Anuário do Instituto de Geociências
Universidade Federal do Rio de Janeiro
https://revistas.ufrj.br/index.php/aigeo/

ISSN 0101-9759
e-ISSN 1982-3908

Mapping the Soil Frontiers with Legacy Soil Data: An Approach for Covering the Lack of Updated Reference Maps of Minas Gerais, Brazil

Mapeando as Fronteiras de Solos com Dados de Legados de Solos: uma Abordagem para Cobrir a Falta de Mapas de Referência Atualizados em Minas Gerais, Brasil

Bruno Nery Fernandes Vasconcelos[1], João Vitor Meza Bravo[2], Jorge Eduardo Ferreira Cunha[3] & Elpídio Inácio Fernandes-Filho[1]

[1]Universidade Federal de Viçosa, Centro de Ciências Agrárias, Departamento de Solos, Viçosa, MG, Brasil
[2]Universidade Federal de Uberlândia, Instituto de Geografia, Uberlândia, MG, Brasil
[3]Universidade Federal de Uberlândia, Programa de Pós-graduação em Agricultura e Informações Geoespaciais, Uberlândia, MG, Brasil
E-mails: brunonery81@gmail.com; jvmbravo@ufu.br; jorgeagronomia@yahoo.com.br; elpidio@ufv.br
Corresponding author: Bruno Nery Fernandes Vasconcelos; brunonery81@gmail.com

Abstract

Digital Soil Mapping (DSM) of large areas is a time-consuming and expensive process, where soil scientists take as many as possible observations to predict soil classes and their attributes. Sometimes, the DSM is made in geographic regions with no updated geographic information, leading the soil scientist to depend on Legacy Soil Data (LSD). However, LSD is not always available at an adequate scale or resolution, forcing soil scientists to find creative solutions. Here we present a method for mapping soil frontiers with no updated reference data. We demonstrate that by combining different LSD sources with adequate predictive environmental covariables, the results could be consistent enough for mapping the soil frontiers of a large geographic region without updated reference data. For doing that, we have adopted the full geographic extension of Minas Gerais state – Brazil – as a study area. Within its extension, Minas Gerais has heterogeneity in soil classes and soil formation processes, phenomena triggered by such a divergent universe of environmental variables. Minas Gerais has no updated soil maps, making it a relevant study case for this research. Thus, we conclude that the Digital Soil Mapping process could be enriched by using different sources of Legacy Soil Data, even when there is no updated reference data.

Keywords: Digital soil mapping; Unmapped geographic regions; Soil profile descriptions

Resumo

O mapeamento digital de solos (DSM) de grandes áreas é um processo demorado e caro, onde os cientistas do solo fazem o maior número possível de observações para prever as classes de solo e seus atributos. Às vezes, o DSM é feito em regiões geográficas onde não há informações geográficas atualizadas disponíveis, levando o cientista do solo a depender de Dados Legados de Solos (DLS). No entanto, os DLS nem sempre estão disponíveis em escala ou resolução adequada, forçando o cientista a encontrar soluções criativas. Aqui apresentamos um método para mapear as fronteiras de solos em casos onde não há dados de referência atualizados. Demonstramos que combinando diferentes fontes de DLS com covariáveis ambientais preditivas adequadas, os resultados podem ser consistentes o suficiente para mapear as fronteiras do solo de uma grande região geográfica sem dados de referência atualizados. Para isso, adotamos toda a extensão geográfica do estado de Minas Gerais – Brasil – como área de estudo. Em sua extensão, Minas Gerais apresenta heterogeneidade nas classes de solos e nos processos de formação do solo, fenômenos desencadeados por um universo tão divergente de variáveis ambientais. Além disso, Minas Gerais não possui mapas de solos atualizados, o que o torna um caso de estudo relevante para esta pesquisa. Assim, concluímos que ao utilizar diferentes fontes de Legacy Soil Data o processo de Mapeamento Digital do Solo pode ser enriquecido, mesmo quando não há dados de referência atualizados.

Palavras-chave: Mapeamento digital do solo; Regiões geográficas não mapeadas; Descrições do perfil do solo

Received: 11 January 2022; Accepted: 28 April 2022

1 Introduction

Digital Soil Mapping (DSM) is a technique that requires considerable effort from those researchers engaged in mapping soil frontiers (McBratney, Mendonça Santos & Minasny 2003; Lagacherie 2008). It depends on several factors connected to the expertise of the soil scientist, especially those of collecting and describing soil data (McBratney, Mendonça Santos & Minasny 2003; Scull et al. 2003; Arrouays, Lagacherie & Hartemink 2017). DSM is also time-consuming, making researchers expend lots of money and effort (Hendriks et al. 2019). Therefore, there is an ever-increasing need to create alternative methods for generating digital soil maps (Lagacherie 2008; Grunwald 2009; Grunwald, Thompson & Boettinger 2011; Rossiter et al. 2015; Arrouays, Lagacherie & Hartemink 2017; Hendriks et al. 2019).

Legacy Soil Data (LSD) is one of those alternative sources of collecting data for soil mapping purposes (Grunwald 2009; Kempen et al. 2009; Sulaeman et al. 2013; Heung, Hodúl & Schmidt 2017; Hendriks et al. 2019; Samuel-Rosa et al. 2020). LSD provides soil scientists with a rich context for improving the quality of its maps (Hendriks et al. 2019). The overall quality of a digital soil map usually depends on the soil scientist's expertise as well as the distribution and amount of soil samples, and the quality of the set of predictor variables, e.g., climate, geomorphology, vegetation (McBratney, Mendonça Santos & Minasny 2003; Scull et al. 2003; Odgers, McBratney & Minasny 2015). The mathematical or statistical model to convert soil sample descriptions into soil maps is also an essential variable within this context (McBratney, Mendonça Santos and Minasny, 2003; Scull et al. 2003; Lagacherie, 2008; Odgers, McBratney & Minasny 2015). Moreover, choosing that model could become an easy task if the soil scientist knows the soil classification and geographic position. Therefore LSD, combined with the soil scientist expertise, and a qualified set of predictor variables, could generate accurate soil maps because it makes available past information from experts (Sulaeman et al. 2013; Odgers, McBratney & Minasny 2015; Heung, Hodúl & Schmidt 2017; Hendriks et al. 2019; Samuel-Rosa et al. 2020).

Here we assume that the use of LSD for soil mapping purposes is possible once the Soil Science community has its standards for collecting soil data and producing its maps (McBratney, Mendonça Santos & Minasny 2003; Lagacherie 2008). These standards are given by the soil profile descriptions, standardized soil classification systems, and soil survey manuals (McBratney, Mendonça Santos & Minasny 2003; Lagacherie 2008; Hendriks et al. 2019). Hendriks et al. (2019) extensively revised cases where LSD could be used and when it could not. These authors considered the scale as the main factor influencing the use (or not) of LSD. Based on the review, Hendriks et al. (2019) indicate that the use of LSD has increased during the last few decades. Somehow it could be associated with the emergence of the Geo Big data, a context where - on the Internet - several geodatasets are free and open to use (Robinson et al. 2017).

This definition helps us argue LSD as an exciting source for DSM purposes once it fulfills gaps – time and space – on the soil maps. From the geospatial information point of view, if an old geodatabase serves some map use purpose, it has enough quality (ISO 2013). That means the geospatial information quality is more dependent on the map use purpose and the map user than the information itself (Griffin et al. 2017; Griffin, Robinson & Roth 2017). Additionally, geospatial data is rarely distributed over the national territory *in developing countries, such as Brazil*. This lack of geoinformation is one of those problems that negatively impact the country's development (Camboim, Bravo & Sluter 2015; Carneiro & Miranda 2020; Sluter et al. 2020). This is a similar scenario compared to that one described by Estes and Mooneyhan in the '90s (Estes & Mooneyhan 1994; Sluter et al. 2020).

Specifically, Brazil has a large parcel of its territory with no soil information (ten Caten et al. 2012). In this context, the Legacy Soil Data could be a valuable input for those researchers interested in representing the spatial distribution of soils and their attributes, especially within developing countries, where there is no geospatial information available (Estes & Mooneyhan 1994; Arrouays, Lagacherie & Hartemink 2017; Hendriks et al. 2019; Samuel-Rosa et al. 2020). That means the Digital Soil Mapping could benefit from the use of legacy soil data once it reduces the costs of mapping soils, enhances the accuracy of the final product, as well as could serve as a predicting variable applied to the definition of sampling method or the quality controlling process (Grunwald 2009; Odgers, McBratney & Minasny 2015; Heung, Hodúl & Schmidt 2017; Hounkpatin et al. 2018; Hendriks et al. 2019; Silva et al. 2019; Lamichhane, Kumar & Adhikari 2021).

Therefore, the research problems leading us here are defined by the following questions: how could we assist the Digital Soil Mapping of large areas by using Legacy Soil Data when there are no updated reference maps? What kind of Legacy Soil Data could benefit the soil mapping when there is no updated geographic information available? We have developed an integrative process by answering these questions, allowing digital soil maps from different sources: (1) collected in soil descriptions made by soil scientists and (2) from older/outdated soil maps.

2 Methodology and Data

2.1 Study Area

We have selected an area as a case study to achieve our goals. The area comprises the Minas Gerais State, Brazil. Figure 1 shows the geographic localization of the Minas Gerais state and its geographic context in South America.

Minas Gerais is one of the 26 Brazilian states with approximately 586.000 km², located in the southeast part of the country (Figure 1), and characterized by a pronounced heterogeneity in its natural resources, mainly related to climatic and physical/geological subjects (de Souza et al. 2015; Pereira et al. 2018).

According to the Köppen climate classification, Minas Gerais has climatic characteristics varying almost like a gradient, considering the north/south direction. The north region predominates the Aw (tropical savanna climate with dry winter season), a climate type that occupies almost 65% of the whole territory. In the central/southern portion of the state, the Cwa (humid temperate climate with dry winter and hot summer) and the Cwb (humid temperate climate with dry winter and moderately hot summer) are the climatic types with dominance (de Sá Júnior et al. 2012; Alvares et al. 2013).

Additionally, the geological characteristics of Minas Gerais comprise lands of the Archean crystalline complex, associated with the São Francisco Craton. In this region lies an important sedimentary basin of the Neoproterozoic, represented by sedimentary rocks of the Bambuí Group. The mobiles belts of Mantiqueira (or orogenic belt) is another relevant tectonic domain in Minas Gerais, including rugged terrains dominated by faulted and folded metamorphic rocks. In this region, mountainous massifs and dissected

Figure 1 Study area, state of Minas Gerais, Brazil.

plateaus developed in a humid tropical climate (Schaefer 2013). There is also a portion of Minas Gerais belonging to the Paleozoic Sedimentary Basin of Paraná, covered by Basalts of the Serra Geral Formation commonly interspersed with sandstones of Botucatu Formation, both from the Cretaceous (Schaefer 2013). Figure 2 shows the relief and the climate of Minas Gerais.

The soil diversity in Minas Gerais is a natural condition: an effect of its climatic and geological heterogeneity. This variation is expressed by the presence of soils on the 13 orders of the Brazilian Soil Classification System. This is a challenging scenario for those soil scientists interested in mapping the soil frontiers of Minas Gerais once there is no detailed and updated geographic information available.

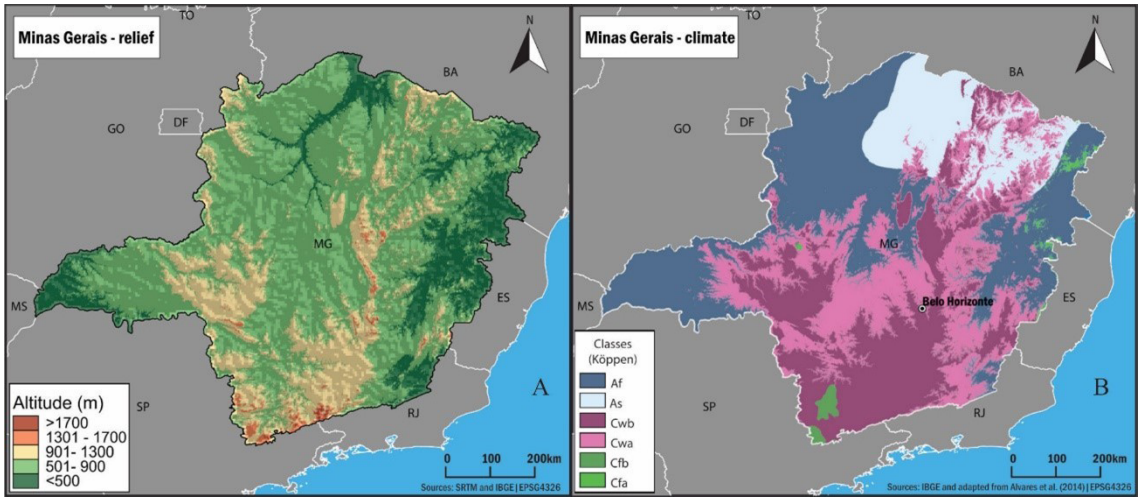

Figure 2 A: Relief of Minas Gerais; B: Climate of Minas Gerais.

Also, the rational planning for a sustainable land occupation is of great importance to our society, and the DSM could benefit the police-making process. For Brazil and, more specifically, Minas Gerais, this importance is highlighted due to high soil heterogeneity combined with increased agricultural production and a high degree of anthropogenic occupation (Souza et al. 2020).

2.2 Experimental Design, Data and Procedures

We have started by selecting the study area, as we presented before. After that, we have carried out an experiment to test the framework we have developed, first generating the maps from the two datasets and then comparing the solutions.

As we mentioned before, the experiment used LSD from different sources to create new soil maps. For designing the first map, we have used legacy data - profiles and soil samples descriptions – from several technical reports and soil surveys, made available by the Brazilian Soil Information System (EMBRAPA 2016), as well as by the Soil Database of Minas Gerais (FEAM 2016), and by the Geoprocessing Laboratory from the Universidade Federal de Viçosa. We have also compiled Legacy Soil Data from scientific reports, such as articles, master's theses, and Ph.D. dissertations. The final compilation consisted of 1856 soil profiles all over the Minas Gerais state, covering the density of 1 sample per 300 km^2 (Figure 3). We have used all the samples available for generating the map.

The second input source of Legacy Soil Data, which generated the second soil map, came from a conventional soil map produced by the team of researchers from the Universidade Federal de Viçosa (UFV 2010). By "conventional soil map," we mean a product designed under traditional methods of soil surveying, following standards, as well as made for a specific purpose. This soil map was produced at a 1:650000 scale, covering 303 mapping units (Figure 3).

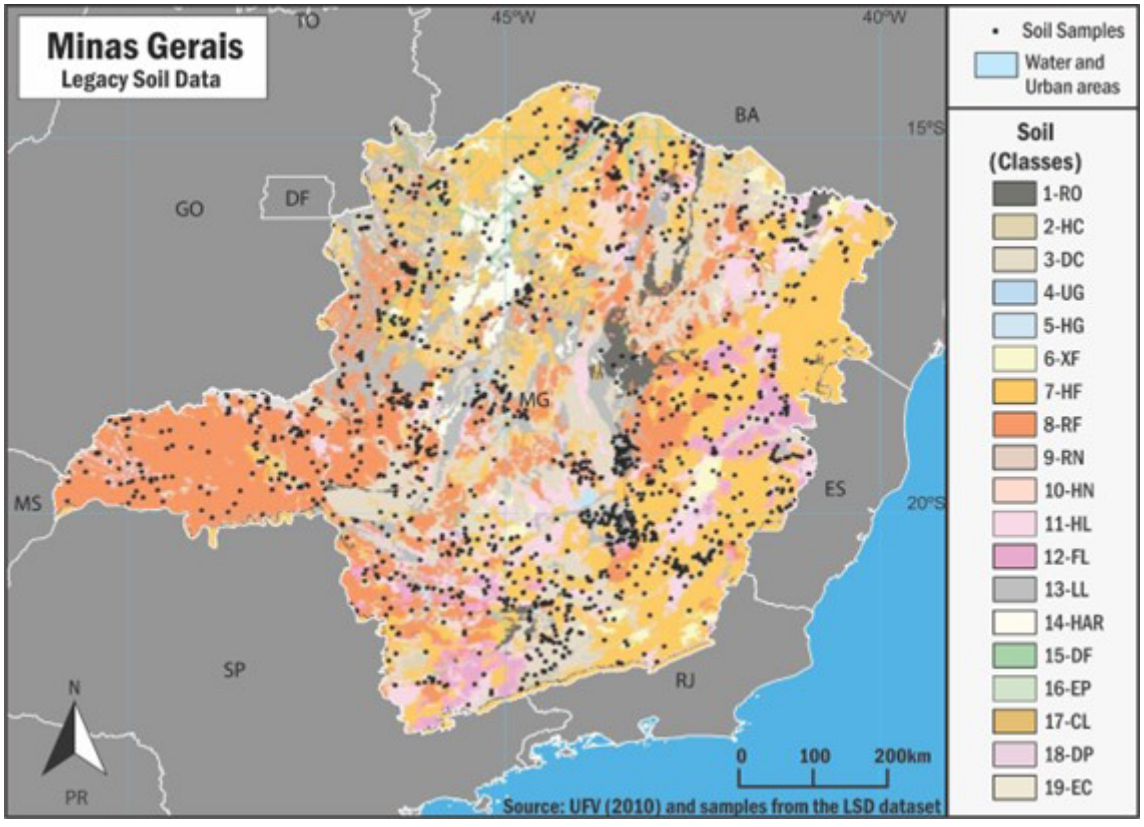

Figure 3 Sampling points derived from the profile descriptions and the conventional Soil Map produced at a scale of 1: 650,000. Note: CL: Chromic Luvisols: DC: Dystric Cambisols; DF: Dystric Fluvisols; DP: Dystric Plinthosols; EC: Eutric Cambisols; EP: Eutric Planosols; FL: Ferric Lixisols; HAC: Haplic Acrisols; HAR: Haplic Arenosols; HF: Haplic Ferralsols; HG: Haplic Gleysols; HL: Haplic Lixisols; HN: Haplic Nitisols; HC: Humic Cambisols; LL: Lithic Leptosols; REG: Regosols; RF: Rhodic Ferrasols; RN: Rhodic Nitisols; RO: Rock Outcrop; UG: Umbric Gleysols; XF: Xantic Ferralsols.

While generating the soil maps, we have also used environmental observations as predictive factors for defining soil classes. Tables 1 and 2 present the environmental factors that we called here "covariables ."The cartographic products we have used to generate the results ranged from approximately 1:100.000 (SRTM) to 1:1.000.000 (Geological maps). This scale variability is a natural condition of the method we employed after gathering information from such a diverse universe of databases.

Regarding the morphometric covariables (relief), we estimated its values using images from the Shuttle Radar Topography Mission (SRTM). We processed these data with routines developed at the R software (R-Saga package). Besides, there was a thematic map from which we delimited the state's geomorphological compartments (geomorphological map). The organism' factors came from NDVI values, calculated with MODIS (Moderate Resolution Imaging Spectroradiometer) images. The parent material data came from the geological map (CODEMIG 2014) and the geodiversity map (CPRM 2016), combined with aero geophysical data of gamma-spectrometry, gravimetry, and magnetometry (CODEMIG 2014). The WorldClim - Global Climate Data database (WorldClim 2015) gave us observations related to the climatic variables.

We have selected the main covariables as an innovative process by removing highly correlated variables (correlation above 95%). After that, we sought to identify - among the categorical variables - those with a high degree of similarity. Finally, we have used the "importance function" of the RandonForest algorithm (R package), which ranks the covariables by importance, enabling us to select a set of 16 covariables presented in Table 2. These were the covariables we have used to model the equation for predicting soil classes and their attributes.

After preparing the database for predicting the soils classes, we made the experiment analysis, as demonstrated in the following section.

2.3 Analysis

We have adopted the same method for generating the soil maps in both cases, i.e., when creating the soil map (1) from the profile and soil sample descriptions (here called dataset "a") and (2) from the outdated soil map (dataset "b"). Then, we could make a relative comparison between these two products.

Table 1 Covariables subdivided by soil formation factors.

Soil formation factors	Covariables
Relief	Geomorphological map – MDE -Real Surface Area - Convergence index - Cross sectional curvature -Flow line curvature - General curvature - Longitudinal curvature - Maximal curvature - Minimal curvature - Plan curvature - Profile curvature - Tangential curvature - Euclidean distance drainage - Diurnal anisotropic heating– Gradient - Hill index – Landforms - Standardized height - Mid slope position - Morphometric protection index - Normalized height -Slope - MRRTF (Multi-resolution ridge top flatness) - MRVBF (Multi-resolution valley bottom flatness) - Slope Height - Mass balance index -WTI (1) - Solar radiation diffuse 1 - Solar radiation diffuse 2 - Solar radiation direct 1 -Solar radiation direct 2 - Solar radiation duration 1 - Solar radiation duration 2 - Solar radiation total 1 -Solar radiation total 2 - Surface specific points - Terrain Ruggedness Index - Terrain Surface convexity - Topographic Position Index - Topographic Wetness Index (TWI) - Valley Index – Valley - Valley Depth - Vector Ruggedness Measure
Climate	Temperature (max, min, mean), Precipitation for the 12 months, and 19 Bioclimatic variables
Parent Material	Geological and Geodiversity maps (1:1.000.000), Concentration of the K, U, Th elements - Ratio between the Th / K, Th/U, U/K (Gammaspectrometry), magnetometry (full strength, vertical derivative, analytical signal), Gravimetry
Organism	NDVI (Normalized Difference Vegetation Index)

Table 2 Covariables to predict soil classes.

Soil formation factors	Covariables
Relief	Geomorphological map - MRRTF (Multi-resolution ridge top flatness) - MRVBF (Multi-resolution valley bottom flatness) - Solar radiation diffuse 1
Climate	Temperature (min) July – Bio 4 (Temperature Seasonality) – Bio 13 (Precipitation in the wettest month)– Bio 18 (Precipitation in the hottest month) – Bio 19 (Precipitation in the coldest quarter)
Parent Material	Geological maps (1:1.000.000) – Geodiversity maps (1:1.000.000) – K element concentration (gamaespectometria) – Concentration ratio of Th/K
Geographic coordinates	Latitude - Longitude

In this context, the first injunction was the geographic position of the samples on the old soil map. We adopted the same distribution of points from the dataset "a" to evaluate the results on the same geographic density and distribution basis. Then, we extracted the soil information from the dataset "b" on a Geographical Information System (QGIS) and generated the new soil map, combining LSD and the covariables until running the classifier algorithm. Therefore, the RandomForest classifier – from the RandomForest package of the R software – was the algorithm we chose for generating the soil maps. The RandomForest is a predictive model where each tree depends on the values of a random vector sampled independently with the same distribution for all trees (Breiman 2001).

Regarding the level of detail of the soil classes, the first element of the mapping units was adopted in the second categorical level of the Brazilian Soil Classification System (suborder). We choose this level because it is an intermediate level of detail compatible with the final mapping scale (1: 500.000) calculated according to the pixel size used (1.000 m).

Then we evaluated the relative accuracy of the predicted soil maps by calculating the error/confusion matrix (Lillesand et al. 2015). The confusion matrix is a method where reference data is presented in columns and the classified data in the lines; the diagonal represents the level of agreement between both maps (Lillesand et al, 2015; Congalton & Green 2019). We have also used the Kappa Index, a discrete multivariate statistic used to measure the agreement between estimated and reference data (Equation 1).

$$k = \frac{N * \sum_{i=1}^{M} M(i,i) - \sum_{i=1}^{M} SL(i) - SC(i)}{N^2 - \sum SL(i) * SC(i)} \qquad (1)$$

where
K= kappa index
N= number of pixels of verification
M= number of classes
SL= partial sum of line i
SC= partial sum of column i

The Kappa index varies between 0 and 1, where 0 (zero) represents total disagreement and one total agreement (Congalton & Green 2019) of the classes compared. In this research, we understand the level of agreement as a quality parameter that indicates a resemblance between the two soil maps we generated.

3 Results and Discussion

Tables 3 and 4 present the confusion matrices given by classifications based on (a) the soil profile data and (b) the legacy soil map, respectively. The columns represent the original data and the lines of the predicted data. Within the main diagonal, we have the number of the relative "correct" answers.

When comparing Tables 3 and 4, we see differences in the magnitude of the diagonal values. Notably, Table 4 presents higher values than Table 3, meaning the classification accuracy made using the legacy soil map has achieved better performance levels. Additionally, in both tables, it is possible to perceive the predominance of the total number of samples present in the three most expressive classes in terms of area (3-DC, 7-HF, and 8-RF). These classes are associated with the lowest error values in the classification accuracy. On the other hand, very low total sample values are observed for some classes such as 4-UG, 5-HG, 9-RN, 22-HAC, 33-REG, and 19-EC, with many of these classes showing maximum error values. This fact indicates a considerable sample imbalance, following the literature (Lillesand et al. 2015; Odgers, McBratney & Minasny 2015). Also, these last classes may be misclassified as more representative classes, as they occupy similar places in terms of relief and climate.

The analysis of the error matrices also allows us to perceive that the greater the dispersion of the number of samples along the column, the greater the difficulty in predicting a particular soil class. This dispersion is probably due to the non-specificity correlation between the soil class, the relief, and the environment, or co-occurrence with more representative classes. This is evidenced mainly in 3-DC, 7-HF, 8-RF, and 11-HL. On the other hand, classes such as 2-HC, 6-XF, and 9-RN present low "dispersion" of samples along the column, which indicates some degree of environmental specificity. Thus, 2-HC and 9-RN are associated with high altitudes or subtropical climates of southern Brazil, being this way, environmentally specific. For the 6-XF, the low occurrence is due to its specificity regarding the source material, which is not typical for most of Minas Gerais territory (Ker 1997).

The fact gives another exciting result: the classification could not detect the soil classes 16-EP, 17-CL, and 18-DP, in both cases. Further, the classes 22-HAC and 33-REG have only been seen in the classification made with the soil profile data. Table 3 also shows that all classes have received samples, even those with a smaller area, which led us to understand that the data from soil profiles was representative. In contrast, while observing Table 4 is noticeable that the classes 5-HG and 19-EC did not receive any samples. Evident, 5-HG is representative of floodplain areas and, thus, does not appear in low-density surveys. 19-EC is not a typical soil for Brazil, occurring in small islands; hence, its observation would

Table 3 Error matrix of the predicted map with the soil profile database.

Soil Classes	2-HC	3-DC	4-UG	5-HG	6-XF	7-HF	8-RF	9-RN	11-HL	12-FL	13-LL	14-HAR	15-DF	22-HAC	33-REG	Error
2-HC	13	3	1	0	0	2	2	0	0	0	2	0	0	0	1	0,46
3-DC	3	162	2	3	1	59	67	2	23	7	6	6	5	1	2	0,54
4-UG	1	4	0	1	0	10	11	0	4	0	0	3	4	0	0	1
5-HG	0	6	2	0	0	5	5	0	1	1	1	0	1	1	0	1
6-XF	0	4	0	0	7	18	12	0	1	0	0	0	0	0	0	0,83
7-HF	0	74	4	5	10	149	69	0	25	11	4	9	5	3	1	0,60
8-RF	2	77	2	2	8	62	238	4	19	6	3	3	3	1	2	0,45
9- RN	0	6	0	0	0	4	15	0	3	3	0	0	0	0	1	1
11- HL	0	39	3	0	0	42	33	3	46	11	3	1	4	1	1	0,75
12-FL	0	16	0	1	0	19	16	1	11	33	0	0	0	0	2	0,67
13-LL	2	24	0	1	0	9	17	0	3	3	1	3	0	0	3	0,98
14-HAR	0	7	0	0	0	13	8	0	2	0	1	30	0	0	2	0,52
15-DF	0	13	2	1	0	7	8	0	7	0	0	2	16	0	0	0,71
22-HAC	0	5	0	1	0	10	5	0	4	1	1	0	0	0	0	1
33-REG	1	13	0	0	0	10	11	1	4	1	5	2	0	0	1	0,98
TOTAL	22	453	16	15	26	419	517	11	153	77	27	59	38	7	16	

Note: Chromic Luvisols: CL; Dystic Cambisols: DC; Dystric Fluvisols: DF; Dystric Plinthosols: DP; Eutric Cambisols: EC; Eutric Planosols: EP; Ferric Lixisols: FL; Haplic Acrisols: HAC; Haplic Arenosols: HAR; Haplic Ferralsols: HF; Haplic Gleysols: HG; Haplic Lixisols: HL; Haplic Nitisols: HN; Humic Cambisols: HC; Lithic Leptosols: LL; Regosols: REG; Rhodic Ferrasols: RF; Rhodic Nitisols: RN; Rock Outcrop: RO; Umbric Gleysols: UG; Xanthic Ferralsols: XF.

Table 4 Predict map error matrix with the legacy soil map database.

Soil Classes	1-RO	2-HC	3-DC	4-UG	5-HG	6-XF	7-HF	8-RF	9-RN	10-HN	11-HL	12-FL	13-LL	14-HAR	15-DF	19-EC	20-WATER	Error
1-RO	12	0	8	0	0	0	3	2	0	0	3	0	12	1	0	0	0	0,71
2-HC	0	10	6	0	0	0	3	0	0	0	0	1	2	0	0	0	0	0,55
3-DC	1	0	230	0	0	0	40	27	1	0	12	9	20	0	4	0	0	0,33
4-UG	0	0	0	0	0	0	3	5	0	0	0	0	0	0	0	0	0	1
5-HG	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
6-XF	0	0	0	0	0	9	10	1	0	0	3	0	0	0	0	0	0	0,61
7-HF	0	0	25	1	0	0	375	50	0	0	14	8	10	1	5	0	1	0,23
8-RF	1	1	17	0	0	0	52	348	2	3	5	1	18	0	1	0	1	0,23
9- RN	0	0	1	0	0	0	3	1	13	0	0	0	0	0	0	0	0	0,28
10-HN	0	0	0	0	0	0	1	5	0	2	2	0	0	0	0	0	0	0,80
11- HL	0	0	15	0	0	0	40	16	0	2	98	5	0	0	0	0	0	0,44
12-FL	0	1	7	0	0	1	21	5	0	0	2	38	0	0	0	0	0	0,49
13-LL	5	1	20	0	0	0	11	26	0	0	1	0	116	1	0	0	1	0,36
14-HAR	1	0	1	0	0	0	8	3	0	0	0	0	2	5	0	0	1	0,76
15-DF	0	0	3	0	0	0	10	3	0	0	1	0	0	0	49	0	1	0,27
19-EC	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
20-WATER	0	0	0	0	0	0	0	8	0	0	1	0	0	1	1	0	10	0,52
TOTAL	20	13	333	1	0	11	580	501	16	7	142	63	180	9	60	0	15	

Note: Chromic Luvisols: CL; Dystic Cambisols: DC; Dystric Fluvisols: DF; Dystric Plinthosols: DP; Eutric Cambisols: EC; Eutric Planosols: EP; Ferric Lixisols: FL; Haplic Acrisols: HAC; Haplic Arenosols: HAR; Haplic Ferralsols: HF; Haplic Gleysols: HG; Haplic Lixisols: HL; Haplic Nitisols: HN; Humic Cambisols: HC; Lithic Leptosols: LL; Regosols: REG; Rhodic Ferrasols: RF; Rhodic Nitisols: RN; Rock Outcrop: RO; Umbric Gleysols: UG; Xanthic Ferralsols: XF.

also need a high-density survey. Therefore, considering the universe of 1856 samples released on the legacy soil map, no sample included these last classes, which allows us to infer specific incompatibility between the reference and the classified data. The Kappa index values differed in agreement with these first results (last columns Tables 3 and 4) for both predicted maps. The lowest kappa value (0.23) was observed for the map made with the soil profile data, and the highest value (0.53) was obtained for the map produced with samples derived from the legacy soil map.

Also, the results presented in the error matrices allow us to indicate the under-sampling issue of some soil classes. In this context, soil classes occupying large geographic regions (e.g., 3-DC, 7-HF, and 8-RF) receive more samples than smaller classes (e.g., 4-UG, 5-HG, 15-DF, and 19-EC). This problem can also be evidenced in the errors associated with these classes, which are always greater in the classes of smaller areas. Consequently, some soil classes are not detected during the classification, decreasing the overall product quality (Hengl et al. 2007; Kim et al. 2012; Barthold et al. 2013; Teske et al. 2015). Still, the error matrices also allowed us to detect the algorithm deficiency. Some soil class behavior (e.g., 3-DC, 7-HF, and 8-RF) evidenced that the classifier has confused - on some level - the soil classes.

In contrast to the last findings, classes such as 2-HC, 6-XF, and 9-RN presented a high level of agreement between the predicted and reference data. These previous facts allow us to understand that there are distinct class clusters.

The results we found could also be interpreted by observing Figure 4. Thus, Figure 4 allows us to visualize the level of agreement of both maps we produced. On the left side, we see the comparison made between the map created with the soil profile data, and the right side shows the comparison between the map produced with soil samples derived from the legacy soil map, which is, in this case, the reference map. It is necessary to highlight that, in both cases, the maps were produced by applying the soil classes source on the predicting model that considered the covariables described in the method.

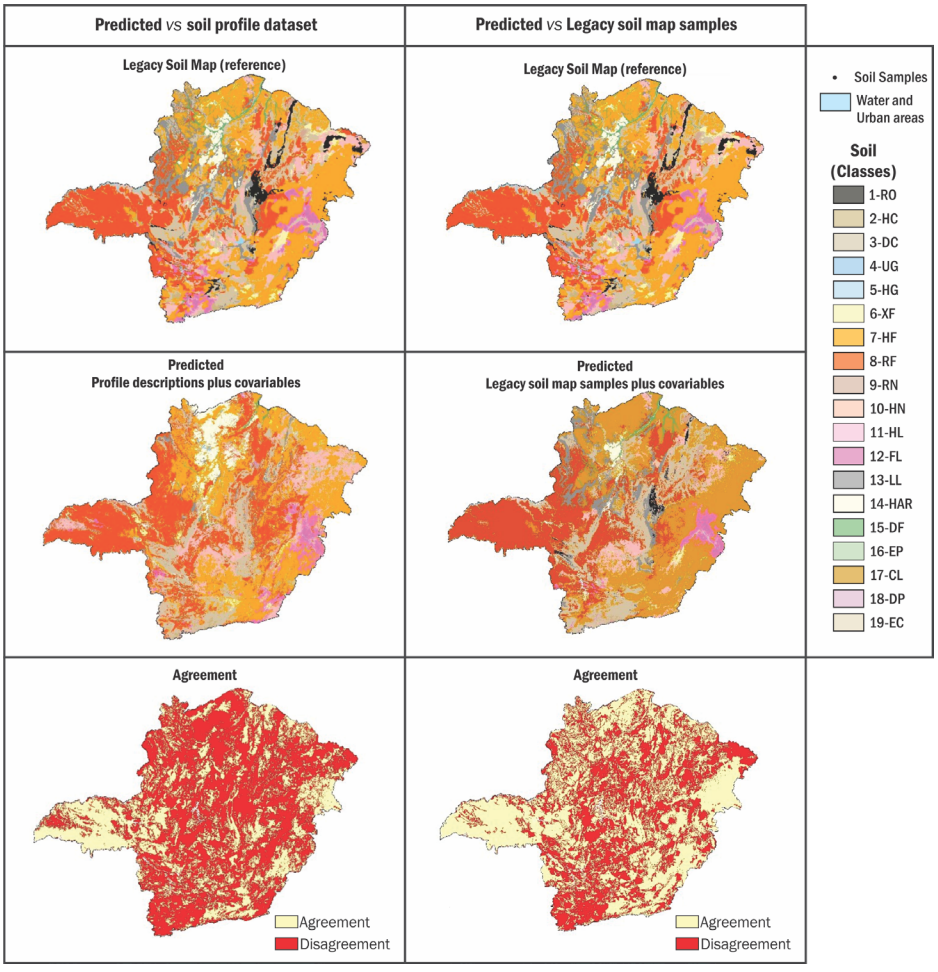

Figure 4 Comparison among the resulting maps and the reference data.

We could observe in Figure 4 that there is a partial agreement of the maps we produced compared to the reference map. That means the map created with the soil profile data has achieved 40% relative accuracy compared to the reference map. It is possible to infer that the soil class homogeneity we see in the west/southwest region of Minas Gerais plays an essential role on this correctness level. That is a geographic region where predominates soils from the class 8-RF.

In contrast, the map built with the soil samples derived from the legacy soil map has achieved 58% agreement. Figure 4 also illustrates the level of agreement between the maps we produced, whose values achieved 51%. Again, the soil homogeneity in the west/southwest region was the key factor to that correctness degree. However, we could infer that soils on the east/northeast quadrant, on that classification, were responsible for the greater level of agreement between the prediction and the reference data. These soils are in the classes 7-HF and 3-DC.

Interestingly, the soil class 1-RO was predicted on the map built with data from the legacy soil map; however, it was not detected in the expected map made with the soil profile dataset. Further, the soil class 14-HAR has had a different behavior in both predicted maps. When observing the map derived from the soil profile dataset, the 14-HAR soil class was overestimated, occupying a space where we originally had the 7-HF soil class (north). In contrast, the 14-HAR soil class was underestimated in approximately 50% of the original area of the reference map when observing the maps we produced with the legacy soil data.

Another interesting result is the presence of classes 22-PA and 33-RR in the map made with the soil profile dataset, in opposition to its non-detection on the predicted map built with the legacy soil map data. It led us to understand a geographic region where researchers engaged in the Digital Soil Mapping should densify the network of soil observations because the soil frontiers have more diversity than the reference data give it.

4 Conclusion

In general, the representativeness of the existing classes in both maps was similar. This result also attests to the potential of using legacy soil map data plus environmental covariables for generating a new soil map. On the other hand, the map resulting from the soil profile database has presented a higher degree of incompatibility when compared with the reference data than the other data source. This is understandable since much of this data was generated in larger work scales, where different soil classes were sampled over short distances.

The accuracy of detection for each class was primarily related to the representativeness of the class in the study area. Secondarily, to the specificity of the relationship between the soil and the environment in which the class occurred, soils with similar environmental occurrences are confused.

Here, the best results are those with the map built based on the legacy soil map. Mainly, the results we found demonstrate the importance of using consistent legacy data - in terms of scale or spatial resolution – with the mapping purpose (Hendriks et al. 2019). Complementary, it is needed to highlight that part of the information on the soil descriptions database was obtained from scientific works that often seek to characterize – precisely - different soils in the surrounding landscape. That is an observation similar to the one made by Carré et al. (2007), which points out that sources like soil descriptions usually present an uneven spatial distribution, generating under- and over-sampled regions, which do not always represent the predominant soil classes inside a geographic area.

Therefore, we conclude that sample distribution and density are core issues impacting the overestimating or underestimating soil class detection. Somehow, it is necessary to adopt different sampling strategies to minimize the negative impacts. An idea we encouraged to test in future research works would be using taxonomic distance tables, which could decrease the classification confusion. Then, greater weights would be used for errors associated with more taxonomically distant classes.

Finally, the results demonstrate that the combination of different Legacy Soil Data sources, and predictive environmental covariables for soil mapping, is a valuable input/method for mapping the soil frontiers in geographic regions with no updated soil data.

5 Acknowledgments

To CNPq for granting a scholarship (PhD) and to the postgraduate program in soils and plant nutrition at UFV - SNP-DPS-UFV

6 References

Alvares, C.A., Stape, J.L., Sentelhas, P.C., Gonçalves, J.L.M. & Sparovek, G. 2013, 'Köppen's climate classification map for Brazil', *Meteorologische Zeitschrift*, vol. 22, no. 6, pp. 711-28, DOI:10.1127/0941-2948/2013/0507.

Arrouays, D., Lagacherie, P. & Hartemink, A. E. 2017, 'Digital soil mapping across the globe', *Geoderma Regional*, vol. 9, pp. 1-4, DOI:10.1016/j.geodrs.2017.03.002.

Barthold, F. K., Wiesmeier, M., Breuer, L., Frede, H.G., Wu, J. & Blank, F.B. 2013, 'Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia', *Journal of Arid Environments*, vol. 88, pp. 194–205.

Breiman, L. 2001, 'Random Forests', *Machine learning*, vol. 45.1, pp. 5–32.

Camboim, S.P., Bravo, J.V.M. & Sluter, C.R. 2015, 'An investigation into the completeness of, and the updates to, OpenStreetMap data in a heterogeneous area in Brazil', *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, pp. 1366–88, DOI:10.3390/ijgi4031366.

Carneiro, A.F.T. & Miranda, C.R. 2020, 'Evolution and trends in research on land administration and cadastre', *Revista Brasileira de Cartografia*, vol. 72, pp. 880–97, DOI:10.14393/rbcv72nespecial50anos-56586.

Carré, F., Mcbratney, A. B. & Minasny, B. 2007, 'Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping', *Geoderma*, vol. 141, no. 1-2, pp. 1–14, DOI:10.1016/j.geoderma.2007.01.018.

CODEMIG – Companhia Desenvolvimento Econômico de Minas Gerais 2014, *Mapa geológico*. Retrieved May 26, 2016. <http://www.portaldageologia.com.br>.

Congalton, R.G. & Green, K. 2019, *Assessing the Accuracy of Remotely Sensed Data*, *Assessing the Accuracy of Remotely Sensed Data*, 3rd edn, CRC Press, Boca Raton.

CPRM – Companhia de Pesquisa de Recursos Minerais 2016. *Mapa Geodversidade*. Retrieved May 26, 2019. <http://www.cprm.gov.br/geobank>.

de Sá Júnior, A., Carvalho, L.G., Silva, F.F. & Alves, M.C. (2012) 'Application of the Köppen classification for climatic zoning in the state of Minas Gerais, Brazil', *Theoretical and Applied Climatology*, vol. 108, no. 1, pp. 1-7, DOI:10.1007/s00704-011-0507-8.

de Souza, J.J.L.L., Abrahão, W.A., da Silva, J., da Costa, L.M. & de Oliveira, T.S. 2015, 'Geochemistry and spatial variability of metal(loid) concentrations in soils of the state of Minas Gerais, Brazil', *Science of the Total Environment*, vol. 505, pp. 338–49, DOI:10.1016/j.scitotenv.2014.09.098.

EMPRABA – Empresa Brasileira de Pesquisa Agropecuária 2016. *Sistema de Informação de Solos Brasileiros*. <https://www.sisolos.cnptia.embrapa.br/>.

Estes, J.E. & Mooneyhan, D.W. 1994, 'Of maps and myths', *Photogrammetric Engineering & Remote Sensing*, vol. 60, no. 5, pp. 517–24.

FEAM – Fundação Estadual do Meio Ambiente 2016. *Banco de Solos de Minas Gerais*. <http://www.feam.br/>.

Griffin, A.L., Griffin, A.L., White, T.M., Fish, C.S., Tomio, B., Huang, H., Sluter, C.R., Bravo, J.V., Fabrikant, S.I., Bleisch, S., Yamada, M.M. & Picanço, P.L. 2017, 'Designing across map use contexts: a research agenda', *International Journal of Cartography*. Taylor & Francis, vol. 3, sup. 1, pp. 90–114, DOI:10.1080/23729333.2017.1315988.

Griffin, A.L., Robinson, A.C. & Roth, R.E. 2017, 'Envisioning the future of cartographic research', *International Journal of Cartography*, vol. 3, sup. 1, pp. 1–8, DOI:10.1080/23729333.2017.1316466.

Grunwald, S. 2009, 'Multi-criteria characterization of recent digital soil mapping and modeling approaches', *Geoderma*, vol. 152, no. 3-4, pp. 195-207, DOI:10.1016/j.geoderma.2009.06.003.

Grunwald, S., Thompson, J.A. & Boettinger, J.L. 2011, 'Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues', *Soil Science Society of America Journal*, vol. 75, no. 4, pp. 1201-13, DOI:10.2136/sssaj2011.0025.

Hendriks, C.M.J., Stoorvogel, JJ., Lutz, F. & Claessens, L.2019, 'When can legacy soil data be used, and when should new data be collected instead?', *Geoderma*, vol. 348, no. 2, pp. 181–8, DOI:10.1016/j.geoderma.2019.04.026.

Hengl, T., Toomanian, N., Reuter, H.I. & Malakouti, M.J. 2007, 'Methods to interpolate soil categorical variables from profile observations: Lessons from Iran', *Geoderma*, vol. 140, no. 4, pp. 417–27.

Heung, B., Hodúl, M. & Schmidt, M.G. 2017, 'Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes', *Geoderma*, vol. 290, pp. 51-68, DOI:10.1016/j.geoderma.2016.12.001.

Hounkpatin, K.O.L., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten, T., Ameleung, W. & Welp, G. 2018, 'Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso)', *Scientific Reports*, vol. 8, no. 1, e9959, DOI:10.1038/s41598-018-28244-w.

ISO 2013, *Geographic information — Data quality ISO/FDIS 19157*, *Iso/Tc 211*.

Kempen, B., Brus, D.J., Heuvelink, G.B.M. & Stoorvogel, J.J. 2009, 'Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach', *Geoderma*, vol. 151, no. 3-4, pp. 311-26, DOI:10.1016/j.geoderma.2009.04.023.

Ker, J.C. 1997, 'Latossolos Do Brasil: Uma Revisão', *Geonomos*, vol. 5, no. 1, pp. 17–40, DOI:10.18285/geonomos.v5i1.187.

Kim, J., Grunwald, S., Rivero, R.G. & Robbins, R. 2012, 'Multi-scale Modeling of Soil Series Using Remote Sensing in a Wetland Ecosystem', *Soil Science Society of America Journal*, vol. 76, no. 6, pp. 2327-41, DOI:10.2136/sssaj2012.0043.

Lagacherie, P. 2008, 'Digital soil mapping: A state of the art', in A.E. Hartemink, A. McBratney & M. Mendonça-Santos, M. (eds), *Digital Soil Mapping with Limited Data*, Springer, Dordrecht, pp. 3-14.

Lamichhane, S., Kumar, L. & Adhikari, K. 2021, 'Updating the national soil map of Nepal through digital soil mapping', *Geoderma*, vol. 139, e108951, DOI:10.1016/j.geoderma.2021.115041.

Lillesand, T.M., Kiefer, R.W. & Chipman, J.W. 2004, *Remote sensing and image interpretation*, New York.

McBratney, A.B., Mendonça Santos, M.L. & Minasny, B. 2003, 'On digital soil mapping', *Geoderma*, vol. 117, no. 1-2, pp.3-52, DOI:10.1016/S0016-7061(03)00223-4.

Odgers, N.P., McBratney, A.B. & Minasny, B. 2015, 'Digital soil property mapping and uncertainty estimation using soil class probability rasters', *Geoderma*, vol. 237-238, pp. 190-8, DOI:10.1016/j.geoderma.2014.09.009.

Pereira, G., Cardozo, F.S., Negreiros, A.B., Zanin, G.D., Costa, J.C., Lima, T.E.R., Rufino, P.R. & Ramos, R.C. 2018, 'Análise Da Variabilidade Da Precipitação Para O Estado De Minas Gerais (1981-2017)', *Revista Brasileira de Climatologia*, vol. 1, pp. 213–29, DOI:10.5380/abclima.v1i0.61028.

Robinson, A.C., Demsar, U. , Moore, A.B., Buckley, A., Jiang, B., Field, K., Kraak, M.-J., Camboim, S.P. & Sluter, C.R. 2017,

'Geospatial big data and cartography: research challenges and opportunities for making maps that matter', *International Journal of Cartography*, vol. 3, sup. 1, pp. 32–60, DOI:10.1080/23729333.2016.1278151.

Rossiter, D.G., Liu, J., Carlisle, S. & Zhu, A.-X. 2015, 'Can citizen science assist digital soil mapping?', *Geoderma*, vol. 259-260, pp. 71-80, DOI:10.1016/j.geoderma.2015.05.006.

Samuel-Rosa, A., Dalmolin, R.S.D., Moura-Bueno, J.M., Teixeira, W.G. & Alba, J.M.F. 2020, 'Open legacy soil survey data in Brazil: Geospatial data quality and how to improve it', *Scientia Agricola*, vol. 77, no. 1, e20170430, DOI:10.1590/1678-992x-2017-0430.

Schaefer, C.E.G.R. 2013, 'Bases físicas da paisagem brasileira: estrutura geológica, relevo e solos. Tópicos em ciência do solo', *Sociedade Brasileira de Ciência do Solo*, vol. 8, no. 1, pp. 221–78.

Scull, P. 2003, 'Predictive soil mapping: A review', *Progress in Physical Geography*, vol. 27, no. 2, pp. 171-97, DOI:10.1191/0309133303pp366ra.

Silva, E.B, Giasson, E., Dotto, A.C., ten Caten, A.,Demattê, J.A.M., Bacic, I.L.Z. & Veiga, M. 2019, 'A regional legacy soil dataset for prediction of sand and clay content with VIS-NIR-SWIR, in southern Brazil', *Revista Brasileira de Ciencia do Solo*, vol. 43, e0180174, DOI:10.1590/18069657rbcs20180174.

Sluter, C.R., Carneiro, A.F.T., Iescheck, A.L., Pontes, D.R. & Gediel, J.A.P. 2020, 'Cartografia e Direito na Formação Territorial e na Configuração da Propriedade no Brasil', *Revista Brasileira de Cartografia*. EDUFU - Editora da Universidade Federal de Uberlandia, vol. 72, pp. 916–39, DOI:10.14393/rbcv72nespecial50anos-56599.

Souza, C.M., Shimbo, J., Rosa, M.R., Parente, L.L., Alencar, A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., Ferreira, L.G., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Ferreira Neto, L.C., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., Azevedo, T. 2020, 'Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine', *Remote Sensing*, vol. 12, no. 17, e2735, DOI:10.3390/RS12172735.

Sulaeman, Y., Minasny, B., McBratney, A.B., Sarwani, M. & Sutandi, A. 2013, 'Harmonizing legacy soil data for digital soil mapping in Indonesia', *Geoderma*, vol. 192, pp. 77-85, DOI:10.1016/j.geoderma.2012.08.005.

ten Caten, A., Dalmolin, R.S.D., Mendonça-Santos, M.L. & Giasson, E. 2012, 'Mapeamento digital de classes de solos: Características da abordagem brasileira', *Ciencia Rural*, vol. 42, no. 11, pp. 1989-97, DOI:10.1590/S0103-84782012001100013.

Teske, R., Giasson, E. & Bagatini, T. 2015, 'Comparação De Esquemas De Amostragem Para Treinamento De Modelos Preditores No Mapeamento Digital De Classes De Solos', *Revista Brasileira de Ciência do Solo*, vol. 39, no. 1, pp. 14-20.

UFV – Universidade Federal de Viçosa; Fundação Centro Tecnológico de Minas Gerais (CETEC-MG); Universidade Federal de Lavras (UFLA); Fundação Estadual do Meio Ambiente (FEAM) 2010. *Mapa de Solos Do Estado de Minas Gerais: legenda expandida*. Fundação Estadual do Meio Ambiente.

WORLDCLIM 2015. *Global Climate Data*. <http://www.worldclim.com.br>.

Author contributions

Bruno Nery Fernandes Vasconcelos: conceptualization; formal analysis; methodology; validation; writing-original draft; writing. **João Vitor Meza Bravo:** methodology; validation; writing; visualization; editing. **Jorge Eduardo Ferreira Cunha:** writing; validation; review. **Elpídio Inácio Fernandes-Filho:** conceptualization; methodology; supervision.

Conflict of interest

The authors declare no potential conflict of interest.

Data availability statement

All data included in this study are publicly available in the literature.

Funding information

CNPq Scholarship – Bruno Nery Fernandes Vasconcelos.

How to cite:

Vasconcelos, B.N.F., Bravo, J.V.M., Cunha, J.E.F. & Fernandes-Filho, E.I. 2023, 'Mapping the Soil Frontiers with Legacy Soil Data: An Approach for Covering the Lack of Updated Reference Maps of Minas Gerais, Brazil', *Anuário do Instituto de Geociências,* 46:49327. https://doi.org/10.11137/1982-3908_2023_46_49327