# DIGITAL SOIL MAPPING APPROACH BASED ON FUZZY LOGIC AND FIELD EXPERT KNOWLEDGE

## Abordagem de mapeamento digital de solos baseado em lógica fuzzy e conhecimento de campo de especialista

**Michele Duarte de Menezes[1], Sérgio Henrique Godinho Silva[2], Phillip Ray Owens[3], Nilton Curi[2]**

**ABSTRACT**

In Brazil, soil surveys in more detailed scale are still scarce and necessary to more adequately support the decision makers for planning soil and environment activities in small areas. Hence, this review addresses some digital soil mapping techniques that enable faster production of soil surveys, beyond fitting continuous spatial distribution of soil properties into discrete soil categories, in accordance with the inherent complexity of soil variation, increasing the accuracy of spatial information. The technique focused here is knowledge-based in expert systems, under fuzzy logic and vector of similarity. For that, a contextualization of each tool in the soil types and properties prediction is provided, as well as some options of knowledge extraction techniques. Such tools have reduced the inconsistency and costs associated with the traditional manual processes, relying on a relatively low density of soil samples. On the other hand, knowledge-based technique is not automatic, and just as the traditional soil survey, the knowledge of soil-landscape relationships is irreplaceable.

**Index terms:** Digital soil mapping, soil prediction, conditioned Latin hypercube sampling, knowledge miner.

**RESUMO**

No Brasil, levantamentos de solos em escalas maiores ainda são escassos e necessários para dar apoio mais adequado ao planejamento de atividades relacionadas a solos e ambientes em áreas menores. Em consequência, este trabalho apresenta algumas técnicas de mapeamento digital de solos que permitem a produção mais rápida de levantamentos de solos, além de ajustar a distribuição espacial contínua das propriedades do solo em categorias discretas, de acordo com a complexidade inerente da variabilidade dos mesmos, aumentando a acurácia de informações espaciais. A técnica aqui enfatizada é baseada em sistemas que empregam o conhecimento de um especialista, sob uso de lógica fuzzy e similaridade de vetores. Para isso, é proporcionada a contextualização de cada ferramenta para a predição de classes de solos e suas propriedades, assim como algumas opções de técnicas para a aquisição de conhecimentos. Tais ferramentas têm reduzido a inconsistência e custos associados aos tradicionais procedimentos manuais, utilizando uma relativamente baixa densidade de amostragem. Por outro lado, a técnica baseada no conhecimento de especialistas não é automatizada, e, assim como no método tradicional de levantamentos de solos, o conhecimento das relações solo-paisagem é insubstituível.

**Termos para indexação:** Mapeamento digital de solos, predição de solos, amostragem em hipercubo latino condicionado, mineração do conhecimento.

**(Received in june 5, 2013 and approved in june 27, 2013)**

## INTRODUCTION

In Brazil, soil surveys in more detailed scale are still necessary because the lack of information or the small-scale existing maps do not adequately support planning and management of agricultural and environmental projects. Soil surveys or sampling schemes in a more detailed scale are common only in small areas, generally to attend specific projects (MENDONÇA-SANTOS; SANTOS, 2007). Since the traditional soil maps are manually produced, even on a GIS basis, and have as limitation the low speed and high production cost (ZHU et al., 2001), digital soil mapping is viewed as an opportunity to optimize soil mapping, employing more quantitative techniques for spatial prediction (MCBRATNEY; SANTOS; MINASNY, 2003), in which the accuracy or uncertainty has been measured and discussed, and that makes the pedologist mental model more explicit. In theory, the basis of predictive soil mapping is similar to traditional soil survey, since it is possible to use knowledge of soil-environment relations to make inferences (SCULL et al., 2003).

Various approaches have been used for fitting quantitative relationships between soil properties or types and their environment, in order to predict them (spatial

[1]Universidade Federal Rural do Rio de Janeiro/UFRRJ – Departamento de Solos – 23890-000 – Seropédica – RJ – Brasil – michele_duarte@ig.com.br
[2]Universidade Federal de Lavras/UFLA – Departamento de Ciência do Solo/DCS – Lavras – MG – Brasil
[3]Purdue University – Agronomy Department – West Lafayette – IN – USA

inference models). The models are divided into data-driven (Pedometry approach) and quantitative soil survey approach (knowledge driven). Pedometry approaches are more quantitative and automatic, mainly based on statistics, geostatistics, machine learning and data mining techniques. A dense scheme of sampling is often required. On the other hand, the knowledge driven approach tries to fit within the conventional soil survey and mapping framework, aiming to effectively utilize the soil scientist's knowledge (SHI et al., 2009).

Soil survey is a paradigm-based science that is based on the application of conceptual soil-landscape models, in which the hypothesis is that the location and distribution of soils in the landscape is predictable (HUDSON; HEUVELINK; ROSSITER, 1992). Such models rely on tacit pedological knowledge, generally acquired by systematic field observation of repeating relationships between soils types or properties and landform position (MACMILLAN; PETTAPIECE; bRIERLEY, 2005). Most of the information about soils is found in soil maps and respective legend or in the mind of the soil surveyor. Hudson, Heuvelink and Rossiter (1992) argued that soil survey was deficient for not expressing the scientific knowledge in a more formal and systematic way.

Thus, this review attempts to elucidate the use of expert systems under fuzzy logic and its application for predicting soil types and properties. Expert systems allow the use of existing data or expert knowledge of the pedologist in conjunction with statistical and mathematical approaches to generate soil information. Besides, they allow fitting continuous spatial distribution of soil properties into discrete soil categories, in accordance with the inherent complexity of soil variation, increasing the accuracy of spatial information (ZHU et al., 2001).

## EXPERT SYSTEMS

According to Dale, McBratney, Russell (1989), expert systems consist of ways to harvesting and engineering knowledge, which allow exploiting the information of soil surveyor acquired through experience. Expert knowledge systems try to capture tacit knowledge and integrate it in the predictive model in order to improve it. Dale, McBratney, Russell (1989) delineated the components of an expert system to soil data: a source (e.g. data or environmental variables), an organizer and an information predictor, and a client to use the information. The predictor includes a knowledge base and an inference engine which operates on the knowledge base. The computer-based knowledge can use the human expert or numerical methods. Such approach is able to exploit soil surveyor knowledge by developing rule-based systems that imitate the surveyor's conceptual model of soil variability (SCULL et al., 2003). The pioneering attempts to apply expert systems in Pedology used the Boolean logic (SKIDMORE et al.,1991), which defines a strict binary decision (true or false, 0 or 1). In terms of soil maps, the soil surveyor has to assign individual soils in the field in only one class (ZHU et al., 2001). The polygons of the maps, also referred to as crisp or Boolean, represent only the distribution of a set of prescribed soil class (central concepts of soils). The same approach is used for soil property maps, where the whole polygon assumes a property value assigned to the mapping unit.

## FUZZY LOGIC

The nature of soil-landscapes are complex, whose changes in soils or properties are often more gradual and continuous, differently to the variation represented by a crisp map (polygon-based) (Figure 1a). There is uncertainty in the boundaries allocation, as well as in the values of the soil properties (LEGROS, 2006) (Figure 1b). Fuzzy logic attempts to represent the uncertainty in the predictor and predicted properties or types, as an alternative that seems more adapted to the imprecise knowledge conveyed by soil surveyors (WALTER; LAGACHERIE; FOLLAIN, 2007), recognizing the concept of partial truth, alternatively to the subjective rigidity imposed on soils.

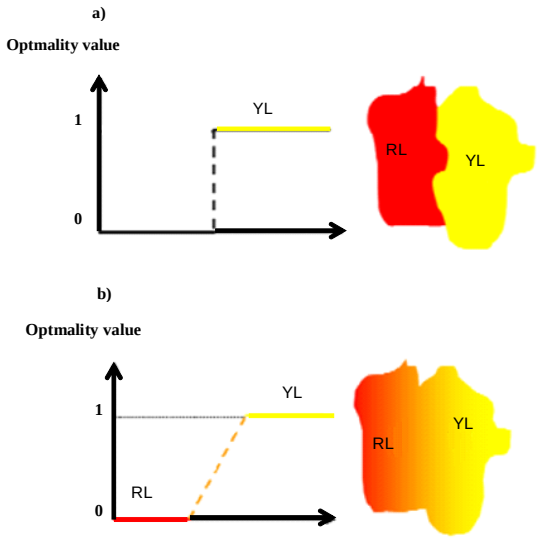


Figure 1 – Lateral distribution of an optimal value under Boolean logic (a) and fuzzy logic (b) related to distribution of Yellow Latosol (YL) and Red Latosol (RL) in the landscape.

Instead of a crisp membership (e.g., entirely Red Latosol or Yellow Latosol, Figure 1a), the idea is that the soils in nature rarely fit exactly the classification types to which they are assigned (ZADEH, 1965). Nevertheless, there is a range of optimal values among classes. The concept of belonging to a set has been modified to include partial degrees of membership. The maximum membership is often 1 and represents the central or modal concept, whereas the 0 value expresses no membership. Values in between this range express different degrees of similarity to the central concept.

Besides the broad application of fuzzy logic in science, Scull et al. (2003) cited two different approaches for soil prediction in a continuous way: the first is based on the fuzzy-k-means classifier, which partitions observations in multivariate space into natural classes. The second is known as the semantic import model, and is used in situations when classification schemes are pre-defined and class limits are relatively well understood. The semantic model is commonly used with expert knowledge and it refers to a data integration concerned with analysis and interpretation of a multi-source spatial data. In geographic analysis, it is frequently required the integration of spatial data with multi-sources (as raster or vector formats, crisp or continuous maps) to answer specific questions about given spatial phenomenon. In this sense, Zhu and Band (1994) presented the first approach which employs knowledge-based semantic data integration, combined with expert system techniques and fuzzy set theory for spatial data integration.

A fuzzy logic based model called similarity vector (ZHU et al., 1997) represents soils at a given location, in which the landscape is perceived as a continuum. The fuzzy logic is used to infer the membership of a soil type from environmental variables, such as parent material, canopy coverage, digital elevation model and its derivative maps. Under fuzzy logic, a soil at a given pixel (*i,j*) is represented by a *n*-element similarity vector: $S_{ij} = (S_{ij}^1, S_{ij}^2, ..., S_{ij}^k, ..., S_{ij}^n)$, where *n* is the number of prescribed soil types over the area and $S_{ij}^k$ is an index which measures the similarity between the local soil at (*i,j*) to the prescribed soil type *k*. $S_{ij}^k$ is soil type *k*. The similarity value is measured according to how close the soil is to centroid concept (between 1 and 0, as already discussed). The more similar a soil is to a prescribed soil type, the higher its similarity value (fuzzy membership).

This methodology has been successfully applied to generate soil maps (crisp maps) (ZHU; BAND, 1994; ZHU et al., 1996; MCKAY et al., 2010) and to predict properties in a continuous way, as depth of A horizon (ZHU et al., 1997), *solum* depth (QUINN; ZHU; BURT, 2005; LIBOHOVA, 2010), drainage classes (MCKAY et al., 2010), A horizon silt and sand contents (QI et al., 2006), soil transmissivity (ZHU et al., 1997) or aquifer recharge potential, which is a spatially distributed phenomenon and closely related to soil-landscape potential.

**SoLIM (Soil-Land Inference Model) and ArcSIE (Soil Inference Engine)**

In order to overcome some limitations of a traditional soil survey, researches and tools have applied knowledge-based techniques and fuzzy logic concepts as a predictive approach, for instance, the softwares SoLIM (ZHU et al., 2003) and ArcSIE (SHI, 2013). They have two major components: a similarity model for representing soil spatial variation and a set of inference techniques for populating the similarity model. The improvements of the last versions also contain means of extracting rules (expert knowledge extraction). Hereafter is provided a review about the potential of some tools to predict soil types and properties.

**ArcSIE**

ArcSIE works as an extension of ArcMAP (ArcGIS - Environmental Systems Resource Institute). There are two inference methods implemented in ArcSIE for calculating fuzzy membership values: rule-based reasoning (RBR) and case-based reasoning (CBR). In other words, rule and case are two types of knowledge supported by ArcSIE. In RBR, rules are created from direct specifications of soil surveyor, while in CBR it represents the knowledge of the soil at a specific location, also called tacit points.

**Ruled-based reasoning with ArcSIE (RBR)**

Rule-based reasoning (RBR) in ArcSIE can be useful when the soil scientist knows the soil-landscape relationships and prescribes, under certain environmental conditions, where a specific soil type is more likely to occur. The premise of this technique is that one or two factors out of the five state factors (parent material, climate, organisms, time and topography, JENNY, 1941) control the distribution of soils on the landscape. For example, when climate, organisms, parent material, and time are relatively constant, the topography would be the greatest driver for soil differentiation. Continuous variation of soils are represented by continuous soil property maps derived from

the similarity vectors (ZHU et al., 1997) and a lower number of sample points is required (only one typical value per soil type). The following steps are required in order to predict soil types or properties (adapted from LIBOHOVA, 2010) (Figure 2).

**Establishing soil-landscape relationships**

In order to establish the soil-landscape relationships, Zhu and Band (1994) used the knowledge drawn by a certified soil scientist in his domain expert, since he was working in the study area. Libohova (2010) used previous soil surveys and block diagrams from the county soil survey to provide visual insight into the soil-landscape model established by the field soil scientist. In Brazil, where soil series have not been established so far, it could be used information from previous soil survey reports and scientific papers, which detailed the topographic sequence of soils. For a better comprehension of spatial distribution of soils, it is required the integration of pedologic studies with other branches of science, especially Geology (stratigraphy), Geomorphology and Hydrology (VIDAL-TORRADO; LEPSCH; CASTRO, 2005). The analysis of the phenomenon studied by these disciplines and tis results can help in pedologic investigations, collaborating for a better soil sampling and interpretations.

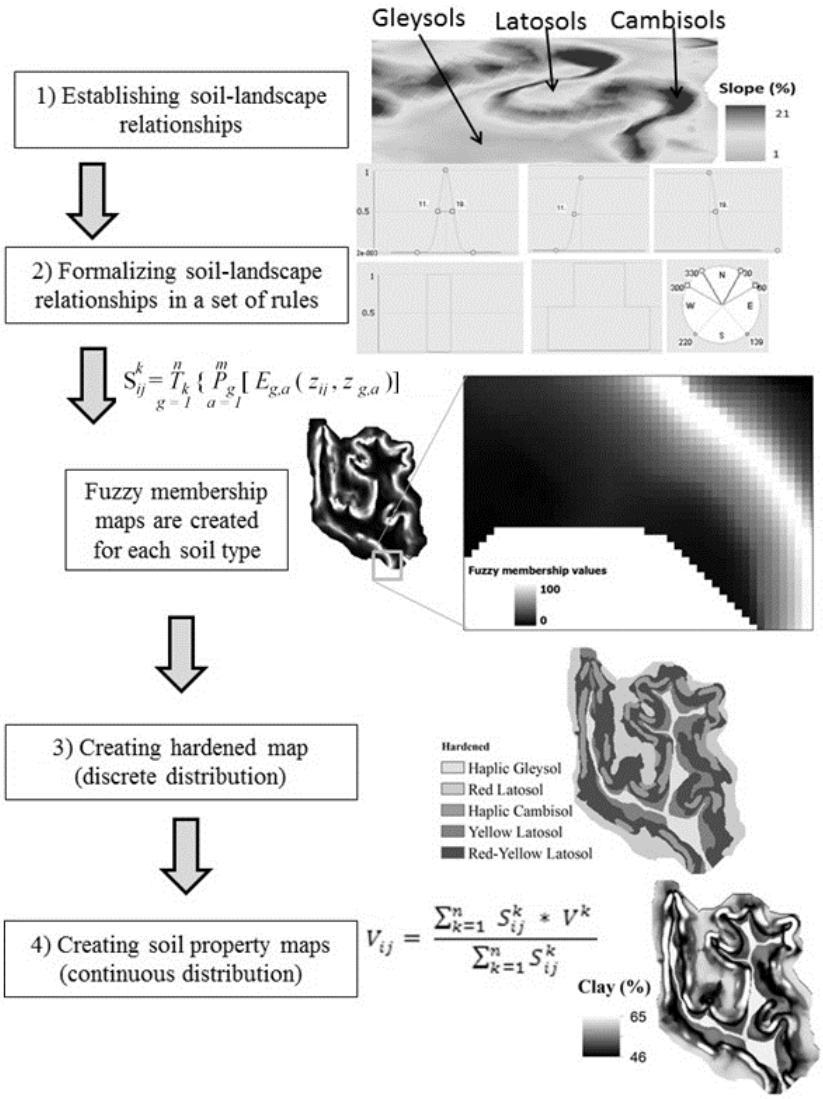


Figure 2 – Steps required for rule-based reasoning.

**Quantifying relationships between soils and terrain attributes and formalizing these relationships in a set of rules that relates to raster maps**

ArcSIE provides tools for soil scientists to formalize the relationships based on pedological knowledge of the local soils. In this case, the inference is based on rules using fuzzy logic. Threshold values are identified and assigned to each soil map unit in a GIS basis. For this, data layers in a raster format that characterize environmental covariates, as terrain attributes, geology, vegetation, climate and others are prepared (SHI et al., 2009). Then, the knowledge about soil-landscape relationships in the first step is qualitatively modeled on a continuous basis in a set of created rules.

The values of the environmental covariates and ranges associated with each soil map class (rules) are used to define membership functions, which in turn are referred to as optimality functions as they define the relationships between the values of an environmental feature and a soil type. The rules are set within the software based on "if-then" statements, in which a central location encompasses the rules that provide 100% probability of meeting the class. As the covariates get further from meeting all the rules, the probability of the location being in that class changes and alters the soil property prediction. The number of rules is not limited and information, such as land-use derived from remotely sensed data, can be inserted as a rule and the predictions altered based on the land use type. The cutoffs are set based on knowledge from a soil scientist who understands the soil-landscape relationships.

The initial output from the inference is a series of fuzzy membership maps in raster format, one for each soil type under consideration (Figure 2). The fuzzy membership values represent the similarities of each place in the landscape to those soil types. The equation below describes how the knowledge of a given soil type is used for a global knowledge in RBR and CBR in order to create fuzzy membership values, represented by three functions (*E*, *P*, and *T*) (SHI et al., 2004):

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

where $S_{ij}^{k}$ is the fuzzy membership value at a location $(i, j)$ for a soil *k*. The *m* is the number of environmental features used in the inference. The *n* is the number of instances for soil type *k*. $Z_{ij,a}$ is the value of the $a^{th}$ environmental feature at location $(i,j)$. $Z_{g,a}$ is the most optimal range given by rule or case *g*, defining the most favorable condition of feature *a* for soil *k*. In RBR it is directly specified by the soil scientist, while in CBR, it is derived by the computer based on the case location and the environmental data layers. *E* is the function for evaluating the optimality value at the environmental features level. If $Z_{ij,a}$ falls into the range of $Z_{g,a}$, *E* returns the maximum optimality value; otherwise, *E* uses a function to derive the optimality value based on the difference between $Z_{ij,a}$ and $Z_{g,a}$. Based on the nature of the environmental covariates used in the prediction, there are five choices for E: *cyclic*, *ordinal*, *nominal*, *raw values*, and *continuous* (bell-shape, z-shape and s-shape continuous curves). *P* integrates the optimality values from individual environmental covariates to generate an overall predicted value for soil *k*. *T* is the function for deriving the final fuzzy membership value for soil *k* at site *(i,j)* based on all the instances for soil *k*.

Using this toolbox, the parameters are adjusted to the curves and explicitly express the mental model of the pedologist. Accomplishing this step, fuzzy membership maps are created (Figure 2, step 2). These maps reveal more details at the spatial level than polygon maps. According to Zhu et al. (1996), the general shapes on the membership images follow the landscape better than the ones on the soil maps where inclusion or exclusion from a region is more based on restrictions derived from the scale of the map than on local conditions. The central concept of the soil type responds to local variations in the apparent soil forming environment (represented by covariables).

**Creating hardened map**

The fuzzy membership maps (Figure 2, step 3) for each soil type are aggregated in order to create a hardened or a defuzzified map, which corresponds to the traditional soil vector map (discrete distribution). For that, the ArcSIE assigns at each pixel the soil type with the highest fuzzy membership value.

**Creating soil property maps**

The soil-landscape relationships are extracted and the characterized environmental conditions are linked through a set of inference techniques to populate the similarity model for a given area (ZHU; MCKAY, 2001). Thus, based on fuzzy membership values, the continuous variation of soil properties can be derived from the similarity vectors, using the following formula (ZHU et al., 1997):

$$V_{ij} = \frac{\sum_{k=1}^{n} S_{ij}^{k} * V^{k}}{\sum_{k=1}^{n} S_{ij}^{k}}$$

where $V_{ij}$ is the estimated potential of recharge value at location *(i,j)*, $V^k$ is a typical value of soil type *k* (e.g. Haplic Cambisol under native forest), and *n* is the total number of prescribed soil types for the area. If the local soil formative environment characterized by a GIS resembles the environment of a given soil category, then property values of the local soil should resemble the property values of the candidate soil type. The resemblance between the environment for local soil at (*i*,*j*) and the environment for soil type *k* is expressed by $S_{ij}^k$, which is used as an index to measure the level of resemblance between the soil property values of the local soil and those of soil category (ZHU et al., 2001). The property value can be any property that shows a recognizable pattern or relationship with the terrain attribute or landscape position (LIBOHOVA, 2010). The higher the membership of a local soil in a given soil type, the closer the property values (potential of recharge) at that location will be to the typical property values (ZHU et al., 2010).

**Case-based Reasoning with ArcSIE (CBR)**

CBR, in general, is a method of solving problems based on similar problems solved in the past. Dutta and Bonissone (1993) define better this type of methodology as the action of solving new problems by identifying and adapting similar problems stored in a library of past experiences.

CBR has been applied to soil science in association with fuzzy logic in order to solve problems related to soil data extrapolation. CBR emerges as an alternative to the RBR, since the formulation of rules to explain soils variability becomes laborious, even possessing the knowledge, motivating a search for alternative solutions, being one of them also provided by ArcSIE.

For instance, from a set of points (ArcSIE also works with lines, polygons and rasters as sources of information) with x, y coordinates distributed within a study area and a set of environmental covariates layers (GIS data layers), ArcSIE can extract information from each environmental covariate layer at the site where each point is located, and then associate the points classified as the same soil type with their environmental covariate values of occurrence. For example, considering two soil types (A and B), each one containing 8 and 10 sample points, respectively, and two environmental covariate layers (elevation and slope). The information obtained would be 8 slope and elevation values for soil A and 10 ones for soil B. Thus, one could predict soil properties in no sampled places according to the relationships between environmental data and soil properties. In this example of CBR use, the "former problems" would be the sampled locations, and from them, other places ("new problems") would be classified based on membership approaches characteristics of fuzzy logic.

It has been noticed that a minimum sample size covering the different combinations among environmental covariates has to be reached to allow the data extrapolation. If not, places with environmental combinations not included in the set of points would not be classified, as in the example of figure 3. In this case, the watershed is located at a mountainous region with dense rain forest vegetation, which hampers the full access to visit and sample soil. Thus, the same property map could be successfully generated with the use of RBR, since this watershed has been intensely studied. Thus, the knowledge could make up the low density of samples.

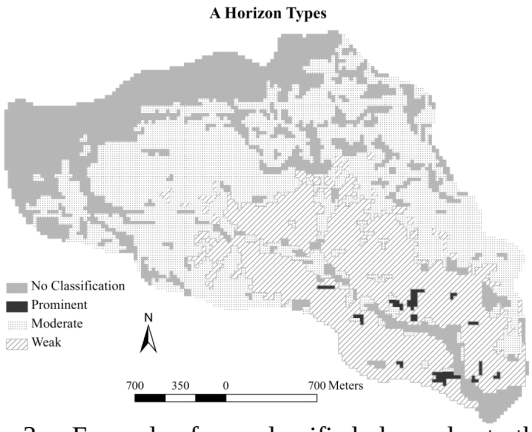


Figure 3 – Example of non-classified places due to the absence of data covering all the environmental features of the interest area.

**CONDITIONED LATIN HYPERCUBE SAMPLING SCHEME**

The necessity of finding out the optimal sampling method in order to adequately represent the soil variability within an area has generated many suggestions by soil scientists for years. Over the past decades, extensive work has been published on sampling schemes for soil mapping (MULDER; BRUIN; SCHALOMAN, 2013). Additionally, especially in developing countries, the number of samples for a soil survey is limited not only by access difficulties, but also by time and money restrictions, which hampers the sampling representativeness of the area and influence the final soil map. Also, this scenery would not allow the

use of CBR for not covering all of the ranges of the environmental covariates.

In this context, Minasny and McBratney (2006) proposed the conditioned Latin Hypercube Sampling (cLHS), derived from Latin Hypercube Sampling (LHS) McKay et al. (1979), and it has been used in soil science and environmental studies for assessing the uncertainty in a prediction model (MINASNY; MCBRATNEY, 2002). LHS is a stratified random procedure that provides an efficient way of sampling variables from their multivariate distributions (MINASNY; MCBRATNEY, 2006). It follows the idea of a Latin square where there is only one sample in each row and each column, generalizing this concept to an arbitrary number of dimensions. Also, the number of samples desired is taken into account at the time of determining the sampling locations. According to Mulder et al. (2013), if *n* is the desired sample size, LHS stratifies the marginal distributions of the covariates into *n* equally probably intervals and randomly samples the multivariate strata such that all marginal strata are included in the sample. However, it may face the issue that sometimes the sampling local may not exist in the field.

In this context, the conditioned Latin Hypercube Sampling (cLHS) adds the condition that the sample chosen must actually occur on the landscape (BRUNGARD; BOETTINGER, 2010). Minasny and McBratney (2006) showed that cLHS closely represented the original distribution of the environment covariates with relatively small sample sizes in a digital soil mapping project in the Hunter Valley of New South Wales, Australia.

Small sample sizes able to represent the soils variability is interesting especially for soil scientists from developing countries, where investments and time availability, area accessibility and former soil information are scarce. However, Mulder et al. (2013) highlight that, while LHS is probability sampling, conditioning the LHS on any constraints and sampling costs leads to a purposive sampling strategy since the inclusion probabilities of locations are modified by the conditioning criteria.

The cLHS may distribute the samples throughout the study area, but, sometimes, some places are very difficult or even impossible to be visited for sampling. To avoid this situation, Roudier, Beaudette and Hewitt (2012) proposed a method for incorporating operational constraints into cLHS. They created a "cost" map representing the cost of reaching every place on the landscape considering terrain and landcover attributes. The mentioned work showed that a cost-constrained LHS is not as optimized as the one without cost-conditionings, but the cost of the produced sampling scheme was reduced, thus providing an alternative to implement it.

We used the cLHS constrained by a cost map (created according to the distance from roads, slope and vegetation cover) to indicate the sampling places for validating a rule-based Cambisol *solum* depth map created through fuzzy logic (RBR) and terrain derivative maps in a watershed of Minas Gerais State, Brazil. An illustration of the sampling locals disposal with and without cost-constraining the sampling scheme is presented in (Figure 4). Also, the cLHS indicated sampling places with different soil properties, such as *solum* depth, soil moisture and color, and amount of pebbles and gravels, providing a good idea of the soil properties distribution along with the landscape features within the study area and, mainly, this sampling scheme reduced the time and investments needed for the field work.

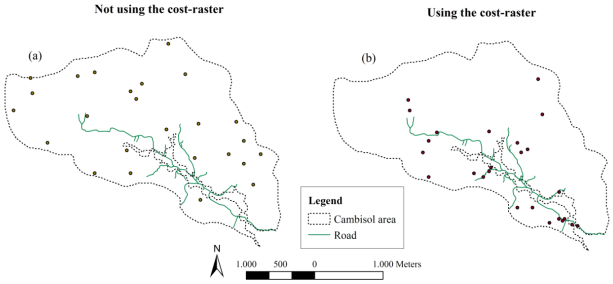


Figure 4 – Conditioned Latin Hypercube Sampling scheme without considering the cost-constrained raster (a) and considering the cost-constrained raster (b) for locating the sampling places in a Cambisol area.

## ACQUIRING INFORMATION FROM EXISTING SOIL MAPS FOR SOIL DATA TRANSFERABILITY

Nowadays, there is a plenty of covariates or layers that can be used to predict soil types and properties, derived from remote sensing, digital elevation models from topographic surveys, geomorphometric variables, analogical or digital soil maps, and others. McKay et al. (2010) investigated potential data layers involved using visual assessment and comparison to known soil locations by expert scientists. Even if the soil-landscape relationships are well known, it could be a hard task to find out which covariate would be more appropriate to tell soil types apart for predictions. From an existing soil map, SoLIM and ArcSIE provide tools for a soil scientist to discover the knowledge implicitly represented by an existing soil map and revise the discovered knowledge.

So, it would be possible to transfer the extracted knowledge to other areas with similar soil-landscape relationships.

Transferability of soil types or rules for predicting properties from one small area to a larger extent can be done if the digital soil mapper knows that the initial area is representative of the larger extent (MCBRATNEY et al., 1993). LAGACHERIE et al. (2001) applied this concept for extrapolating French Mediterranean soilscapes (combination of soil-forming factors in a buffer neighbor can be expressed as a vector composition of elementary landscape classes of different sizes). McKay et al. (2010) applied an accurate transferability of knowledge-based model to predict soil series and drainage classes between similar soil-landscape relationship areas. Such concept along with knowledge mining, fits with the scenery of soil surveys in Brazil, where detailed and semidetailed types are available in small areas to support local specific agricultural and environmental projects (MENDONÇA-SANTOS; SANTOS, 2007), but the necessity of more detailed soil maps in large extensions still remains. Hereafter two ways of extracting knowledge are presented.

**SoLIM Knowledge Miner**

According to Bui (2004), soil maps represent the structured mental soil-landscape model. One way to exploit such information is provided by SoLIM software (ZHU; BAND, 1994; ZHU et al. 1996; ZHU, 1997; ZHU et al., 1997). The knowledge acquisition tool allows the users to extract pixels information from the terrain derivative maps for each polygon (mapping unit). In this context, occurrence rules for each soil type could be formulated by a soil expert in association with SoLIM knowledge acquisition tool and then transferred to a similar area to identify the places more likely to find similar soil types.

One potential application of that is in areas with limited or no soil data availability, but with some soils similarity, especially in terms of environmental factors that influence the soil formation, to another area with already existing soil maps. They could be used as a source of data for predicting soil information (MCBRATNEY et al., 2003). From an existing map, which contains the surveyor knowledge about the distribution of soils on the landscape, and employing GIS data, models could be adjusted through the analysis of terrain derivative maps, such as slope, wetness index, aspect, profile curvature and so forth, which are supposed to explain the different soil types occurrence in an area based on the catena concept (MILNE, 1935): soil profiles occurring on topographically associated landscapes will be repeated on similar landscapes. This should permit soil data transferability as a manner of assuming soil patterns in the no-data area, based on soil scientist knowledge and soil-landscape models. Zhu et al. (2001) states that the soil-landscape concept contends that if one knows the relationships between each soil and its environment within an area, then one is able to infer what soil might be at each location on the landscape by assessing the environmental conditions at that point.

For instance, it is well-known that the Gleysols are more likely to occur in low elevation and concave places, with high water accumulation (RESENDE et al., 2007), but it should be difficult to tell the values of wetness indexes or concavity in order to separate those places from the surrounding areas. Likewise, Cambisols are more likely to be found under steep relief, but how steep the topography should be in an area of interest to determine the places representative of Cambisols could be hard to tell. Thus, a tool proposed by Zhu et al. (1997) that extracts the values of those terrain derivatives could help to understand soil types occurrence pattern and, hence, to extrapolate soil types distribution from a mapped area to a similar one which does not have soil data.

SoLIM software contains a knowledge acquisition tool which allows the users to extract pixels information from the terrain derivative maps for each polygon. Regarding a soil map, polygons should represent different mapping units. Thus, through the use of terrain derivative maps, SoLIM provides a way to acquire soil information from environment characteristics, helping to comprehend how the soil data were extrapolated to non-sampled places. This tool also generates graphics from the values of each terrain derivative map for each mapping unit. This would inform the user whether the mapping units are overlapping or not for each terrain derivative map. This latter result would allow the user to classify an area with no soil data based on environmental similarities of different areas through correlations between soil types and terrain attributes.

As an example, a watershed located in Nazareno county, in Minas Gerais State, Brazil, contains Latosols (Oxisols) in association with Cambisols (Inceptisols) on high lands, and Gleysols in low elevation areas (MOTTA et al., 2001). Using the SoLIM tool, it was possible to extract the pixel values of altitude above the channel network (AACN) map over the soil units, as shown on figure 5. Both curves are not presenting large overlapping areas: Gleysols, as expected, present lower AACN values, basically inferior to 10, the contrary of Red Oxisols. This graphic setting those curves apart contributes to a better understanding of the soil types correlation to AACN. In this context, occurrence rules for each soil type could be formulated by a soil expert in association with SoLIM knowledge acquisition tool and then transferred to a similar area to identify the places more likely to find similar soil types.

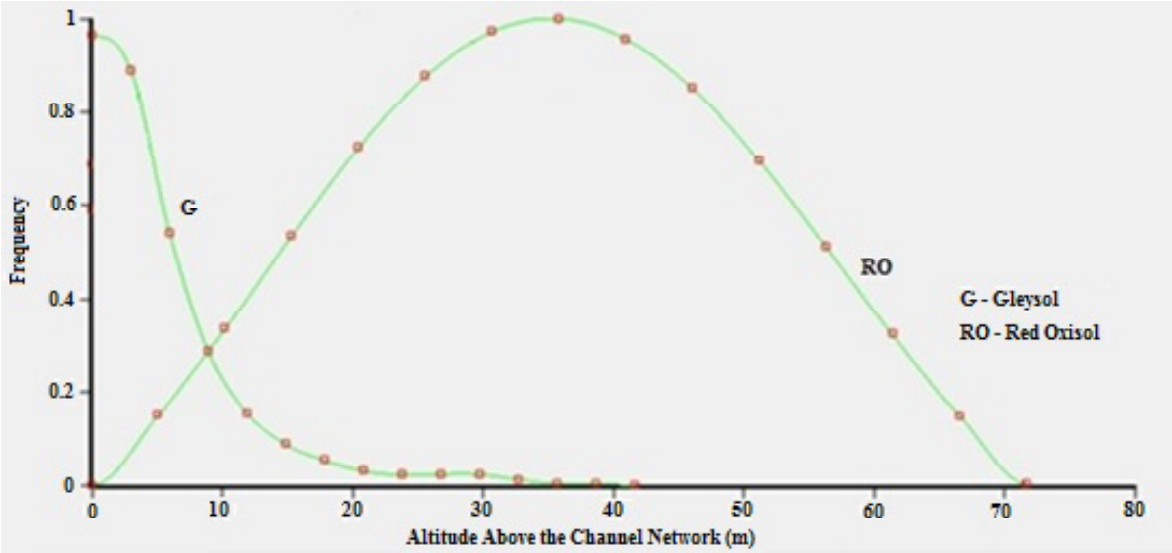


Figure 5 – Graphic showing the pixel frequency distribution (from 0 to 1) for Gleysols and Red Oxisols over altitude above the channel values.

**Boxplots**

Boxplots are another way to visualize the differences between pixel values of terrain derivative maps for different soil types and also to verify how adequate the extraction of information from existing maps was. They may show the overlapping values and present the differences or similarities of quartiles and medians according to different terrain derivatives and, thus, it makes it possible to identify the best environmental covariate for predicting soil properties.

In order to illustrate this identification tool, Figure 6 shows boxplots of four different mapping units (1, 2, 3 and 4) and four terrain derivatives (slope, profile curvature, wetness index and AACN) of a watershed in Minas Gerais State, Brazil. They were created using the R software (R DEVELOPMENT CORE TEAM, 2013).

Analyzing the boxplots, some overlapping of ranges in values can be seen for slope data although the medians are well separated. Wetness index boxplots for mapping units 1 and 2 are entirely overlapping in values, as well as for 3 and 4 ones, indicating that this terrain attribute would not succeed in separating all the mapping units occurrence. The least overlapping of values is pursued for better understanding the mapping methodology to represent the soils distribution on the landscape.

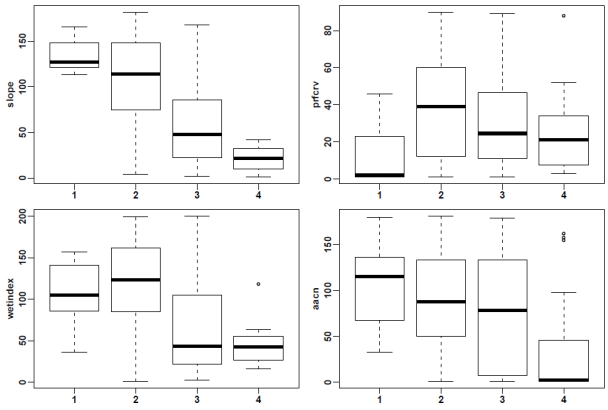


Figure 6 – Boxplots for terrain derivatives and mapping units. prfcrv - profile curvature, wetindex - wetness index, aacn - altitude above the channel network.

**FINAL CONSIDERATIONS**

The tools presented in this review have a potential for faster production of soil surveys, since the techniques reduce the inconsistency and costs associated with the traditional manual processes (ZHU et al., 2001). Also, when compared with pedometric approaches, a low density of soil samples is necessary. On the other hand, knowledge-based technique is not automatic, and just as the traditional soil survey, the knowledge of soil-landscape relationships is necessary,

and its use has been considered as efficient as economical (HUDSON, 1992; MCMILLAN; MOON; COUPE, 2007).

As raised by Hudson (1992), the soil survey has so far failed in not expliciting the mental model of the soil surveyor. Expliciting the rules in functions to get optimality values, as well as the use of knowledge miner techniques in order to utilize the legacy data, can contour this limitation of traditional soil survey. Once the knowledge is explicit, extracted or established in reference areas, the transferability to larger areas within same soil-landscape relationships should be tested (MCKAY et al., 2010), as an opportunity to raise the geographic expression of surveyed areas, very much needed in Brazil.

Since fuzzy membership maps represent soil types and can be viewed as a non-linear transformation of environmental variables based on expert knowledge of a soil-landscape model (ZHU et al., 2010), its use as an auxiliary in soil property prediction should be more explored. One example of such application is related to pedometric prediction methods. Those that do not incorporate the use of auxiliary variables (interpolation relying only on point observations) have been outperformed by hybrid methods (interpolation relying on point observations combined with interpolation based on regression of the target variable on spatially exhaustive auxiliary information). Hybrid methods explore the extra information when there is auxiliary information (maps of covariates related to terrain, land use, and others) able to explain part of variation (HENGL; HEUVELINK; ROSSITER, 2007). In this sense, Zhu and Lin (2010) compared maps generated from linear regression and environmental variables with regression using fuzzy membership maps as auxiliary. The non-linearity and complexity inherent to the steeper terrain with more variable soil types were well captured by a set of soil membership maps, which can be used to describe model and non-linear variation of soil property values. The linear regression using environmental variables would be more appropriate to be used on gently rolling landscapes, where soil-environment model is simple and stable over space.

Finally, the mapping tools presented in this work show the advantages of associating them to the field expert-knowledge in order to enhance the final results quality. Along with that, however, it is worthy to remind that these tools should be used on soil surveys and mapping to assist the field work (and never in order to replace it), mainly because the soil variability is not completely predictable, which makes this field activity irreplaceable for soil mapping.

**ACNOWLEDGEMENTS**

The authors thank CNPq for sponsoring both the Ph.D. of the first author and the Science without Boarders participation of the second author, in which the developed projects were results of an extensive partnership between Purdue University (USA) and UFLA.

**REFERENCES**

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. **Digital Soil Mapping**: Bridging Research, Environmental Application, and Operation (Progress in Soil Science 2). New York: Springer Science+Business Media B.V., v. 2, 2010, p. 67-75.

BUI, E.N. Soil survey as a knowledge system. **Geoderma**, Amsterdam, v.120, n.1–2, p. 17–26, 2004.

DALE, M.B.; MCBRATNEY, A.B.; RUSSELL, J.S. On the role of expert systems and numerical taxonomy in soil classification. **Journal of Soil Science**, hoboken, v.40, n.2, p.223–234, 1989.

DUTTA, S.; BONISSONE, P.P. Integrating case and rule-based reasoning. **International Journal of Approximate Reasoning**, New York, v.8, p.163-203, 1993.

HENGL, T.; HEUVELINK, G.; ROSSITER, D.G. About regression-kriging: from equations to case studies. **Computer and Geosciences**, London, v.33, n.10, p.1301-1315, Oct. 2007.

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

JENNY, H. **Factors of soil formation**. New York: McGraw-Hill, 1941, 281p.

LAGACHERIE, P. et al. Mapping of reference area representativity using a mathematical soilscape distance. **Geoderma**, Amsterdam, v.101, p.105-118, 2001.

LEGROS, J.P. **Mapping of the Soil**. Enfield: Science Publishers, 2006, 411p.

LIBOHOVA, Z. **Terrain attribute soil mapping for predictive continuous soil property maps**. Ph.D. Thesis. West Lafayette: Purdue University, 2010, 122 p.

MCBRATNEY, A.B. et al. Digital soil mapping. In: HUANG, P.M.; YUNCONG, L.; SUMNER, M.E. (eds). **Handbook of soil sciences**: properties and processes. 2nd ed. Boca Raton: CRC Press, 1993, 1442p.

MCBRATNEY, A.B.; SANTOS, M.L.M.; MINASNY, B. On digital soil mapping. **Geoderma**, Amsterdam, v.117, n.4, p.3-52, Jun, 2003.

MCKAY, M.D.; BECKMAN, R.J.; CONOVER, W.J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. **Technometrics,** Alexandria, v.21, p.239–245, 1979.

MCKAY, J. et al. Evaluation of the transferability of a knowledge-based soil-landscape model. In: BOETTINGER, J.L. et al. (eds.). **Digital soil mapping:** bridging research, environmental application, and operation. London: Springer, 2010, p. 165-177.

MACMILLAN, R.A.; MOON, D.E.; COUPE, R.A. Automated predictive ecological mapping in a Forest Region of B.C., Canada, 2001–2005. **Geoderma**, Amsterdam, v.140, p.353–373, 2007.

MACMILLAN, R.A.; PETTAPIECE, W.W.; BRIERLEY, J.A. An expert system for allocating soils to landforms through the application of soil survey tacit knowledge. **Canadian Journal of Soil Science**, Ottawa, v.85, p.103-112, 2005.

MENDONÇA-SANTOS, M.L.; SANTOS, H. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M. (eds). **Developments in Soil Science**. New York: Elsevier, v.31, p.39-54, 2007.

MENEZES, M. D. **Levantamento pedológico de hortos florestais e mapeamento digital de atributos físicos do solo para estudos hidrológicos**. Ph.D. Thesis. Lavras: UFLA, 2011, 225p.

MILNE, G. Some suggested units of classification and mapping particularly for East African soils. **Soil Research**, Vitoria, v.4, p.183-198, 1935.

MINASNY, B; MCBRATNEY, A.B, Uncertainty analysis for pedotransfer functions. **European Journal of Soil Science**, Hoboken, v.53, 417–430, 2002.

MINASNY, B.; MCBRATNEY, A.B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers and Geosciences**, Oxford, v.32, n.9, p.1378–1388, 2006.

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.; BRUIN, S.; SCHAEOMAN, M.E. Representing major soil variability at regional scale by contrained Latin Hypercube Sampling of remote sensing data. International **Journal of Applied Earth Observation and Geoinformation**, Amsterdam, v.21, p.301-310, 2013.

QI. F.; et al. Fuzzy soil mapping based on prototype category theory. **Geoderma**, Amsterdam, v.136 p.774–787, 2006.

QUINN, T.; ZHU, A.X., BURT, J.E. Effects of detailed soil spatial information on watershed modeling across different model scales. **International Journal of Applied Earth Observation and Geoinformation,** Amsterdam, v.7, p.324–338, 2005.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing. Available at: <http://www.r-project.org> Accesses in: 24 May. 2013.

RESENDE, M. et al. **Pedologia**: base para distinção de ambientes. 5.ed. Lavras: UFLA, 2007. 322 p.

ROUDIER P.; BEAUDETTE, D.E.; HEWITT, A.E. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Proceedings: 5th Global Workshop on Digital Soil Mapping 2012: **Digital Soil Assessments and Beyond**, Sydney, p.10-13, 2012.

SCULL, P. et al. Predictive soil mapping: a review. **Progress in Physical Geography**, London, v.27, n.2, p.171-197, 2003.

SHI, X. **ArcSIE user's guide**. Available in: <http://www.arcsie.com/index.htm> Acessed on: Jun 4, 2013.

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

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

SILVA, S.H.G. **Cambisol (Inceptisol) solum thickness mapping based on expert knowledge with limited data from a watershed in Minas Gerais, Brazil**. Monograph. Lavras: UFLA, 2013. 41p.

SKIDMORE, A.K. et al. Use of an expert system to map forest soil from a geographical information system. **International Journal Geographical Information Science,** Wageningen, v.5, p.431-445, 1991.

VIDAL-TORRADO, P.; LEPSCH, I.F.; CASTRO, S.S. Conceitos e aplicações das relações pedologia-geomorfologia em regiões tropicais úmidas. In: VIDAL-TORRADO, P. et al. **Tópicos em Ciência do Solo**. Viçosa: Sociedade Brasileira de Ciência do Solo, 2005, v.4, p.145-192.

WALTER, C.; LAGACHERIE, P.; FOLLAIN, S. Integrating pedological knowledge into digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTS, M. (eds). **Digital Soil Mapping** - an Introductory Perspective Developments in Soil Science, New York: Elsevier, v. 31, 2007, p.281-300.

ZADEH, L.A. Fuzzy sets. **Information and Control**, v.8, p.338–353. 1965. Available in: < http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf>. Accessed on May 29, 2013.

ZHU, A.X. A similarity model for representing soil spatial information. **Geoderma**, Amsterdam, v.77, p.217–242. 1997.

ZHU, A.X., BAND, L.E. A knowledge-based approach to data integration for soil mapping. **Canadian Journal of Remote Sensing**, Kanatan, v.20, p.408– 418, 1994.

ZHU, A.X. et al. Automated soil inference under fuzzy logic. **Ecological modeling**, Amsterdam, v.90, n.2, p.123-145, 1996.

ZHU, A.X. et al. Derivation of soil properties using a soil land inference model (SoLIM). **Soil Science Society of American Journal**, Madison, v.61, n.2, p.523-533, Feb. 1997.

ZHU, A.X. et al. **SoLIM**: A New Technology For Soil Mapping Using GIS, Expert Knowledge and Fuzzy Logic. 2003. Available in: <http://solim.geography.wisc.edu/pubs/Overview2007-02-16.pdf>. Accessed on: May 29, 2013.

ZHU, A.X. et al. Soil mapping using GIS, expert knowledge, and fuzzy logic. **Soil Science Society of American Journal**, Madison, v.65, n.5, p.1463-1472, Apr./May 2001.

ZHU, Q.; LIN, H.S. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. **Pedosphere**, London, v.20, n.5, p.594-606, Sept. 2010.

ZHU, A.X.; MCKAY, D.S. Effects of spatial detail of soil information on watershed modeling. **Journal of Hydrology**, Amsterdam, v. 248, n. 4, p. 54-77, July 2001.

ZHU, A.X.; QI, F.; MOORE, A.; BURT, J.E. Prediction of soil properties using fuzzy membership values. **Geoderma**, Amsterdam, v.158, n. 3/4, p.199-206, Sept. 2010.