II Pedometrics
Nov, 24th to 27th, 2021
Brazil

Soil data curation with the help of an expert system for soil classification

VAZ, Glauber J.[1]; SILVA JÚNIOR, Adalberto F.[2]; SILVA NETO, Luís de França[3]
[1] Embrapa Informática Agropecuária, glauber.vaz@embrapa.br; [2] Universidade Federal Rural de Pernambuco, adalbertofrancisco75@gmail.com; [3] Embrapa Solos, luis.franca@embrapa.br

Thematic Session: Pedometrics: innovation in tropics

Abstract
We used an expert system for soil classification to curate soil data and improve its quality. The system is the first to accurately classify soil profiles to the fourth level of the current version of the Brazilian Soil Classification System (SiBCS). We analyzed 94 soil profiles using the expert system, which guided the necessary changes on soil data to make it consistent with the corresponding classifications. About 45% of soil profiles did not require data treatment, and most changes were related to horizon symbols. Even after data treatment, changes in classification were necessary for almost 40% of the profiles on at least one categorical level. Therefore, using an expert system for soil classification can help identify inconsistencies in data and classifications of soil profiles, in addition to guiding the necessary changes. It can also help improve the SiBCS.
Keywords: data quality; soil profiles; digital tool.

Introduction

Increasing the quantity and quality of soil data and information is essential for improving soil resource governance. It is one pillar of action for the Global Soil Partnership (GSP) (FAO, 2021), which aims to improve soil governance to guarantee healthy and productive soils as well as supporting the provision of essential ecosystem services. The National Soil Program of Brazil (Pronasolos) (POLIDORO et al., 2016) was proposed to provide richer information on Brazilian soils for decision making. The long-term program has five main lines of action, one of which deals with database and soil information.

Soil classification is an essential component of soil science. The Brazilian Soil Classification System—in Portuguese, *Sistema Brasileiro de Classificação de Solos* (SiBCS)—is the official taxonomic system for soil classification in the country. It is structured in the form of a taxonomic key up to the fourth categorical level. It also contains recommendations of qualifiers for the fifth level and suggested properties for the sixth categorical level (DOS SANTOS et al., 2018).

The correct classification relies on the consistency and completeness of soil data, which involves dozens of soil attributes. Vaz et al. (2019) developed an expert system for automatic soil classification and analyzed data from a widely used soil database, comparing the results of the system with the classifications registered in the database. They showed the need for greater data curation of available databases under the supervision of soil scientists and presented the system as a powerful tool to assist with this activity. However, their analysis was limited to the first level of SiBCS, and did not include data curation. In the present study, we used the same

expert system as Vaz et al. (2019) to examine soil data to the first four levels of SiBCS, as well as took steps to curate the data and improve its quality.

Methodology

The expert system we used to analyze the soil profiles is based on the rules of SiBCS for its first four categorical levels. The classification provided by the system only considers the current version of SiBCS (DOS SANTOS et al., 2018).

We formed a team of soil and computer scientists to analyze the data with the help of the expert system. Once completely validated, the software should provide correct classifications in all cases since the data provided is correct and complete. In some situations, the software can generate wrong classifications due to the considerable complexity and the number of possible classes in the system. When this occurs, the software is corrected and starts to produce the expected result. Therefore, all the automatic classifications generated by the system in the present study were correct and verified by soil scientists.

We analyzed 94 soil profiles from the states of Pernambuco and Rio Grande do Norte in Brazil. These samples were collected during the GeoTab Project, which aimed to organize soil data from the Brazilian coastal tablelands and update the classifications of the profiles. These are available in '.doc' files, meaning that the data needed to be processed in order to be generated in the format required by the expert system. We did it using an app called SmartSolos.

After obtaining the automatic classification for a given soil profile, we compared it with the recorded one. When the classes were different, we analyzed the data to check its consistency and the rules of the software in order to verify its correctness. The source of such differences could be errors in software, soil attribute data, or classification. For each case, we made the necessary changes.

Results and discussion

Table 1 shows the number of profiles analyzed for each first categorical level (order) of SiBCS. The 'Classification' columns provide the number of soil profiles from each order that were classified by the expert system according to the records made previously by soil scientists. The 'Data' columns indicate whether data treatment was required in order to obtain a correct classification.

The 'Ln' columns give the number of profiles whose records were correctly classified to the n^{th} level. For example, the classification of nine out of 25 *argissolos* was consistent with verified records to the fourth level, while 15 *argissolos* had correct classifications to the third level, but not the fourth. Finally, one profile that was actually an *argissolo* had been labeled with entirely different classes.

Table 1: The classifications and the consistency of data for the analyzed soil profiles.

Order	# Profiles	Classification					Data			
		L0	L1	L2	L3	L4	OK	Horizon	Addition	Update
Argissolos	25	1	0	0	15	9	12	9	1	3
Cambissolos	11	0	0	1	2	8	4	5	2	1
Chernossolos	2	0	0	0	0	2	1	1	1	0
Espodossolos	3	0	0	0	0	3	3	0	0	0
Gleissolos	8	0	0	3	0	5	6	0	2	0
Latossolos	11	0	0	1	2	8	8	2	1	0
Luvissolos	5	2	0	0	1	2	2	3	0	0
Neossolos	13	1	0	3	0	9	4	9	7	0
Nitossolos	1	0	0	0	0	1	0	0	0	1
Organossolos	1	0	0	0	0	1	0	1	0	0
Planossolos	7	2	0	0	1	4	1	6	0	0
Plintossolos	4	0	0	0	0	4	1	3	0	0
Vertissolos	2	0	0	1	0	1	1	1	0	0
Unknown	1	1	0	0	0	0	0	1	0	0
Total	94	7	0	9	21	57	43	41	14	5

The ‘Data’ column group indicates the changes, if any, required for each classification:

- OK: data were consistent; therefore, no change was made.
- Horizon: changes in the horizon symbols.
- Addition: additional data were needed.
- Update: updates in some attributes.

In order to arrive at a consistent classification, profiles occasionally required changes to horizon symbols, attribute updates, or additional data. The sum of numbers in the ‘Data’ columns is not necessarily equal to the number of profiles examined from the corresponding order, as is the case of the ‘Classification’ columns. This might occur, for example, when a single profile requires changes in both horizon symbol and attribute update.

After data treatment, the system classified 60.6% (57/94) of all profiles in a manner consistent with the records at all four levels. Meanwhile, 22.3% (21/94) of profiles were consistent with the third level, with errors only arising in the fourth. In most of these cases, the registered class at the fourth level is no longer valid. As such, these errors were largely caused by incompatibilities across SiBCS versions, and the records had not yet been updated. In 9.6% (9/94) of profiles, only the first and second levels were correct. In 7.5% (7/94) of profiles, the classification was completely different from the original. Therefore, some change in classification was necessary for almost 40% (37/94) of the profiles.

To obtain a correct classification, data must be correct and complete. No data treatment was required for 45.7% (43/94) of profiles. Of the profiles that did require changes, most needed only the adjustment of symbol horizons, which can be quickly done by a specialist. Updating obsolete symbols to the current standard and adding a missing suffix were the most common changes. In 14.9% (14/94) of the profiles, it was necessary to add data that a specialist would be able to distinguish but were not explicitly registered. In some cases, it was necessary to replicate the dry color in other horizons or to add an attribute indicating, for example, cohesive qualifier, fluvic qualifier, or alterable primary materials. Data not related to horizon symbols only had to be updated in 5.3% (5/94) of cases, generally for a single attribute. Thus, incorrect attribute values were corrected after analysis by a domain specialist who identified the inconsistencies in the data. In many cases, they were only recognized because the classification obtained by the system was not equal to the one recorded—furthermore, the results from the expert system provided indications of the necessary changes.

It is important to note that one profile was classified by the system as “unknown” for the first level. The current version of SiBCS considers the predominance (>50%) of activity clay in the B horizon to classify *luvissolos* and *argissolos*. However, in the profile classified as “unknown”, 50% of the B horizon had low-activity clay and 50% high-activity clay. Therefore, it is not classified either as a *luvissolo* or as an *argissolo*. This demonstrates another benefit of the expert system, namely its ability to validate SiBCS rules using software.

Conclusions

Analyzing soil profiles with an automatic soil classification tool makes it easier to identify errors in data or classification of soil profiles and allows more reliable data curation. Additionally, the system can identify areas for improvement in the SiBCS.

References

DOS SANTOS, H. G. et al. **Sistema brasileiro de classificação de solos**. Brasília, DF: Embrapa, 2018.

FAO. **Global soil partnership.** Disponível em: <www.fao.org/global-soil-partnership>. Acesso em: 23 set. 2021.

POLIDORO, J. C. et al. **Programa Nacional de Solos do Brasil (PronaSolos)**. Rio de Janeiro: Embrapa Solos, 2016.

VAZ, G. J. et al. Uma API para a classificação de solos do Brasil. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 12., 2019, Indaiatuba. **Anais...** Ponta Grossa: SBIAGRO, 2019. p. 63-72.