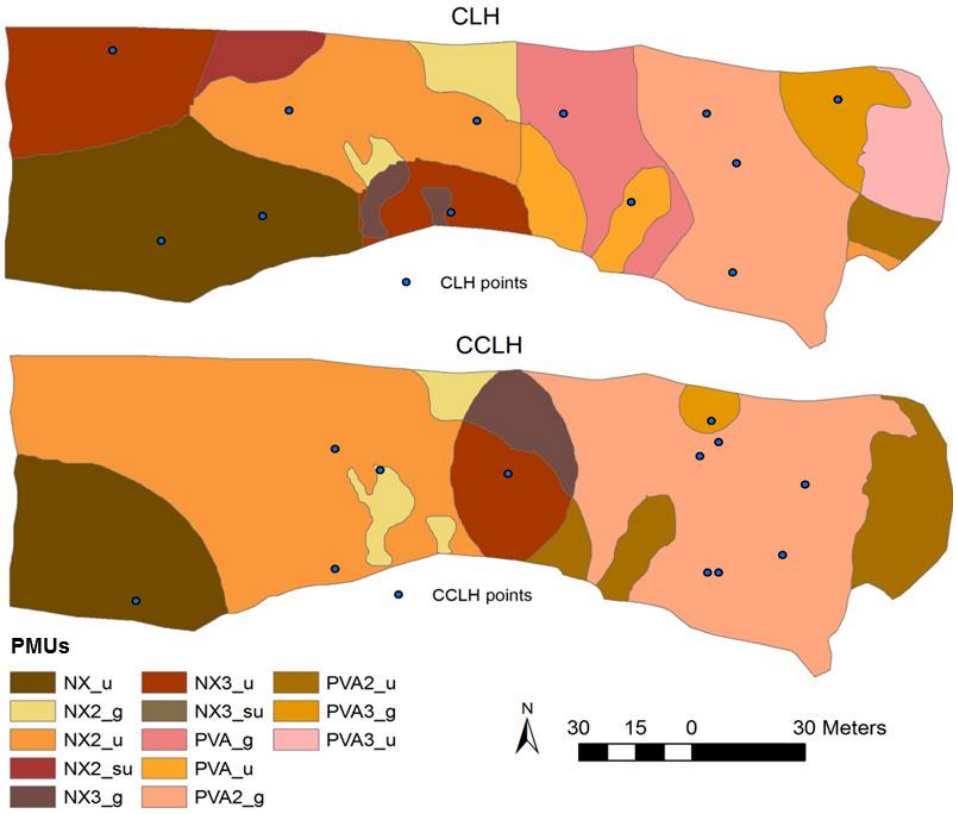

Figure 4 - Soil maps generated with a support of CLH and CCLH.

According to the soil map created with support of the sampling locations indicated by the CLH, 13 pedological mapping units (PMUs) were identified, against 8 by CCLH (PMUs discussed later), being none of these latter different from the ones found by CLH. It demonstrates that CLH was able to capture a greater variability of soil properties than CCLH due to its characteristic of allocating the samples with no cost constrain.

In addition, to validate the maps, confusion matrices were used to calculate the global index (GI) and Kappa index. The soil maps with support of CLH showed GI of 84.6% (10 correctly classified samples out of 13) and Kappa index of 81.2, which corresponds to an excellent classification, according to Landis and Koch (1977). In contrast, the map created from CCLH samples presented GI of 69.2% (9 correctly classified samples) and Kappa index of 60.6, equivalent to a substantial classification, as proposed by Landis and Koch (1977). These results confirm that the sampling systems influenced the resultant final maps, which may probably lead to inappropriate uses of soils where the information shown on the maps differs from the reality.

2.3.3 Mapping of Soil A Horizon Thickness Through CLH and CCLH Systems

In both maps generated according to the sampling systems, the higher and steeper areas (Figure 2) presented thin horizon thickness (Figure 5). In most gentle slope places, A horizon is often thicker than those of the steepest locations. These results are in agreement with Świtoniak (2014), who found out that the most intensive removal of topsoil has occurred in the highest, steep and convex parts of the hill. However, on the map generated from the CLH there is a

region with estimated thicker A horizon (around 40 cm of thickness) than the same region on the map generated from the CCLH (around 20 cm). This difference has probably occurred due to the sampling by CLH covers areas of more difficult access, which were not contemplated by the CCLH sampling sites.

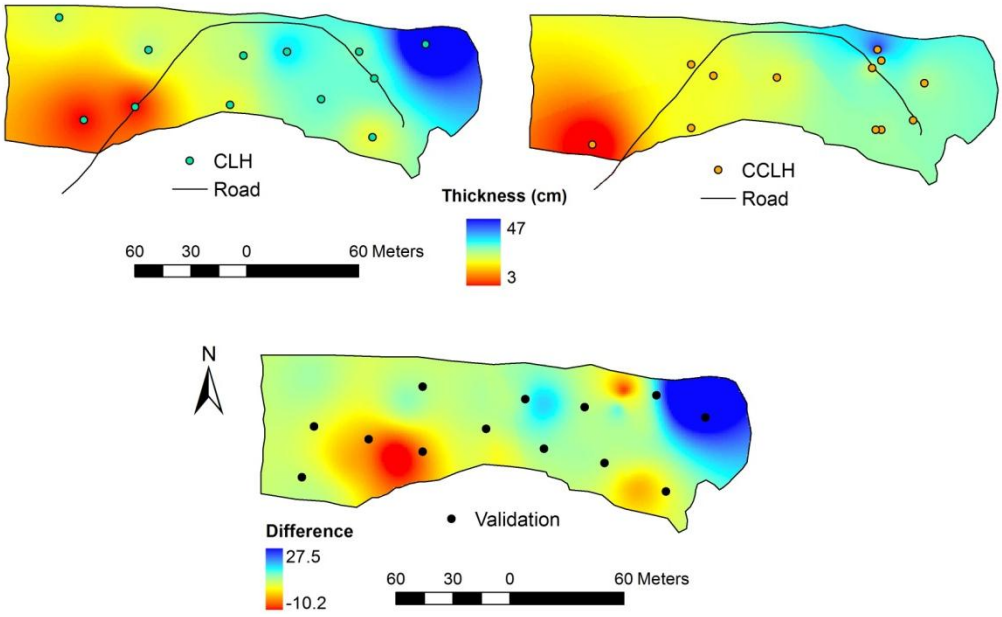

Figure 5 - Soil A horizon thickness maps per sampling system and the validation points highlighting the places where the A horizon thickness values were more divergent between both maps.

Figure 5 also represents a map of the difference between the two others originated from each sampling system. It is noted that these differences (in red and blue on the map) are mainly explained by the presence of samples at these

sites in only one of the sampling systems. This fact led to different estimates of the thickness values by IDW, which takes into account the distance among observations to spatialize information (to estimate data for the non-observed places based on the observed places). Thus, as the distance from a certain observation site to a non-observed site, whose information will be estimated, increases, the similarity between the characteristics of those locales decreases, thus being the estimates sensitive to distance (Webster and Oliver, 2007).

Through the field validation points to assess the accuracy of each map generated by the two sampling systems, the real values of the A horizon thickness were obtained and compared with the estimated values for their respective locations (Figure 6). CLH information spatialized by IDW method provided the greatest power to represent the data (R^2 = 0.902) and the lowest RMSE (5.65) in relation to CCLH (R^2 = 0.539; RMSE = 7.952). Furthermore, it was observed that the values of real and estimated thickness by IDW based on CLH are closer to 1:1 ratio than in the corresponding graph for CCLH (Figure 6), constituting a lower error across the whole range of observed values. On this graphic for CCLH, it can be noticed that for values of A horizon thickness smaller than 20 cm, there was an overestimation of this soil property and for values greater than 20 cm, there was an underestimation of values.

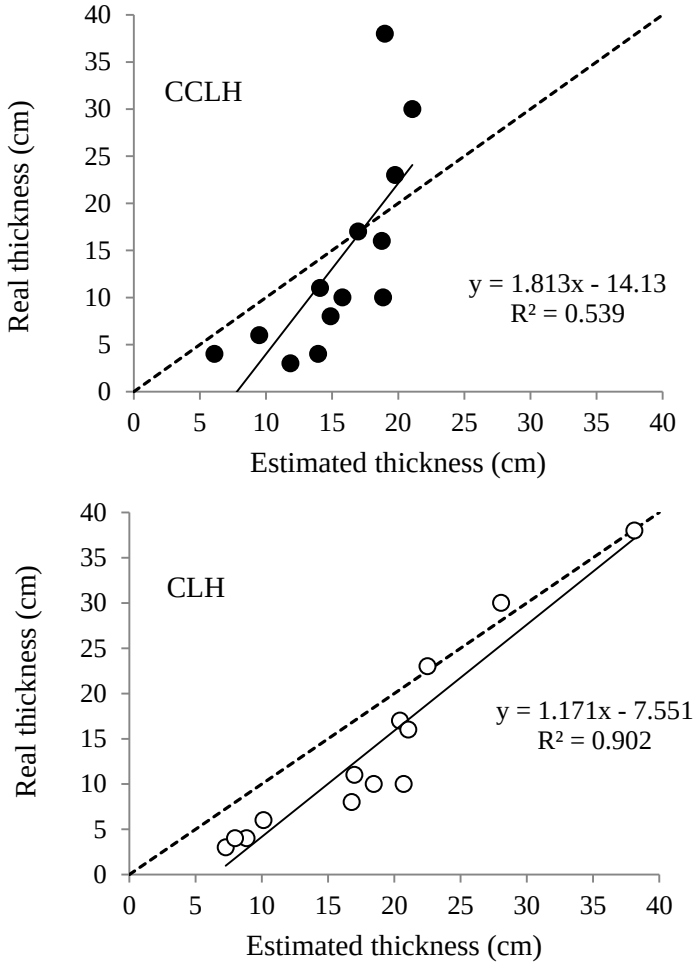

Figure 6 - 1:1 ratio for real and estimated A horizon thickness obtained by means of CCLH (above) and CLH (below).

According to the data in Figure 6, CLH tends to show better results from a practical point of view, for both aiding decision-making on the most

appropriate management for each segment of the landscape and directing strategies for practices of soil and water conservation, especially where the A horizon is thin. This detailed information is critical, particularly in Brazil, where, due to the intense weathering-leaching processes suffered by most soils, organic matter is the most important fraction in generation of soil charges, due to its direct relationship with effective CEC and potential yield of soils (Goedert, 1983; Lopes and Cox, 1977).

Furthermore, assessing the variability of the terrain attributes included in the sampling for each system (Table 2), it was found that the CCLH encompassed a smaller range of values of slope, elevation and wetness index than the CLH. Thus, analyzing the statistical parameters representing the variability of data on the thickness of A horizon obtained by each system, both the coefficient of variation (CV) and standard deviation (SD) were higher for the CLH, by not having any restriction in regard to the difficulty of sampling in some locations on the landscape (Table 3). This resulted, for this system, in greater representativeness of the variability of the terrain attributes that tend to influence soil properties, which is reflected in the maps of A horizon thickness made for each sampling system (Figure 5). However, there was no statistical difference in the mean values of the A horizon thickness in both systems (p value = 0.577, CV = 50.98), by analysis of variance with sampling system as factor of variation (Table 3).

Table 3 - Variability and comparison between the sampling systems to assess soils A horizon thickness.

	Mean (cm)	Standard Deviation (cm)	CV (%)
CCLH[1]	16.67a	6.01	36.03
CLH[2]	18.75a	11.27	60.09

[1]Cost-constrained Conditioned Latin Hypercube; [2]Conditioned Latin Hypercube. Means followed by the same letter in column do not differ statistically according to Scott Knot test at 5% of probability (CV=50.98% and n=24). Standard deviation and coefficient of variation were calculated per sampling system (n=12).

2.3.4 Final Detailed Soil Map

From the field information provided by the two sampling systems, a final soil map of the study area was generated (Figure 7). Two soil classes at suborder level were found, Haplic Nitosols (NX) and Red-Yellow Argisols (PVA) (correspondent to Ustalfs and Ustults in US Soil Taxonomy, respectively), occurring in distinct positions on the landscape. While NX appears in high and sloping segments, PVA is located in the lower portions of the landscape with gentle topography.

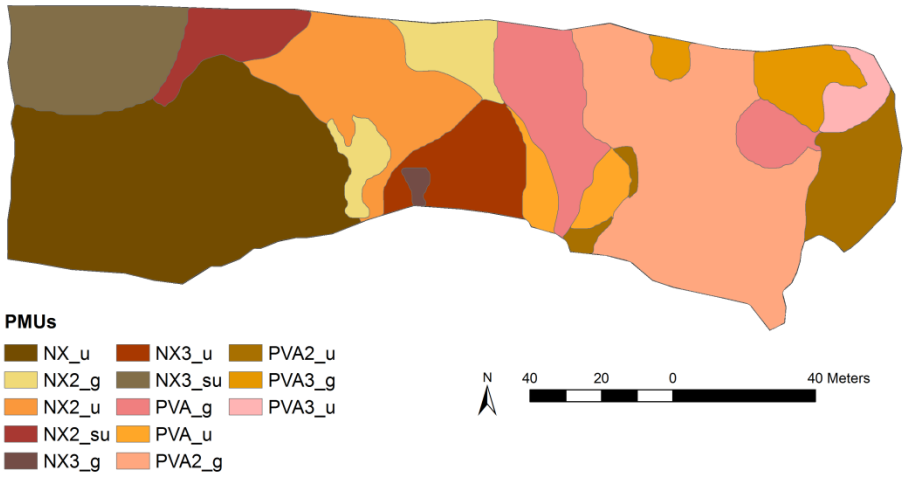

Figure 7 - Soil map of the study area at ultra-detailed scale. PMUs description in Table 4.

Both soil classes at the suborder level were divided into 13 PMUs according to the type of A horizon, presence or absence of gravel, fertility, texture, native vegetation, and relief phase (Table 4). The PMU NX on undulate (8-20% gradient) slope (NX_u) occupies 27% of the study area, followed by PVA2 on gentle slope (0-8% gradient) (PVA2_g), covering 21.1% of the area. It was also noted that both PMUs contain gravel, which can be taken as a common soil property in the study area. The fact that the PMU NX_u is found on the highest regions of the landscape and on undulate relief tends to favor the natural erosion of these soils, which is associated with the presence of weak (thin) A horizon at these sites.

Table 4 - Pedologic mapping units (PMUs) found in the study area.

PMU_relief	PMUs description	Area (%)
NX_u	NX eutrophic, weak A, clayeytexture, gravelly, semiperennial tropical forest, undulated relief	27.0
NX2_g	NX eutrophic, moderate A, clayey texture, semiperennial tropical forest, gentle relief	4.2
NX2_u	NX eutrophic, moderate A, clayey texture, semiperennial tropical forest, undulated relief	9.2
NX2_su	NX eutrophic, moderate A, clayey texture, semiperennial tropical forest, strongly undulated relief	3.2
NX3_g	NX eutrophic, moderate A, clayey, gravelly, semiperennial tropical forest, gentle relief	0.5
NX3_u	NX eutrophic, moderate A, clayey texture, gravelly, semiperennial tropical forest, undulated relief	5.2
NX3_su	NX eutrophic, moderate A, clayey texture, gravelly, semiperennial tropical forest, strongly undulate relief	8.4
PVA_g	PVA dystrophic, moderate A, sandy clay loam, semiperennial tropical forest, gentle relief	8.0
PVA_u	PVA dystrophic, moderate A, sandy clay loam, semiperennial tropical forest, undulated relief	2.6
PVA2_g	PVA dystrophic, moderate A, sandy clay loam, gravelly, semiperennial tropical forest, gentle relief	21.1
PVA2_u	PVA dystrophic, moderate A, sandy clay loam, gravelly, semiperennial tropical forest, undulated relief	5.4
PVA3_g	PVA dystrophic, prominent A, sandy clay loam, gravelly, semiperennial tropical forest, gentle relief	3.6
PVA3_u	PVA dystrophic, prominent A, sandy clay loam, gravelly, semiperennial tropical forest, undulated relief	1.6

NX = Haplic Nitosol; PVA= Red-Yellow Argisol; g - gentle relief(0-8% slope);

u - undulated (8-20% slope); su - strongly undulated relief (>45% slope).

The final soil map (Figure 7) was related to the A horizon thickness maps developed with support of the two sampling systems (Figure 8). It was noted that the A horizon thickness map generated by the support of CLH is closer to the PMUs data present in the soil map than the one generated according to CCLH. Only at the boundaries of the PMUs there were major differences between the types of A horizon. However, as the soils and their properties vary gradually along the landscape as a continuum, the boundaries between PMUs represent only transition trends of the properties, not the real (field) boundaries between them.

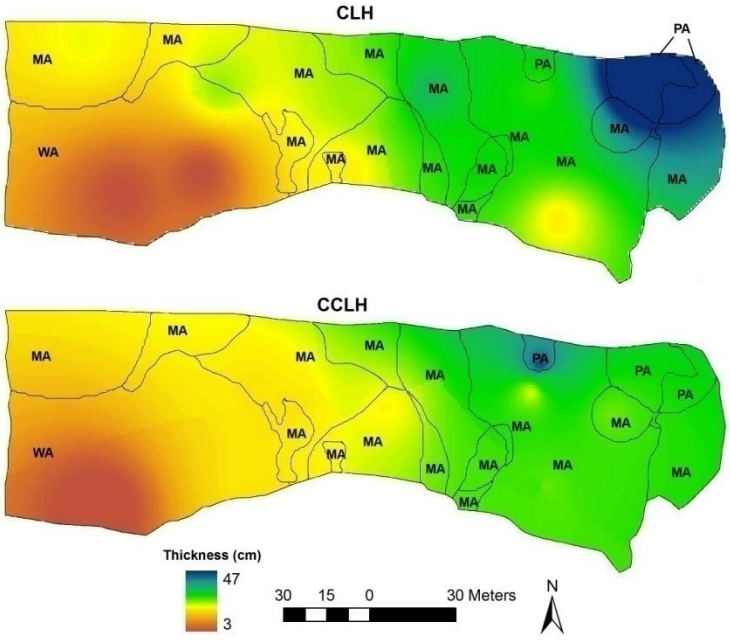

Figure 8 - PMUs limits of the soil maps over the A horizon thickness maps generated according to CLH and CCLH information. WA = weak A; MA = moderate A; PA = prominent A.

By the fact that information obtained through the CLH is more representative of the variability of soils across the landscape, the soil class (or soil property) maps generated with support of this sampling system would allow for more accurate definitions of the potential uses of soils for each section of the landscape. This fact was also obtained by Rad et al. (2014) that employed CLH to map soils of 85,000 ha in Iran and concluded that a few samples were sufficient to capture great variability of soils and that the use of environmental variables as basis for CLH determine the sampling locales presented high correspondence with soil property variability.

In contrast, the fact that the samples have been allocated throughout the area by CLH could turn into a difficulty because hilly and distant areas were covered, requiring more time and investments to conduct the field work. An aspect that diminished this limiting condition of sampling throughout the area was the fact of this study had been carried out in a reference area (Favrot, 1989), which means it is representative of the soils occurring on that region of study, where detailed soils surveys can be properly performed (Lagacherie et al., 2001). According to the field experience of the authors in tropical conditions, soil surveying in large regions normally includes areas of difficult or even impossible access, such as mountainous and/or swampy areas, which would not allow for a detailed sampling through the whole region. In this sense, the CCLH, although not being as representative as the CLH (Roudier et al., 2012), facilitated the fieldwork and also allocated the samples based on the variability of the terrain attributes. Mulder et al. (2013) used conditioned Latin hypercube constrained to places of easy access and found out that this sampling scheme could adequately represent variability of soil properties with small number of samples, although spatial correlation could not be found, being characterized as

a time and cost efficient scheme. This latter consideration consists an important aspect, especially for tropical conditions, where limited funds for field work and difficulties of access are quite common (Menezes et al., 2013), which commonly hamper or even makes it impossible to collect soil samples throughout the study area (Silva et al., 2014; Cambule et al, 2013).

Finally, the authors do not know any study that employed a comparison between CLH and CCLH sampling systems in such comprehensive way to soil survey, which represents a practical alternative mainly for tropical conditions, where fundings for those activities are scarce (Ker et al., 2012).

2.4 CONCLUSIONS

Soil maps are necessary for infrastructure and best management practices for land-use. Soil surveys are expensive and time consuming because field evaluation and sampling are one of the most costly portions of the project. More low cost and efficient sampling methods are necessary while maintaining the statistical rigor. This research indicates that the CLH has better spatial representation of the variability of the A horizon thickness and enables greater accuracy in separating the pedological mapping units in the study area than CCLH. However, the CCLH, despite revealing lower accuracy than the CLH, reduces the time and investment required in the field work. A cost-benefit analyses may be necessary to identify the best method to use for an initial soil survey.

2.5 REFERENCES

Ashtekar, J.M. and P.R. Owens. 2013. Remembering Knowledge: An Expert Knowledge Based Approach to Digital Soil Mapping. Soil Horizons, p.1-6.doi:10.2136/sh13-01-0007

Böhner, J., McCloy, K.R., Strobl, J. (2006): SAGA – Analysis and Modelling Applications. GöttingerGeographischeAbhandlungen, v.115, 130p.

Brungard, C.W. and J.L. Boettinger. 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In: J.L. Boettinger et al. (eds.), Digital Soil Mapping, Progress in Soil Science 2, Springer Neatherlands, Dordrecht, Neatherlands. p.67-75. doi:10.1007/978-90-481-8863-5_6

Cambule, A.H., D.G. Rossiter, J.J. Stoorvogel. A methodology for digital soil mapping in porrly-asscessible areas. Geoderma, 192:341-353. doi: 10.1016/j.geoderma.2012.08.020

Congalton, R.G., and K. Green.2009. Assessing the accuracy of remotely sensed data:Principles and practices. 2nd ed. CRC Press, Boca Raton, FL.

Embrapa – Empresa Brasileira de Pesquisa Agropecuária. 2013. Sistema Brasileiro de Classificação de Solos. 3rd ed. Rio de Janeiro, 353 p.

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

Goedert, W.J. 1983. Management of the Cerrado soils of Brazil: a review. J. Soil Sci. 34:405–428. doi: 10.1111/j.1365-2389.1983.tb01045.x

Ker, J.C., N. Curi, C.E. Schaefer, and P. Vidal-Torrado. 2012. Pedology. Brazilian Soil Science Society (SBCS), Viçosa, 343p.

Lagacherie, P., J.M. Robbez-Masson, N. Nguyen-The, and J.P. Barthe`s. 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma. 101:105–118. doi: 10.1016/S0016-7061(00)00101-4

Lagacherie, P. and Voltz, M. 2000. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma. 97:187-208. doi: 10.1016/S0016-7061(00)00038-0

Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics. 33(1):159–174. doi: 10.2307/2529310

Li, J. and Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record, 137 p.

Lopes, A.S. and F.R. Cox. 1977. A Survey of the Fertility Status of Surface Soils Under “Cerrado” Vegetation in Brazil. Soil Sci. Soc. Am. J. 41:742–747. doi:10.2136/sssaj1977.03615995004100040026x

McBratney, A.B., M.L. Mendonça-Santos, and Minasny, B. 2003.On digital soil mapping. Geoderma. 117:3-52. doi: 10.1016/S0016-7061(03)00223-4

McKay, M.D., R.J. Beckman, and W.J. Conover. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics.21:239–245. doi: 10.2307/1268522

Mendonça-Santos, M.L., A.B. McBratney, and B. Minasny. 2007. Soil prediction with spatially decomposed environmental factors. In: Lagacherie, P., A.B. McBratney, and M. Voltz (Eds.). Developments in Soil Science, 31:269-278. doi: 10.1016/S0166-2481(06)31021-5

Menezes, M.D., S.H.G. Silva, P.R. Owens, N. Curi. 2013. Digital soil mapping approach based on fuzzy logic and field expert knowledge. Cienc. agrotec., 37(4):287-298. doi: 10.1590/S1413-70542013000400001

Minasny, B., and A.B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32:1378–1388. doi:10.1016/j.cageo.2005.12.009

Mulder, V.L., S. de Bruin, and M.E. Schaepman. 2013. Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. Int. J. Appl. Earth Obs. 21:301-310. doi: 10.1016/j.jag.2012.07.004

R Development Core Team. 2009. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http:// www.r-project.org (accessed 30 April 2014). Roudier, P. clhs: a R package for conditioned Latin hypercube sampling. http://cran.r-project.org/web/packages/clhs/clhs.pdf. 2012.

Rad, M.R.P., N. Toomanian, F. Khormali, C.W. Brungard, C.B. Komaki, and P. Bogaert. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma. 232-234:97-106. doi: 10.1016/j.geoderma.2014.04.036

Roudier, P. 2012. clhs: a R package for conditioned Latin hypercube sampling. http://cran.r-project.org/web/packages/clhs/clhs.pdf. Accessed 20 April 2014.

Roudier, P., D.E. Beaudette, and A.E. Hewitt. 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: Minasny, B, B.P. Malone, A.B. McBratney (eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia, 10-13 April 2012. CRC Press, Boca Raton, FL. 482p.

Santos, R.D., R.C. Lemos, H.G. Santos, J.C., Ker, L.H.C. Anjos, and S.H.Shimizu. 2013. Manual de descrição e coleta de solo no campo. 6a.ed. Viçosa, SBCS, 100p.

Santos, W.J.R., N, Curi, S.H.G., Silva, S. Fonseca, E. Silva, and J. J Marques. 2014. Detailed soil survey of an experimental watershed representative of the Brazilian Coastal Plains and its practical application. Ciênc. agrotec., 38(1):50-60. doi: 10.1590/S1413-70542014000100006

Silva, S.H.G., P.R. Owens, M.D. Menezes, W.J.R Santos, N. Curi. A Technique for Low Cost Soil Mapping and Validation using Expert Knowledge on a Watershed in Minas Gerais, Brazil. Soil Sci. Soc. Am. J.78(4)1310-1319.doi: 10.2136/sssaj2013.09.0382

Soil Survey Staff. 1999. Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys. 2ed. USDA-SCS, Washington, DC, USA. Agriculture Handbook No. 436. 871p.

Świtoniak, M. 2014. Use of soil profile truncation to estimate influence of accelerated erosion on soil cover transformation in young morainic landscapes, North-Eastern Poland. Catena 116:173–184. doi: 10.1016/j.catena.2013.12.015

Webster, R and M.A. Oliver.2007. Geoestatistics for environmental scientists. 2nd Edition. John Wiley & Sons Ltd, Cornwall, England. 315p. doi: 10.1002/9780470517277

Xu, C.G., H.S. He, Y.M. Hu, Y. Chang, X.Z. Li, and R.C. Bu. 2005. Latin hypercube sampling and geostatistical modeling of spatial uncertainty in a spatially explicit forest landscape model simulation. Ecol. Model., 185(2–4):255–269. doi: 10.1016/j.ecolmodel.2004.12.009

Zhu, A.X., L. Band, R. Vertessy, and B. Dutton. 1997. Derivation of soil properties using a soil land inference model (SoLIM). Soil Sci. Soc. Am. J. 61(2):523-533. doi: 10.2136/sssaj1997.03615995006100020022x

3. ARTICLE 2. 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

***Article prepared according to the rules of Geoderma.**

ABSTRACT

Diverse projects are being carried out worldwide focusing on development of more accurate soil maps and one of the most valuable sources of data are the existing soil maps. This work aimed to (i) compare two data mining tools, KnowledgeMiner and decision trees, to retrieve legacy soil data from a detailed soil map, (ii) to create and validate the predicted soil maps in the field with the objective to identify the best method for modeling and refining soil maps, (iii) extrapolating soils information to the surrounding similar areas and (iv) to assess the accuracy of this soil map. The study was carried out in Minas Gerais state, Southeastern Brazil. From a detailed soil map, information of 12 terrain attributes was retrieved from the entire polygon of each mapping unit of the map (MUP) and from a circular buffer around the sampled points (CBP). KnowledgeMiner and decision trees were employed to retrieve information per soil class and soil maps were created per method. A field validation of 20 samples was chosen by a cost-constrained conditioned Latin hypercube sampling scheme and the accuracy of all maps was assessed using a global index, Kappa index, and errors of omission and commission. The KnowledgeMiner MUP map had a greater accuracy than the other methods,

being even more detailed than the original map, accounting for 80% of global index and a Kappa index of 0.6524. The information extracted by KnowledgeMiner provided rules for mapping the watershed surroundings with 70.97% of global index and a kappa index of 0.5586. Legacy soil data extracted by KnowledgeMiner from a detailed soil map and used to model soil class distribution outperformed decision trees, promoted improvements on the existing soil map, and allows for the creation of a soil map for the surroundings of the study area.

Keywords: Digital soil mapping; Knowledge acquisition; Fuzzy logic; Data mining; Tropical soils; Cost-constrained conditioned Latin hypercube sampling scheme.

3.1 INTRODUCTION

The global search for more detailed soil maps has gained increasing importance in the last two decades (Mendonça-Santos and Santos, 2007; McBratney et al., 2006; Hartemink and McBratney, 2008). Diverse projects are being conducted and focusing on the development of more accurate soil maps than existing ones, such as the AfSoilGrids250m (Hengl et al., 2015) that is creating soil property maps for Africa at 250 m resolution, GlobalSoilMap (Arrouays et al., 2014), which aims to make a new digital soil map of the world at a fine resolution, and SoilGrids1Km (Hengl et al., 2014), the first output for a series of finer resolution maps of soil properties and classes to be produced in the future. This fact is associated with diverse technological advances in recent

years, such as powerful electronic devices, the ease of accessing digital information, and satellite data availability, from which pedologists can utilize to their advantage.

Some of the most useful tools available are digital elevation models (DEM) found at different resolutions that provide great information and from which terrain attributes, such as slope, curvature and topographic wetness index, can be derived. Many works have applied these parameters to predict soil properties and classes (Moore et al., 1993; McBratney et al., 2000; McBratney et al., 2003; Behrens et al., 2010; Jafari et al., 2014; Vaysse and Lagacherie, 2015). These works consisted of studying the relief as major driver for soil differentiation, considering the other soil forming factors (climate, organisms, parent material and time) (Jenny, 1941) as relatively constant in the study area.

A more recent advance from the Jenny's model for soil formation (clorpt) is the SCORPAN (*soil = f* (*soil, climate, organisms, relief, parent material, age, n*), in which soils can be predicted from the classic five factors proposed by Jenny (1941) plus available information about the soils (s), such as existing maps, and soils spatial position (n). This model proposed by McBratney et al. (2003) allows for a more quantitative description of the relationships between soil and other referenced factors and it stresses that existing soil information (legacy data) could also be used to refine soil maps. In accordance with this fact, Silva et al. (2014) suggested that maps are made with the best tools and data available at the time they are created, but it does not impede that they can be updated as soon as more information is acquired in the future.

From this point of view, existing soil maps in Brazil, of which most were created prior to the advent of digital soil mapping, could be refined with

current available tools. Soil maps represent the pedologist's mental model about soils variability across the landscape (Bui, 2004). Many digital mapping tools can retrieve this knowledge from existing maps and correlate it with environmental factors, such as geology, topography, and vegetation (Taghizadeh-Mehrjardi et al., 2015).

Among data mining tools, decision trees are one of the most commonly used for digital soil mapping (Lagacherie and Holmes, 1997; Giasson et al., 2011; Häring et al., 2012; Kempen et al., 2015). They can identify the conditions that characterize each soil class according to different environmental variables with reasonable accuracy (e.g. 65-88% of overall accuracy, and a Kappa index of 0.44-0.51, according to Scull et al. (2005)), in which the data set is divided into more homogeneous subsets (Moran and Bui, 2002).

The procedure of decision trees starts at a root node, where the algorithm identifies the optimal split based on an exhaustive search of all possibilities, in order to maximize the average purity of the two nodes, employing the splitting or impurity function called Gini index (Loh, 2011). Nodes are locales where trees split the data set; terminal nodes are called leaves. A leaf node (predicted soil class, for example) is created when the decision tree reaches a stopping criterion (condition defined by the algorithm implemented in the trees, e.g. when the maximum tree depth is reached, when the splitting criteria is smaller than a threshold, among others (Rokach and Maimon (2008)), which makes the tree stops splitting nodes. Otherwise, the aforementioned step is, in turn, applied to each child node.

Decision trees are simple to understand and can identify the most representative variables for prediction (Bou Kheir et al., 2010). This results in a

consistent supervised way to aid in the comprehension of the pedologist's mental model encrypted in soil maps. Also, no assumptions are made regarding the underlying distribution of values of the predictor variables (non-parametric) (Friedman et al., 2000) and decision trees are able to search all possible covariates as splitters in the decision nodes. However, this exhaustive search approach has disadvantages, such as the one reported by Loh (2011), which is the greater chance of selecting the covariates that have more distinct values, if everything else is equal, affecting the integrity of inferences drawn from the tree structure. Henderson et al. (2005) used decision trees in Australia to predict soil pH and other properties based on terrain and climatic variables, at a 250 m resolution, and Lacarce et al. (2012) combined regression trees with geostatistics to predict Pb stocks in soils in France.

Another tool more recently created is the KnowledgeMiner that is part of the Soil-Land Inference Model (SoLIM) software (Zhu et al., 2001). It employs Kernel density to extract environmental variables information from each polygon on the map and then provides statistical indexes, such as minimum, maximum, mean, mode, median and standard deviation, in order to characterize those polygons (map units) and help the user to define the optimal environmental conditions for each map unit to occur.

It considers each cell value of a terrain attribute raster and a numerical interval that contains that cell value. Then, it counts the number of cells within each polygon that has values contained in that interval (SoLIM, 2007). This number of cells will be used to generate the frequency distribution curves (Kernel density), which allow the user to identify the most appropriate value of terrain attributes (e.g. slope gradient values) to individualize each soil class.

These data mining procedures may contribute to disaggregate polygons of the original map to create more detailed soil maps (Bui and Moran, 2001; Thompson et al., 2010; Nauman and Thompson, 2014; Subburayalu et al., 2014), however, a soil map whose polygons present more than one soil class (inclusions) may hinder these inferences.

Combining the need for more detailed soil maps in Brazil, where most of them are at a 1:750,000 scale due to increased funding limitations (Giasson et al., 2006), with the feasibility of using digital soil mapping tools, it has brought to light an economical alternative to obtain soil data: the usage of data mining tools to rescue information embedded on existing soil maps to improve those maps in a digital environment at a lower cost. Thus, this work had as objectives: (i) to compare two data mining tools, KnowledgeMiner and decision trees, to retrieve legacy soil data from a detailed soil map of a watershed in Minas Gerais, Southeastern Brazil; (ii) to create and validate these soil maps in the field, identifying the best method for refinement of soil maps; (iii) to extrapolate that extracted legacy data to the surrounding similar areas of this watershed, which present similar environmental conditions; and (iv) to validate this map.

3.2 MATERIALS AND METHODS

3.2.1 Study area and source of data

The study was developed at Marcela Creek Watershed, located in Nazareno county, state of Minas Gerais, Southeastern Brazil (Figure 1), between

the latitudes 21°14'27" and 21°15'51" S and longitudes 44°30'58" and 44°29'29" W. The climate of the study area is Cwa (warm temperate), according to Köppen classification, having dry winters and warm and rainy summers, presenting a mean annual precipitation of 1,300 mm, a mean annual temperature of 19.7°C and area of 485 ha.

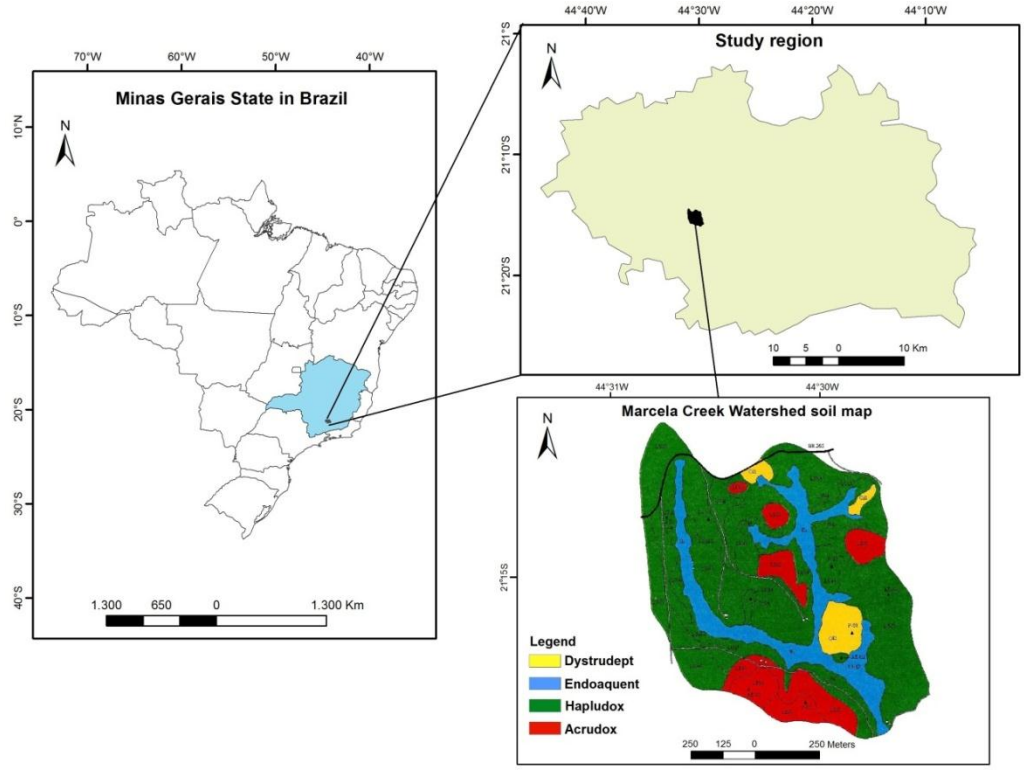

Figure 1 - Location of Marcela Creek Watershed in Minas Gerais state, Southeastern Brazil.

This watershed was chosen due to its great agricultural potential (Silva et al., 2013), high water yield capacity and potential for electric energy generation (Beskow et al., 2013), and for being representative of the Mantiqueira Fields physiographical region. Its water drains into the Itutinga/Camargos hydroelectric power plant reservoir, which is a very important source of electric energy for Southeastern Brazil. Whereas water management is a governmental concern, the knowledge of soils and their distribution are important since soils exert an influence on water movement in different ways (Mello and Curi, 2012). Additionally, there is an important environmental issue in this region: the native vegetation (cerrado and forest) has been rapidly replaced by extensive pasture or crops (more recently) promoting intense land degradation (Alvarenga et al., 2012). This fact could impair the maintenance of hydrological functions of both the study area (Beskow et al., 2013) and its surroundings.

This watershed was mapped by very experienced pedologists and published by Motta et al. (2001), at a scale of 1:12,500, through intensive field work, including description of soil profiles, collection and laboratory analyses of soil samples, making up the basic source of information for the development of this current work. The soil classes found were Hapludox (Hx), Acrudox (Ax), Dystrudept (Dt) and Endoaquent (Et), according to Soil Taxonomy (Soil Survey Staff, 1999).

A 30 m Aster (version 2) DEM, which is the best DEM resolution freely available for Brazil, was obtained from the website http://gdem.ersdac.jspacesystems.or.jp/, preprocessed in order to make it hydrologically consistent, and used to create 12 terrain attributes (TA) on SAGA GIS software (Bohner et al., 2006): slope gradient, topographic wetness index

(WI) (Beven and Kirkby, 1979), longitudinal curvature, cross-sectional curvature, multiresolution index of valley bottom flatness (mrvbf) and multiresolution index of ridge top flatness (mrrtf) (Gallant and Dowling, 2003), vertical distance to channel network (VDCN), hillshade, slope aspect, valley depth, and SAGA wetness index (SWI), which differs from WI for being calculated based on a modified catchment area, resulting in a more realistic representation of water accumulation potential in some portions of the landscape than WI (Olaya and Conrad, 2009). These TA were selected based on works on digital soil mapping that evaluated them in modeling soil classes and properties (Moore et al., 1993; Kim and Zheng, 2011; Brown et al., 2012; Adhikari et al., 2013; Malone et al., 2014; Vaysse and Lagacherie, 2015) and, thus, it was possible to learn the typical environmental conditions of each soil mapping unit from the soil map (legacy data).

The soil classes distribution on the landscape (legacy data from the soil map) was related to the TA and extracted from the map in two different ways, considering: a) each mapping unit polygon entirely (MUP) (6221 training pixels) and b) a circular buffer created within a distance of 100 m of radius from the sampled points (CBP) (406 training pixels) of the soil map produced by Motta et al. (2001) (profile descriptions legacy data). Soil prospections at 97 places were performed throughout the area and the collected samples were submitted to analyses of particle size distribution, soil fertility and sulfuric acid digestion, allowing for soil classification at those places and creation of the soil map (Motta et al., 2001). Then, at the most representative places of each soil class, chosen after an intensive field campaign throughout the area, 9 complete soil profile descriptions were performed, including collection of samples in each

soil horizon that were subjected to complete physical and chemical laboratory analyses. A GPS device was used for acquisition of the geographic coordinates at those locations. The buffers of the CBP method were created around the soil profiles because they were chosen in locales representative of each soil class, representing their typical forming conditions, while the other observations were mostly used to help in the delineations of the polygons on the soil map. That distance from the collected points was chosen in a way that most points could be sampled without crossing the border of the mapping unit to which it belonged.

The assumption was that either the polygon or the buffer of the points should adequately represent the main characteristics of the terrain attributes for the soil class it represents. Polygons should represent the dominant soil class in the map units and, thus, the environmental conditions commonly associated with each soil class. Regarding the buffer of the points, the places where profile descriptions were made represent typical locales of each soil class. In this case, areas close to them should present the most appropriate conditions for each soil class to occur.

3.2.2 KnowledgeMiner

Box plots were generated to identify the terrain attributes that best distinguish each soil class in order to use only the best terrain attributes in KnowledgeMiner. They were created in R software (R Development Core Team, 2009). Box plots represent the range of values, besides the median, minimum,

maximum, 1st and 3rd quartiles, and the inter-quartile range of the data set, which aid in the comparison of data (Brungard and Boettinger, 2010). In this work, box plots allow for the analyses of the variability of terrain attribute values for the soil classes. Thus, the more individualized the values of a terrain attribute are for a soil class in comparison to the other soil classes, shown on box plots, the better that terrain attribute is to distinguish that soil class from the others (Brown et al., 2012; Gessler et al., 1995).

Afterwards, KnowledgeMiner, a tool set of the SoLIM project (Zhu et al., 2001), was employed to extract the pixel values of each terrain attribute for both MUP and CBP from the map. It consists of identifying the optimal (typical) value of each TA for each soil class through calculation of the frequency distribution curves from TA values occurring in each MUP or CBP of the map. Optimal values represent the typical environmental condition of each TA for each soil class to occur, and they are given by the mode of the TA values for each soil class. Then, the optimal values and those curves can show whether any of those environmental layers can explain the pedologist's mental model employed to distinguish a soil class from another during the original mapping process (Menezes et al., 2013). KnowledgeMiner also provides statistical indexes for each polygon on the map, such as minimum, maximum, mean, and standard deviation besides the typical value (mode).

KnowledgeMiner employs the kernel density estimation method for calculating the frequency distribution curves, which show the frequency of different cell values of terrain attributes within each polygon (soil map unit) on the map (SoLIM, 2007). This method considers the value of each observation (e.g. a cell of a terrain attribute), a numerical range that contains that value and

the frequency of that value will be given by the number of cells having values that fall into that numerical interval. It can be expressed by the formula below (SoLIM, 2007):

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n}.\ K\left(\frac{x-x_i}{h}\right) \quad (1)$$

where: f(x) is the density, h is the degree of smearing or bandwidth, which is a number that defines a numerical range starting from a minimum value "A" in which the number of cells (pixels) that falls into this range will determine the frequency value for "A", K represents the kernel function, (x - xi) is the distance between two points, and n is the number of independent observations. Through this analysis, it is possible to identify the optimal values of terrain attributes for characterizing each soil class (modal value).

After the selection of the most important attributes, based on analyses through the aforementioned box plots, the values extracted by KnowledgeMiner were inserted into ArcSIE (Shi et al., 2009), an ArcGIS extension that uses fuzzy logic to identify the places within the area of interest that corresponds to each soil class environmental condition. For each condition, ArcSIE generates a membership map showing the degree of similarity of each pixel to that condition through a similarity vector Sij (S^1_{ij}, S^2_{ij},...S^k_{ij},...S^n_{ij}), where S^k_{ij} is the similarity value between the soil at (*i*,*j*) location and the soil class k, and n is the number of soil classes (Zhu et al., 2001). The degrees of similarity range from 0 (very different) to 1 (very similar). For that calculation, it is necessary to insert into ArcSIE the typical value of a TA for that soil class (v1), and the deviations of this value (w1 and w2), which are calculated by multiplying the standard

deviation of the TA values by 0.2 and represent the number to be subtracted (w1) or added (w2) to the typical value (v1) which will represent 50% of the degree of similarity with the typical conditions to that soil class to occur, while v1 represents the 100% of degree of similarity (typical condition to that soil class to occur). The greater is the similarity, the greater the chance of that place to contain the same soil class of the driving condition.

After classifying each pixel according to their similarities to every soil class found in the area, the membership maps (one per soil class), which shows the degree of similarity of each pixel with every soil class, are generated and, in sequence, all of the membership maps are joined to form a final soil map. In this procedure, each pixel of the final map will represent the soil class that has the greatest similarity to that place in terms of TA values.

3.2.3 Decision Trees for knowledge discovery

Decision trees, another data mining technique commonly used to map soil classes and properties worldwide (Scull et al., 2005; Bou Kheir et al., 2010; Tehrany et al., 2013), were compared with KnowledgeMiner in terms of individualization of soil classes, since several works have employed this mapping technique to predict soil classes also in Brazil (Crivelenti et al., 2009; Giasson et al., 2011; Caten et al., 2012; Giasson et al., 2013). However, most of these studies were performed at a lesser detailed scale, in Southern Brazil, and few comparisons have been made with other methods, which is performed in

this current work with KnowledgeMiner, in a different region of the country, and based on a more detailed scale soil map.

In this task, information of the 12 TA for both the MUP and CBP was collected. With this set of data, decision trees were created using the software R (R Development Core Team, 2009), and the package rpart (Therneau et al., 2015) with the classification and regression trees (CART) algorithm (Breiman et al., 1984) to develop the instances for the occurrence of each soil class. Afterwards, from those instances the soils maps were created by algebra of maps in the software ArcGIS 10.1 (ESRI).

3.2.4 Validation of the soil maps

In order to validate the maps generated from KnowledgeMiner extraction, followed by spatialization using ArcSIE, and from decision trees for both the MUP and CBP methods, 20 field prospections were performed. The validation places were selected using the cost-constrained conditioned Latin hypercube sampling scheme (CCLH) (Roudier et al., 2012). The conditioned Latin hypercube (CLH) is a stratified random procedure, proposed by Minasny and McBratney (2006), consisting of an efficient method for defining sampling locations, taking into account the variability of environmental covariates that are related to the attribute to be mapped, such as terrain attributes related to soil classes, assuming that the location to be sampled must exist on the landscape (Brungard and Boettinger, 2010). However, the places to be sampled are

generally sparsely distributed in the area, which makes the field work more costly or impractical (Silva et al., 2015).

In order to overcome this issue, Roudier et al. (2012) proposed the CCLH, which differs from the CLH for evaluating the difficulty that someone should face to reach the sampling locations in the field. Thus, this sampling system preferably chooses places of easy access, still taking into account the variability of environmental covariates (e.g. terrain attributes) in this procedure. The terrain attributes that were indicated as being important for distinguishing soil classes from the box plot analyses were used as basis for the CCLH to choose the location of validation samples, since the study area present regions that are difficult to access, mainly due to lack of roads.

CCLH requires a map of the cost of reaching every place within the area in order to determine the sample locations in places of easy access. This map was created combining the slope gradient with the distance from the road, which were the most limiting factors to access the whole area. Distance from the road was created using the Euclidean distance tool in ArcGIS software (ESRI) using a set of roads created from analysis of a RapidEye satellite image as the input data. Then, to both distance from the road and slope gradient rasters, a weight (cost) was assigned to each pixel according to the distance (the farther from the road the pixel is, the greater the cost) and the slope gradient (the steeper the relief, the more time in reaching that place). Those rasters were added to each other using the Map Algebra tool in ArcGIS (ESRI) and the resulting map (cost map) showed the cost (difficulty) of reaching every pixel within the area. Thus, taking this cost map into account, CCLH determined the best places to be

sampled, preferring those with easy access (close to roads and presenting gentle relief).

Aiming to assess the quality of the predictive maps, four indexes were calculated: global index, Kappa index, user's accuracy and producer's accuracy. Global index, also known as overall accuracy, represents the number of samples whose soil classes identified in the field matches the soil classes presented on the soil map divided by the total number of samples. Adhikari et al. (2014) employed global and other indexes to assess the accuracy of a soil class map of Denmark.

Kappa index is calculated taking into account the number of soil classes, the number of correctly classified samples and the total number of samples (Congalton and Green, 1999), as follows:

$$Kappa = \frac{Po - Pe}{1 - Pe} \quad (2)$$

where *Po* is the proportion of correctly classified samples, and *Pe* is the probability of random agreement. The results range from -1 to 1, although they commonly are found between 0 and 1, and they indicate increasing accuracy as the values get closer to 1 (Landis and Koch, 1977). Lacoste et al. (2011) used the Kappa index to assess the accuracy of soil parent material prediction maps and as well as Brungard et al. (2015) who mapped soil classes in western USA and used that index and others to evaluate machine learning techniques performance.

User's accuracy shows the probability of a place classified as a soil class on the map to match the soil class found in the field, whereas, a producer's accuracy expresses the probability of a point of a soil class being correctly

classified on the map (Congalton, 1991), considering that an adequate map has values for those two accuracies close to one (100%) (Behrens et al., 2010). Those indexes are presented by the formulas below:

$$User's\ accuracy = \frac{Xii}{\sum_{i=1}^{r} Xij} \quad (3)$$

$$Producer's\ accuracy = \frac{Xjj}{\sum_{j=1}^{r} Xij} \quad (4)$$

where X_{ii} and X_{jj} indicate the number of correctly classified samples and X_{ij} represents the sum of samples of a soil class in a row (user's accuracy) or column (producer's accuracy) of a confusion matrix. Those indexes were also used by Collard et al. (2014) to assess the accuracy of a soil class map in France created from a reconnaissance soil map using regression models, and by Bou Kheir et al. (2008) to verify the quality of distribution maps of soil and bedrock susceptible to gully erosion, based on a decision tree model.

3.2.5 Extrapolation of the soil map information and its validation

After defining the best method for extracting soils information from the map through the field validation, the information obtained from the legacy data was extrapolated to an area of 1771.9 km², which surrounds the studied watershed (4.85 km²) (reference area) and that contains similar environmental

conditions (Curi et al., 1994). The map of this area was also validated using 31 profile descriptions obtained from Giarola (1994), Araújo (2006), UFV-CETEC-UFLA-FEAM (2010), and the Brazilian Agricultural Research Corporation (EMBRAPA) data set (available at http://www.sisolos.cnptia.embrapa.br), which is the number of points available for this area. Global index and Kappa index were calculated in order to assess the accuracy of this extrapolation and, thus, the quality of the created soil map of the region.

3.3 RESULTS AND DISCUSSION

3.3.1 KnowledgeMiner for mapping soils

Figure 2 shows box plots generated from diverse TA for both MUP and CBP to help visualization to identify the best environmental covariates for individualizing each soil class. In box plots, the difference between soil classes and their respective TA values, even in studies of different nature, is assessed by visual analysis (Gessler et al., 1995; Brown et al., 2012; Caten et al., 2012; Teske et al., 2014), since box plots were designed for that. The more separated the values of a TA are for a soil class in relation to the other soil classes, the better that TA is for soil class individualization. Häring et al. (2012) also employed box plots to identify the differences between values of TA for two soil classes and then applied them to models. Contrary to decision trees, which select and use the most important covariates for predictions, KnowledgeMiner only determines the optimal value of each TA for each soil class, and the user is

responsible for selecting the most appropriate terrain attributes to distinguish each soil class from another, according to their environmental conditions.

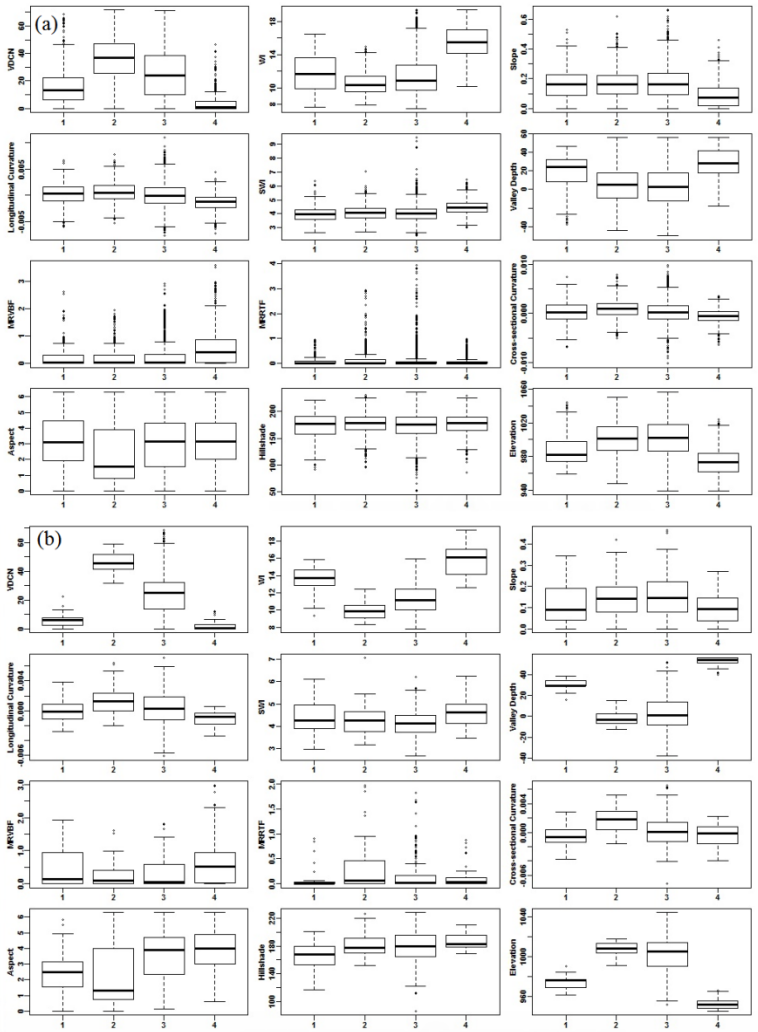

Figure 2 - Box plots of the terrain attributes based on each soil class extracted from the entire polygon (a) and from the buffer of the points (b). 1 = Hapludox, 2 = Acrudox, 3 = Dystrudept, 4 = Endoaquent. VDCN = vertical distance to channel network, WI- wetness index, SWI - SAGA wetness index, MRVBF - multiresolution index of valley bottom flatness, MRRTF - multiresolution index of ridge top flatness.

In Figure 2 it can be seen that for both MUP and CBP from the 12 TA, five of them (VDCN, valley depth, WI, slope, and elevation) presented greater potential to distinguish at least one soil class. VDCN and valley depth calculate the vertical distance from each pixel to the channel network through interpolation of the base level and subsequently subtraction of these values from the original elevation values (Conrad et al., 2015); WI takes into account the slope and specific catchment area, being defined by $ln(a/tan\ b)$, where a = ratio of upslope contributing area per unit contour length and b = the local slope (Beven and Kirkby, 1979), and expresses the places with higher tendency of accumulating water; slope is the ratio of changes in elevation; and elevation is the representation of altitudes in the area. In this sense, since VDCN and valley depth have similar calculation procedures (Conrad et al., 2015), but VDCN still takes into account channel network information and has been more commonly reported as a good soil property predictor (Adhikari et al., 2013; Bishop et al., 2015; Taghizadeh-Mehrjardi et al., 2015), contrary to valley depth, whereas elevation only represents variations of altitude without showing the proximity of each cell to channels, which is important for soil class prediction, only VDCN among these three TA was employed for the analyses. Thus, VDCN, slope, and WI were the TA considered most important for distinguishing the soil classes in the study area.

Table 1 presents the optimal values of TA for each soil class extracted for MUP and CBP methods defined by KnowledgeMiner. The more individualized is the optimal value of a TA for a soil class, the better that TA is in distinguishing that soil class from others. This information is expressed as frequency distribution curves by KnowledgeMiner, whose peaks identify the

optimal value and these peaks assist the user to visualize the degree of overlapping of the TA values for each soil class. In Table 1 it can be observed that VDCN is the one TA that has more different optimal values for distinguishing soil classes, while slope and WI are more adequate to individualize the Et, since for the other soil classes the optimal values of these two TA are similar.

For both MUP and CBP, it is noted that, as expected, Et occurs in the lowest portions of the landscape and closer to the channel network (MUP = 3.7 m and CBP = 1.1 m), which is in agreement with the greatest wetness index compared to the other soils (15.4 and 16.9) and gentler slope (7% and 7.7%). The fact that Oxisols and Inceptisols tend to occur in an intricate pattern in the region of this study may have hindered their individualization, in accordance with the similar values of slope and wetness index found for those soils. Curi et al. (1994), characterizing the soils of the Mantiqueira Fields, region of this current study, identified that it is not uncommon to find Inceptisols associated with Oxisols in this area, where Oxisols tend to be shallower than typical Oxisols and sometimes having properties intermediate between Oxisols and Inceptisols. However, Inceptisols ideally tend to occur in areas with shorter slope lengths and more linear hills, which are common on the lower part of the backslope, right above the channel network. This could also be verified by VDCN values of this soil class, having them intermediate VDCN values between Et and the Oxisols (Table 1).

Table 1 - Optimal values of the terrain attributes extracted from KnowledgeMiner for each soil class.

Soil Class	MUP[1]			CBP[2]		
	VDCN[3]	Slope (%)	WI[4]	VDCN	Slope (%)	WI
Dystrudept	16.6	13.5	11.8	11.8	13.9	12.2
Endoaquent	3.7	7.0	15.4	1.1	7.7	16.9
Hapludox	25.5	14.0	11.4	22.8	16.8	11.6
Acrudox	35.7	13.8	10.5	44.2	15.4	10.0

[1]Entire mapping unit polygon extraction methods and [2]Buffer around the sampled points extraction method, [3]Vertical distance to channel network, [4]Wetness index.

Both Ax and Hx presented similar values of slope and wetness index. In spite of that, Ax commonly occupies the highest areas of the landscape, implying a relatively better drainage than Hx, which is reflected by their color, since hematite requires a drier condition to form than goethite (Curi and Franzmeier, 1984). Thus, some TA could also capture this tendency: VDCN for Ax is greater than for Hx and wetness index for Ax is lesser, which indicates that the latter places have less tendency of accumulating water than those of Hx.

Table 2 - Values extracted from KnowledgeMiner and inserted into ArcSIE to predict the occurrence of soil classes.

Soil Class	TA[1]	MUP[2]			CBP[3]			Curve shape
		v1	w1	w2	v1	w1	w2	
Dystrudept	VDCN[4]	17	10	8	11.8	6.8	6.8	Bell
	WI[5]	--	--	--	12.2	1.9	1.9	Bell
Endoaquent	VDCN	3	--	2	1.1	--	1.5	Z
	Slope	7	--	4	--	--	--	Z
	WI	--	--	--	16.9	1.2	--	S
Hapludox	VDCN	25	10	10	22.8	5.1	5.1	Bell
	WI	--	--	--	11.6	1.4	1.4	Bell
Acrudox	VDCN	35	13	--	44.2	5.8	--	S
	WI	10	2	2	10.0	0.3	0.3	Bell

[1]Terrain Attributes, [2]Entire mapping unit polygon extraction method and [3]Buffer around the sampled points extraction method, [4]Vertical distance to channel network, [5]Wetness index.

Figure 3 illustrates the predicted soil maps of Marcela Creek Watershed for both the MUP and CBP methods by extracting data from the original soil map and Table 3 presents the areas occupied by each soil class. On both

predicted maps, the areas of Dt, Et, and Ax increased in relation to the original map (Figure 1), whereas, the area of Hx was reduced.

Although some inconsistencies can be seen on the predicted map for MUP, such as the Dt crossing areas of Et in three places and the discontinuities of the Et in other places different from the ones already cut by Dt, the general distribution of the soil classes were more similar to the original map than that for the CBP predicted map. However, on the latter, the continuity of the Et was maintained, but the Dt area was considerably increased, while the Hx area was reduced (see validation of the maps in section below). Similar to these changes in soil class areas, Calderano Filho et al. (2014) found an overestimation of some soil class areas in the predicted map in comparison with the original soil map at Serra do Mar, Brazil, using geology, TA derived from a DEM and other remote sensing data as variables for predictions.

Table 3 - Areas occupied by each soil class in the study area on the different maps.

Soil class	Area (ha)		
	Original map	Predicted MUP	Predicted CBP
Dystrudept	21.29	60.6	145.7
Endoaquent	75.34	94.18	83.9
Hapludox	325.38	221.83	132.1
Acrudox	63.89	108.50	126.7

[1]Entire mapping unit polygon and [2]Buffer around the sampled points extraction methods.

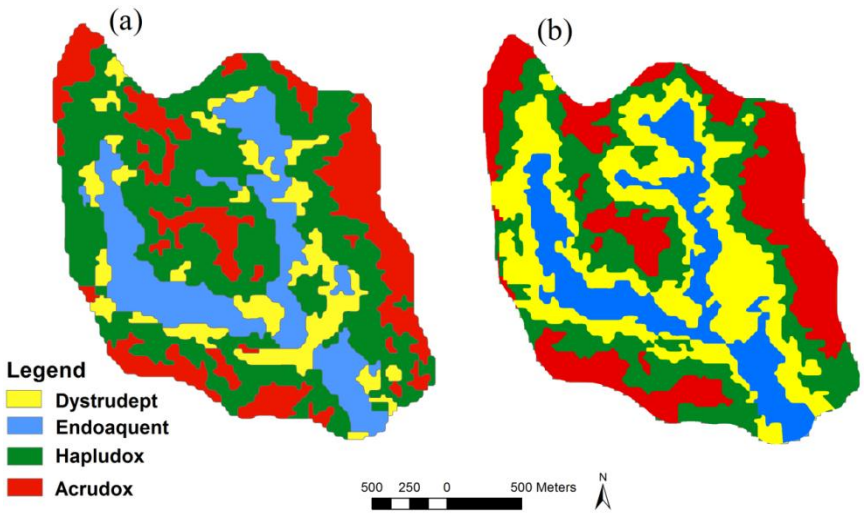

Figure 3 - Predicted soil maps by extracting information based on the original map from the entire polygons -MUP- (a) and from the buffers of the sampled points -CBP- (b) using KnowledgeMiner for Marcela Creek Watershed.

3.3.2 Decision Trees for mapping soils

In order to use another method for data mining to express the thresholds for each soil class, decision trees were created for both the MUP and CBP methods (Figure 4). Although 12 TA were used as input data for the creation of decision trees, only 5 TA were chosen as adequate splitters by the decision tree algorithm for MUP to distinguish the soil classes, these included WI, VDCN, longitudinal curvature, aspect and valley depth, while only 2 (VDCN and valley

depth) were used as splitters by the CBP decision tree. The importance of each TA was calculated using the rpart package (Therneau et al., 2015) in R and is presented in Table 4, in which the greater the importance, the better is that variable to distinguish the soil classes. Usage of lesser variables, chosen by models for predictions, rather than the total number of input variables is not uncommon (Henderson et al., 2005; Odgers et al., 2014). Bou Kheir et al. (2010) and Jafari et al. (2014), employing many environmental covariates for modeling and predicting soil properties, also found that the covariates have different importance in modeling and they stated that this is the reason for some covariates being more commonly used for predictions than others.

Table 4 - Importance of the terrain attributes used on each decision tree.

Terrain attribute	MUP (%)	CBP (%)
Wetness Index	26	11
Vertical distance to channel network	22	22
Digital elevation model	17	24
Valley depth	16	26
Slope	7	2
Aspect	5	2
Multiresolution index of valley bottom flatness	3	3
Longitudinal curvature	2	1
Multiresolution index of top ridge flatness	1	2
Cross-sectional curvature	1	7
Hillshade	0	2
Total	100	100

Comparing the CBP with MUP, the percentage of correctly classified points was 93.1% for CBP and 73.2% for MUP and root node errors were, respectively, 39.65% and 33.21%. Those values of correctly classified points are greater than the ones found by Caten et al. (2012) when evaluating different methods and amount of data to create decision trees for mapping soils in Southern Brazil.

It was noticed that the Dt was not predicted by the MUP decision tree, probably because in this region Dt occurs in very similar landscapes with those where Oxisols are common. In the field, even the slight differences in their landforms are not very helpful for pedologists to distinguish those soil classes on the landscape. Oxisols and Dt, respectively, are related to convex and linear landforms, however, this trend is not very well defined in this region due to these soils occur in intimate geographical association on the landscape (Curi et al., 1994), where Inceptisols are developed from erosion of ancient Oxisols (Resende et al., 2014). The two TA representing the longitudinal and the cross-sectional curvatures that were among the 12 TA available for modeling and that were expected to help the decision tree modeling could not capture a distinction of the landforms of both soils (Figures 2), whereas, those soils were distinguished by KnowledgeMiner. Caten et al. (2012) and Giasson et al. (2013) found that only the more complex decision trees were capable of predicting all the soil classes in their study areas. Contrary to these studies, the CBP decision tree, although simpler than the MUP, could individualize all the soil classes for capturing more specific environmental conditions of each soil class, even when considering that the Ax area predicted was very much reduced in comparison with the original map.

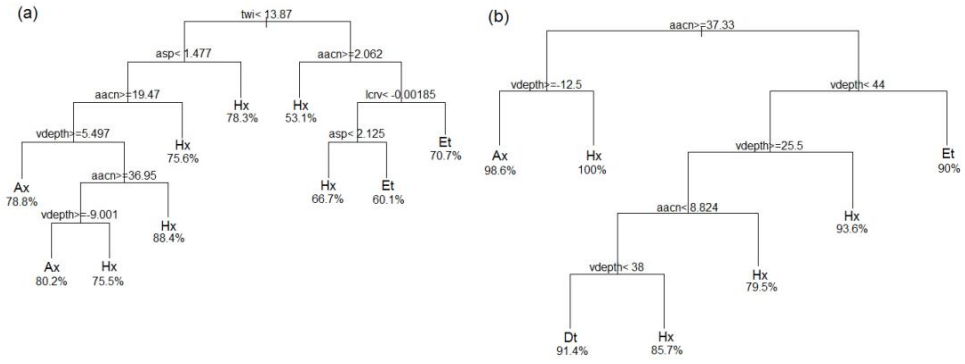

Figure 4 - Decision trees developed for predicting the soil classes for MUP (a) and CBP (b) methods in Marcela Creek Watershed. Percentage under the soil class symbols indicates the amount of correctly classified points in that instance.

From the analysis of the decision trees, the predicted soil maps were created (Figure 5). As expected through the analysis of the MUP decision tree, the Dt was not represented in the predicted map. On both prediction maps, the areas of Ax and Et were reduced, while the Hx area increased (Table 5). Dt was not identified on the MUP map, as expected, but on the CBP map its area increased in comparison to the original map. The discontinuities observed for the Et (Figure 3a) in the predicted map made using KnowledgeMiner were also seen in this map, but with greater frequency and greater distances between polygons. This soil class on CBP map was constrained to the lowest areas of the watershed. Most of the area on both maps was occupied by Hx. Also, in general, both maps turned out very different from the original map. Visually comparing

them with the original map, KnowledgeMiner maps presented greater similarities to it than the decision tree maps.

Table 5 - Areas of each soil class on the original and predicted maps created by decision trees.

Soil class	Area (ha)		
	Original map	Predicted MUP[1]	Predicted CBP[2]
Dystrudept	21.29	--	43.9
Endoaquent	75.34	56.3	33.6
Hapludox	325.38	403.0	407.4
Acrudox	63.89	19.1	0.78

[1]Entire mapping unit polygon and [2]Buffer around the sampled points extraction methods.

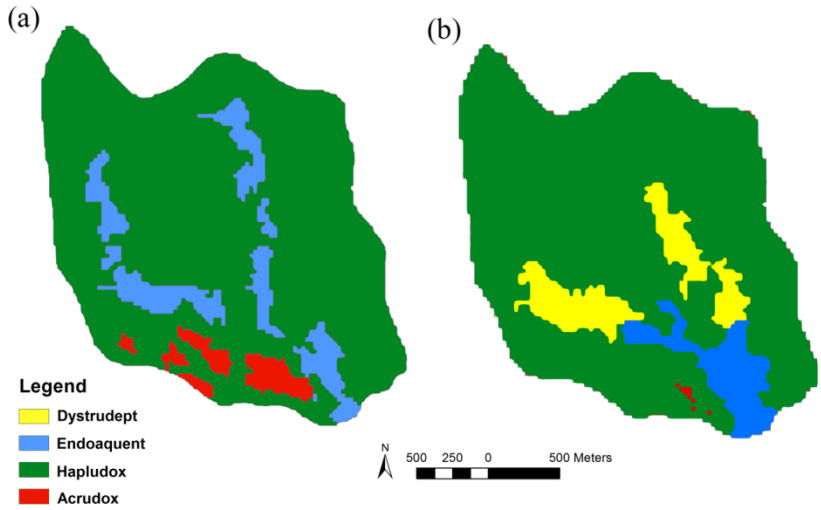

Figure 5 - Maps created from extraction of information of the original map by the entire polygon (a) and by buffers of the points (b) with thresholds identified by decision trees for Marcela Creek Watershed.

3.3.3 Validation of the original and prediction maps

Validation of the predicted maps was performed at 20 places defined by the cost-constrained conditioned Latin hypercube (Roudier et al., 2012). The places chosen by this sampling system captured considerable variability of the soils of the study area, which means that different environmental conditions and all the soil classes reported in previous works were observed, as found by Silva et al. (2015) and Mulder et al. (2012), who employed CCLH to map soil

properties. Furthermore, none of the soil classes found during the field validation differed from those represented in the original map.

In order to verify the accuracy of the original map, it was validated with 20 prospections and presented a global index of 75% and Kappa index of 0.4681, corresponding to a moderate classification, according to Landis and Koch (1977). According to the Brazilian Pedology Technical Manual (IBGE, 2007), soil maps must have at least 70% accuracy and 30% of inclusions to be acceptable. These results confirm that this map represents well the variability of the soil classes in the watershed, being an adequate source of soils information for works that intend to extract and then apply this information in areas with similar soil and environmental patterns.

In the MUP map created from KnowledgeMiner procedure, 16 out of the 20 prospections (80%) were correctly predicted by the soil map and resulted in a Kappa index of 0.6537, equivalent to a substantial classification according to Landis and Koch (1977), while the CBP map had 50% of overall accuracy and a Kappa index of 0.2063, a fair classification. Overall, validating the maps generated based on decision trees, the MUP correctly predicted 11 out of 20 prospections (55%) and had a Kappa index of 0.0674, whereas, the CBP map had a 50% global index and 0.0148 for the Kappa index, with both being classified as having a slight agreement between these maps and the original map.

Producer and user accuracies as well as omission and commission errors were calculated as other parameters for comparison (Table 6). It is observed that the map generated from MUP with KnowledgeMiner had the greatest producer's and user's accuracy and the least errors for all the soil classes, in agreement with the other validation parameters.

Table 6 - Producer's and user's accuracy and omission and commission errors for the methods evaluated for mapping soils.

Soil Map	Soil Class	Producer Accuracy	Omission Error	User Accuracy	Commission Error
KnowledgeMiner MUP	Dt	100.0	0.0	50.0	50.0
	Hx	83.3	16.7	83.3	16.7
	Ax	100.0	0.0	80.0	20.0
	Et	33.3	66.7	100.0	0.0
Mean		79.2	20.8	78.3	21.7
KnowledgeMiner CBP	Dt	16.7	83.3	50.0	50.0
	Hx	70.0	30.0	58.3	41.7
	Ax	66.7	33.3	40.0	60.0
	Et	0.0	100.0	0.0	100.0
Mean		38.3	61.7	37.1	62.9
Decision Trees MUP	Dt	0.0	0.0	0.0	100.0
	Hx	58.8	41.2	83.3	16.7
	Ax	0.0	0.0	0.0	100.0
	Et	33.3	66.7	100.0	0.0
Mean		23.0	27.0	45.8	54.2
Decision Trees CBP	Dt	0.0	100.0	0.0	100.0
	Hx	56.3	43.8	75.0	25.0
	Ax	0.0	0.0	0.0	100.0
	Et	33.3	66.7	100.0	0.0
Mean		22.4	52.6	43.8	56.3

Dt - Dystrudept, Hx - Hapludox, Ax - Acrudox, Et - Endoaquent.

These results indicate that KnowledgeMiner in association with ArcSIE could predict the soil classes of the study area with greater accuracy than both the original map and the decision tree maps, which confirms a gain in details in comparison to the existing map and also in relation to the maps generated by decision trees, although being more time consuming to analyze than decision trees. Other works whose authors used SoLIM or ArcSIE to map soil classes and properties also obtained highly accurate maps as a result, such as Ashtekar and Owens (2013), who successfully utilized SoLIM software to map soil classes and loess depth in Indiana (USA) and compared the results with corn yield. Menezes et al. (2014) utilized ArcSIE to generate a soil class map and then a solum depth map for a watershed in Brazil, as a rule-based procedure aided by expert knowledge, to compare a digital with conventional soil maps and found a greater accuracy using the digital mapping procedure. Akumu et al. (2015) used fuzzy-logic in ArcSIE to map soil textural classes in Canada, with slope, wetness index, elevation, curvature and other TA as input variables, obtaining a fine resolution textural map for an area of 4,300 km². The authors could not find published works that used KnowledgeMiner, although this tool set is part of the SoLIM software (Zhu et al., 2001).

Decision trees are dependent on the strength of the relationships between environmental variables and soil properties (Greve et al., 2010) and, taking into account the fact that Dystrudepts and Oxisols of the study area are located on very similar landscape positions, it may have contributed to their lesser quality of predictions found. Both the longitudinal and the cross-sectional curvatures included into the set of TA available for decision tree modeling in this work were not successful in capturing differences that enabled the distinction of these

soils, which are the TA that were expected to contribute to differentiate those soils, according to the knowledge of pedologists who performed soil surveys and mapping in this study region. This constraint was also encountered by Bui and Moran (2003), re-mapping soils from Murray-Darling basin, Australia, who stated that soil prediction becomes difficult in places where the spatial pattern of the soil units are not well captured by decision trees.

The MUP extraction method had a better performance to express the soil patterns than CBP not only by the KnowledgeMiner, but also by the decision trees, resulting in greater accuracy indexes, although the MUP by decision trees could not predict the Dt. CBP was tested due to the fact that in soil surveys, profile descriptions and sample collections are performed at places where the local soil can represent the map unit, thus, its surrounding areas should reflect the common conditions for that soil class to occur, contrary to the polygon boundaries. It is in agreement with the catena concept (Milne, 1935) that soils occurring on similar landscape positions should be similar, as long as the other soil forming factors (Jenny, 1941) are the same, but it did not work well using the CBP method either by KnowledgeMiner or with decision trees. Future investigations on this method should include different buffer distances from the sampling places.

3.3.4 Extrapolation of the soil information to surrounding areas with similar environmental conditions

The rules of soils occurrence obtained from the detailed soil map and extracted by KnowledgeMiner with MUP, for presenting the greatest accuracy, were used to extrapolate that information to an area of 1,771.9 km², which contained similar environmental conditions and parent material as the studied watershed (Curi et al., 1994). Decision trees were not used for the extrapolation procedure since the maps resulted from both MUP and CBP methods presented not only lower accuracy than the KnowledgeMiner MUP map, but also a lower global index than the 70% threshold recommended by the Brazilian Pedology Technical Manual (IBGE, 2007), and the best decision tree map (MUP) could not predict the occurrence of one of the soil classes (Dt) even within the reference watershed. The predicted soil map of the entire area is presented in Figure 6.

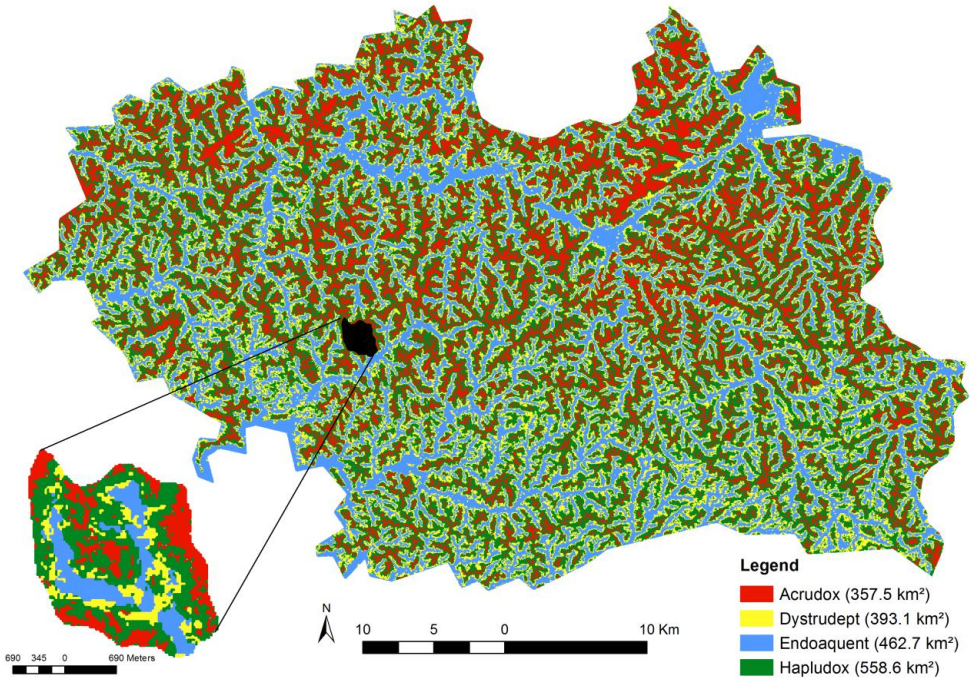

Figure 6 – Soil map of the region that is represented by Marcela Creek watershed generated by extrapolation of information extracted by KnowledgeMiner from the detailed soil map of the watershed.

The dominant soil classes are Hx (31.53%), Et (26.11%), Dt (22.18%), and Ax (20.18%). The accuracy of this soil map accounted for 70.97% of global index and a Kappa index of 0.5586, equivalent to a moderate classification according to Landis and Koch (1977). These results are a little worse than those obtained for the Marcela Creek watershed (80% of global index and 0.6537 for a Kappa index), however, considering the maximum of 30% of inclusions

acceptable in soil maps according to the Brazilian Pedology Technical Manual (IBGE, 2007) and the extension of the area for extrapolation (1771.9 km²) being 365 times larger than the reference area (4.85 km²), these results can be considered adequate. It indicates that this data mining tool associated with both ArcSIE and a good quality soil map as a source of data can create accurate soil maps for areas with similar conditions as the reference area. Reference areas (Favrot, 1989) contain soils representative of the surrounding areas and can be used to create soil surveys to help to understand the soils of the region of study where data are lacking. Lagacherie and Holmes (1997) found that a reference area, containing a large scale (detailed information) soil map, can be used as source of information to map other areas, however, this new map should not be expected to represent the scale as large as the one of the reference area map, but it can be used for future works and for guiding activities. These findings are important mainly for countries that have limited financial support for soil survey activities, such as Brazil (Mendonça-Santos et al., 2007; Silva et al., 2014).

3.3.5 Final considerations

It is known that polygon soil maps, in general, are not pure units, which means that the same polygon may contain more than one soil class (inclusions) (Soil Survey Staff, 1993). This could make it more difficult for the mapper to successfully use the tools tested in this work to obtain information from less detailed scale soil polygon maps in order to refine the existing map, mainly using decision trees, which are more adequate when there is more homogeneity

of soil classes within a map unit (polygon) (Bui and Moran, 2001). This fact may prevent the mapper from understanding in details not only the pedologist's mental model embedded within the map, but also the relation between a soil class and its typical environmental conditions for future refinement of that map. In these cases, disaggregation of polygons may be an alternative to identify different patterns within the same polygon (Kerry et al., 2012; Holmes et al., 2015). However, using KnowledgeMiner to determine the optimal values of environmental covariates for each map unit of Marcela Creek watershed soil map, it seems that inclusions are less considered and this tool could adequately capture the dominant conditions of the main soil class per polygon (MUP method), constituting an adequate alternative for data mining of soil legacy data.

Furthermore, as the methods employed on this work rely on relationships among soil classes and terrain attribute rasters derived from a DEM, the quality of these spatial data is fundamental and, if not appropriate, can negatively influence the final results (Sorensen and Seibert, 2007; Thompson et al., 2001). It should also be taken into account the spatial resolution of the DEM. Hengl (2006) exposes several methods for determining the most appropriate pixel size of an area of interest, while Smith et al. (2006) stated that the best spatial resolutions and neighborhood sizes are variable according to terrain features. Cavazzi et al. (2013) found that not always very fine spatial resolution promotes the greatest accuracy and McBratney et al. (2003) presented the ranges of spatial resolutions for different soil mapping scales, being values between 10 and 40 m appropriate for detailed soil maps, in accordance with the 30 m Aster DEM used in this work.

The findings of this work are in agreement with the statements of the GlobalSoilMap project (Arrouays et al., 2014) and Minasny and McBratney (2010) that the use of existing maps as source of information is one of the best alternatives to create more detailed digital soil maps in a faster and economic way.

3.4. CONCLUSIONS

Legacy soil data extracted by KnowledgeMiner from a detailed soil map and used to model soil classes distribution in Marcela Creek watershed outperformed decision trees and also made improvements on the existing soil map, due to helping to understand the pedologist's mental model embedded on the map.

This mapping technique allows creating a soil map of the area with similar environmental conditions surrounding the watershed, being this area 365 times larger than the reference watershed, with reasonable accuracy and, furthermore, it can contribute to low cost detailed soil mapping of similar areas mainly in developing countries, where the lack of financial investments in soil surveys is not uncommon.

3.5. REFERENCES

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. Soil Sci. Soc. Am. J. 77, 860. doi:10.2136/sssaj2012.0275

Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214-215, 101–113. doi:10.1016/j.geoderma.2013.09.023

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 239-240, 13–24. doi:10.1016/j.geoderma.2014.09.021

Alvarenga, C.C., Mello, C.R. de, Mello, J.M. de, Silva, A.M. da, Curi, N., 2012. Índice de qualidade do solo associado à recarga de água subterrânea (IQS RA) na Bacia Hidrográfica do Alto Rio Grande, MG. Rev. Bras. Ciência do Solo 36, 1608–1619. doi:10.1590/S0100-06832012000500025

Araújo, A.R., 2006. Solos da Bacia do Alto Rio Grande (MG): Base para estudos hidrológicos e aptidão agrícola. UFLA, Lavras.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. d. L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P. a., Thompson, J. a., Zhang, G.-L., 2014. GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties, in: Advances in Agronomy. pp. 93–134. doi:10.1016/B978-0-12-800137-0.00003-0

Ashtekar, J.M., Owens, P.R., 2013. Remembering Knowledge: An Expert Knowledge Based Approach to Digital Soil Mapping. Soil Horizons 54, 1–6. doi:10.2136/sh13-01-0007

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. doi:10.1016/j.geoderma.2009.07.010

Beskow, S., Norton, L.D., Mello, C.R., 2013. Hydrological Prediction in a Tropical Watershed Dominated by Oxisols Using a Distributed Hydrological Model. Water Resour. Manag. 27, 341–363. doi:10.1007/s11269-012-0189-8

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69. doi:10.1080/02626667909491834

Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 242, 238–249. doi: 10.1016/j.geoderma.2014.11.026

Bohner, J., McCloy, K.R., Strobl, J., 2006. SAGA – Analysis and Modelling Applications. Göttinger Geographische Abhandlungen.

Bou Kheir, R., Chorowicz, J., Abdallah, C., Dhont, D., 2008. Soil and bedrock distribution estimated from gully form and frequency: A GIS-based decision-tree model for Lebanon. Geomorphology 93, 482–492. doi:10.1016/j.geomorph.2007.03.010

Bou Kheir, R., Greve, M.H., Bøcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands

using classification-tree based models: The case study of Denmark. J. Environ. Manage. 91, 1150–1160. doi:10.1016/j.jenvman.2010.01.001

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, 1st ed. Chapman and Hall/CRC, New York.

Brown, 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. doi:10.2136/sssaj2011.0205

Brungard, C.W., Boettinger, J.L. 2010. Conditioned Latin hypercube sampling: Optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Netherlands, pp. 67–75. doi: 10.1007/978-90-481-8863-5_6

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68–83. doi:10.1016/j.geoderma.2014.09.019

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17–26. doi:10.1016/j.geoderma.2003.07.006

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. doi:10.1016/S0016-7061(02)00238-0

Bui, E.N., Moran, C.J., 2001. Dissagregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79–94.

Calderano Filho, B., Polivanov, H., Chagas, C. da S., Carvalho Júnior, W., Barroso, E.V., Guerra, A.J.T., Calderano, S.B., 2014. Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar. Rev. Bras. Ciência do Solo 38, 1681–1693.

Caten, A. Ten, Dalmolin, R.S.D., Ruiz, L.F.C., 2012. Digital soil mapping: strategy for data pre-processing. Rev. Bras. Ciência do Solo 36, 1083–1092. doi:10.1590/S0100-06832012000400003

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma 195-196, 111–121. doi:10.1016/j.geoderma.2012.11.020

Collard, F., Kempen, B., Heuvelink, G.B.M., Saby, N.P. A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Reg. 1, 21–30. doi:10.1016/j.geodrs.2014.07.001

Congalton, R. G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35-46. doi: 10.1016/0034-4257(91)90048-B.

Congalton, R. G., Green, K., 1999. Assessing the accuracy of remotely sensed data: Principles and practices. New York, Lewis Publishers, 160p.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991-2007, doi:10.5194/gmd-8-1991-2015.

Crivelenti, R.C., Coelho, R.M., Adami, S.F., Oliveira, S.R.D.M., 2009. Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo. Pesqui. Agropecu. Bras. 44, 1707–1715. doi:10.1590/S0100-204X2009001200021

Curi, N., Chagas, C. da S., Giarola, N.F.B., 1994. Distinction of agricultural environments and soil-pasture relationships in Mantiqueira Fields, MG., in: Carvalho, M.M., Evangelista, A.R., Curi, N. (Eds.), Developments of Pastures in Physiographical Region of Campos das Vertentes, MG. EMBRAPA: CNPGL, p. 127.

Curi, N., Franzmeier, D.P., 1984. Toposequence of Oxisols from the Central Plateau of Brazil. Soil Sci. Soc. Am. J. 48, 341–346. doi: 10.2136/sssaj1984.03615995004800020024x

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

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann. Stat. 28, 337–407. doi:10.1214/aos/1016218223

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347–1370. doi:10.1029/2002WR001426

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modelling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. 9, 421–432. doi:10.1080/02693799508902047

Giarola, N.F.B., 1994. Levantamento pedológico, perdas de solo e aptidão agrícola das terras na região sob influência do reservatório de Itutinga/Camargos (MG). ESAL, Lavras.

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. doi:10.1590/S0103-90162006000300008

Giasson, E., Hartemink, A.E., Tornquist, C.G., Teske, R., Bagatini, T., 2013. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado. Cienc. Rural 43, 1967–1973. doi:10.1590/S0103-84782013001100008

Giasson, E., Sarmento, E.C., Weber, E., Flores, C.A., Hasenack, H., 2011. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci. Agric. 68, 167–174. doi:10.1590/S0103-90162011000200006

Greve, M.H., Greve, M.B., Kheir, R.B., Bøcher, P.K., Larsen, R., McCloy, K., 2010. Comparing Decision Tree Modeling and Indicator Kriging for Mapping the Extent of Organic Soils in Denmark, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Netherlands, pp. 267–280. doi:10.1007/978-90-481-8863-5

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in

Bavarian forest soils. Geoderma 185-186, 37–47. doi:10.1016/j.geoderma.2012.04.001

Hartemink, A.E., McBratney, A., 2008. A soil science renaissance. Geoderma 148, 123–129. doi:10.1016/j.geoderma.2008.10.006

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D. A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398. doi:10.1016/j.geoderma.2004.06.007

Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. 32, 1283–1298. doi:10.1016/j.cageo.2005.11.008

Hengl, T., de Jesus, J.M., MacMillan, R. a., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., 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, e105992. doi:10.1371/journal.pone.0105992

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., 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. doi:10.1371/journal.pone.0125814

Holmes, K.W., Griffin, E.A., Odgers, N.P. 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. Soil Res. 53, 865-880. doi:10.1071/SR14270

IBGE, 2007. Manual Técnico de Pedologia, 2nd ed. IBGE, Rio de Janeiro.

Jafari, A., Khademi, H., Finke, P.A., Van de Wauw, J., Ayoubi, S., 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point

dataset in an arid region, southeastern Iran. Geoderma 232-234, 148–163. doi:10.1016/j.geoderma.2014.04.029

Jenny, H., 1941. Factors of soil formation: A System of Quantitative Pedology. McGraw-Hill Book Co., Inc., New York.

Kempen, B., Brus, D.J., Vries, F. De, 2015. Geoderma. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands 242, 313–329.

Kerry, R, Goovaerts, P., Rawlins, B.G., Marchant, B.P. 2012. Geoderma. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at regional scale. 170, 347–358. doi: 10.1016/j.geoderma.2011.10.007

Kim, D., Zheng, Y., 2011. Scale-dependent predictability of DEM-based landform attributes for soil spatial variability in a coastal dune system. Geoderma 164, 181–194. doi:10.1016/j.geoderma.2011.06.002

Lacarce, E., Saby, N.P.A., Martin, M.P., Marchant, B.P., Boulonne, L., Meersmans, J., Jolivet, C., Bispo, A., Arrouays, D., 2012. Mapping soil Pb stocks and availability in mainland France combining regression trees with robust geostatistics. Geoderma 170, 359–368. doi:10.1016/j.geoderma.2011.11.014

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90–99. doi:10.1016/j.geomorph.2011.06.026

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geogr. Inf. Sci. 11, 183–198. doi:10.1080/136588197242455

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

Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1, 14–23. doi:10.1002/widm.8

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232-234, 34–44. doi:10.1016/j.geoderma.2014.04.033

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, Geoderma. doi:10.1016/S0016-7061(03)00223-4

McBratney, A.B., Minasny, B., Viscarra Rossel, R., 2006. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. Geoderma 136, 272–278. doi:10.1016/j.geoderma.2006.03.051

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327. doi:10.1016/S0016-7061(00)00043-4

Mello, C.R. De, Curi, N., 2012. Hydropedology. Ciência e Agrotecnologia 36, 137–146. doi:10.1590/S1413-70542012000200001

Mendonça-Santos, M.L., Santos, H.G., 2007. The State of the Art of Brazilian Soil Mapping and Prospects for Digital Soil Mapping. in: Lagacherie, P.,

McBratney, A.B., Voltz, M. (Eds.). Digital Soil Mapping: An Introdutory Perspective 31, 39–54. doi:10.1016/S0166-2481(06)31003-3

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. doi:10.1590/0103-9016-2013-0416

Menezes, 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, 287-298.

Milne, G., 1935. Some suggested units of classification and mapping particularly for East African Soils. Soil Res. 4, 183–198.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Netherlands, pp. 429–436. doi:10.1007/978-90-481-8863-5

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil Attribute Prediction Using Terrain Analysis. Soil Sci. Soc. Am. J. 57, NP. doi:10.2136/sssaj1993.572NPb

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16, 533–549. doi:10.1080/13658810210138715

Motta, P.E.F., Curi, N., Silva, M.L.N., Marques, J.J.G. de S. e M., Prado, N.J.S., Fonseca, E.M.B., 2001. Levantamento pedológico detalhado, erosão dos solos, uso atual e aptidão agrícola das terras de microbacia piloto na região sob influência do reservatório da hidrelétrica de Itutinga/Camargos-MG. CEMIG, Belo Horizonte.

Mulder, V.L., de Bruin, S., Schaepman, M.E., 2012. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 21, 301–310. doi:10.1016/j.jag.2012.07.004

Nauman, T.W., Thompson, J. A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399. doi:10.1016/j.geoderma.2013.08.024

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregation and harmonisation of soil map units through resampled classification trees. Geoderma 214, 91–100. doi:10.1016/j.geoderma.2013.09.024

Olaya, V., Conrad, O., 2009. Geomorphometry in SAGA., in: Hengl, T., Reuter, H. I. (Eds.), Geomorphometry - Concepts, Sorftware, Applications, Developments in Soil Science, Elsevier Amsterdam, pp. 293-308.

R Development Core Team, 2009. R: A language and environmental for statistical computing.

Resende, M., Curi, N., Rezende, S.B., Corrêa, G.F., Ker, J.C. 2014. Pedologia: Base para a distinção de ambientes. Editora UFLA, Lavras.

Roudier, P., Hewitt, A.E., Beaudette, D.E., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints., in: McBratney, A.B. (Ed.). 5th Global Workshop on Digital Soil Mapping 2012: Digital Soil Assessments and Beyond. Sydney, pp. 227–231.

Rokach, L., Maimon, O. 2008. Data mining with decision trees - Theory and Applications. World Scientific Publishing Co., Singapore.

Scull, P., Franklin, J., Chadwick, O. A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Modell. 181, 1–15. doi:10.1016/j.ecolmodel.2004.06.036

Shi, X., Long, R, Dekett, R., Philippe, J., 2009. Integrating different types of knowledge for digital soil mapping. Soil Sci. Soc. Am. J. 73, 1682-1692. doi: 10.2136/sssaj2007.0158

Silva, S.H.G., Owens, P.R., Duarte de Menezes, M., Reis Santos, W.J., Curi, N., 2014. A Technique for Low Cost Soil Mapping and Validation Using Expert Knowledge on a Watershed in Minas Gerais, Brazil. Soil Sci. Soc. Am. J. 78, 1310-1319. doi:10.2136/sssaj2013.09.0382

Silva, S.H.G., Owens, P.R., Silva, B.M., César de Oliveira, G., Duarte de Menezes, M., Pinto, L.C., Curi, N., 2015. Evaluation of Conditioned Latin Hypercube Sampling as a Support for Soil Mapping and Spatial Variability of Soil Properties. Soil Sci. Soc. Am. J. 79, 603-611. doi:10.2136/sssaj2014.07.0299

Silva, V.A., Curi, N., Marques, J.J.G., Carvalho, L.M.T. De, Santos, W.J.R., 2013. Soil maps, field knowledge, forest inventory and Ecological-Economic Zoning as a basis for agricultural suitability of lands in Minas Gerais elaborated

in GIS. Ciência e Agrotecnologia 37, 538–549. doi:10.1590/S1413-70542013000600007

Smith, M.P., Zhu, a. X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137, 58–69. doi:10.1016/j.geoderma.2006.07.002

SoLIM, 2007. Knowledge Miner 1.0 User Manual, 46p. Available at http://solim.geography.wisc.edu/software/. Last accessed: May 25, 2015.

Soil Survey Staff, 1993. Soil Survey Manual. USDA, Washington, D.C. Handbook No 18.

Soil Survey Staff, 1999. Soil taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys., 2nd ed, Office. USDA-SCS, Washington, DC.

Sorensen, R., Seibert, J., 2007. Effects of DEM resolution on the calculation of topographical indices: TWI and its components. Journal of Hydrology 347, 79-89. doi: 10.1016/j.jhydrol.2007.09.001

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345. doi:10.1016/j.geoderma.2013.08.018

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253-254, 67–77. doi:10.1016/j.geoderma.2015.04.008

Tehrany, M.S., Pradhan, B., Jebur, M.N., 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J. Hydrol. 504, 69–79. doi:10.1016/j.jhydrol.2013.09.034

Teske, R., Giasson, E., Bagatini, T., 2014. Comparação do uso de modelos digitais de elevação em mapeamento digital de solos em Dois Irmãos, RS, Brasil. Rev. Bras. Ciência do Solo 38, 1367–1376. doi:10.1590/S0100-06832014000500002

Therneau, T., Atkinson, B., Ripley, B., 2015. R package " rpart ." Available at https://cran.r-project.org/web/packages/rpart/rpart.pdf. Last access: May 15, 2015.

Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: Effects on terrain attribute calculation and quantitative soil-landscae modeling. Geoderma 100, 67-89. doi: 10.1016/S0016-7061(00)00081-1

Thompson, J.A., Prescott, T., Moore;, A.C., Bell, J.S., Kautz, D., Hempel, F., Waltman, S.W., Perry, C.H., 2010. Regional Approach to Soil Property Mapping using Legacy Data and Spatial Disaggregation Techniques. in: Gilkes, R.J., Prakongkep, N. (Eds.). Proceedings of 19th World Congress Soil Science, Soil Solutions for a Changing World, 1-6 August, Brisbane, Australia, 17-20.

UFV-CETEC-UFLA-FEAM., 2010. Mapa de solos do Estado de Minas Gerais: legenda expandida. Belo Horizonte: Fundação Estadual do Meio Ambiente.

Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-

Roussillon (France). Geoderma Reg. 4, 20–30. doi:10.1016/j.geodrs.2014.11.003

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. Soil Sci. Soc. Am. J. 65, 1463–1472. doi:10.2136/sssaj2001.6551463x