DOI: http://dx.doi.org/10.1590/1678-992X-2017-0430

ISSN 1678-992X

SCIENTIA AGRICOLA

Research Article

Open legacy soil survey data in Brazil: geospatial data quality and how to improve it

Alessandro Samuel-Rosa[1]*, Ricardo Simão Diniz Dalmolin[2], Jean Michel Moura-Bueno[2], Wenceslau Geraldes Teixeira[3], José Maria Filippini Alba[4]

[1]Universidade Tecnológica Federal do Paraná, Prolongamento da Rua Cerejeira, s/n – 85892-000 – Santa Helena, PR – Brasil.
[2]Universidade Federal de Santa Maria – Depto. de Solos, Av. Roraima, 1000 – 97105-900 – Santa Maria, RS – Brasil.
[3]Embrapa Solos, R. Jardim Botânico, 1024 – 22460-000 – Rio de Janeiro, RJ – Brasil.
[4]Embrapa Clima Temperado, Rod. BR-392, km 78 – 96010-971 – Pelotas, RS – Brasil.
*Corresponding author <alessandrorosa@utfpr.edu.br>

Edited by: Silvia del Carmen Imhoff

Received December 21, 2017
Accepted July 28, 2018

ABSTRACT: Spatial soil data applications require sound geospatial data including coordinates and a coordinate reference system. However, when it comes to legacy soil data we frequently find them to be missing or incorrect. This paper assesses the quality of the geospatial data of legacy soil observations in Brazil, and evaluates geospatial data sources (survey reports, maps, spatial data infrastructures, web mapping services) and expert knowledge as a means to fix inconsistencies. The analyses included several consistency checks performed on 6,195 observations from the Brazilian Soil Information System. The positional accuracy of geospatial data sources was estimated so as to obtain an indication of the quality for fixing inconsistencies. The coordinates of 20 soil observations, estimated using the web mapping service, were validated with the true coordinates measured in the field. Overall, inconsistencies of different types and magnitudes were found in half of the observations, causing mild to severe misplacements. The involuntary substitution of symbols and numeric characters with similar appearance when recording geospatial data was the most common typing mistake. Among the geospatial data sources, the web mapping service was the most useful, due to operational advantages and lower positional error (~6 m). However, the quality of the description of the observation location controls the accuracy of estimated coordinates. Thus, the error of coordinates estimated using the web mapping service ranged between 30 and 1000 m. This is equivalent to coordinates measured from arc-seconds to arc-minutes, respectively. Under this scenario, the feedback from soil survey experts is crucial to improving the quality of geospatial data.

Keywords: Free Brazilian Repository for Open Soil Data, PronaSolos, Pedometrics, soil data recovery, digital soil mapping

Introduction

A large quantity of soil data has already been produced in Brazil as part of soil surveys and research projects on the various aspects of soil science. A considerable proportion of these data has been used only once and faces the risk of never being used ever again. A number of initiatives have been undertaken to compile some of the existing soil data – for example, Chagas et al. (2004), Cooper et al. (2005), Ottoni et al. (2014) and Simões et al. (2015). Due to a variety of reasons – choice of methods and technologies, specificity of goals and policies, and limited funding – these initiatives have not presented a permanent solution to the problem of safeguarding and promoting the reusability of all kinds of soil data.

The under use of soil data represents an inefficient use of the scarce investments made in soil surveys, and holds back the advancement of soil knowledge. As such, Brazilian soil scientists have recently engaged in the design of a free soil data repository that uses community build standards, follows open data policies, and presses for ease of access, maintenance and use. The Free Brazilian Repository for Open Soil Data (febr, www.ufsm.br/febr) is a centralized repository that stores accessible open soil data in a standardized and harmonized format for various applications. These include the new Brazilian national soil survey program (PronaSolos), which aims at producing up-to-date, detailed spatial soil information using digital soil mapping (DSM) (Polidoro et al., 2016).

In general, DSM requires soil observations to be accompanied with geospatial data, the spatial coordinates and the coordinate reference system (CRS). These are needed for sampling environmental covariates, e.g. satellite images, to compute their correlation with soil properties. They are also needed for computing the spatial autocorrelation of soil properties. Once these (auto)correlation structures are known, they can be used to make soil predictions. However, as regards legacy soil data, soil data produced decades ago left for future generations to use, we frequently found the geospatial data to be missing or significantly erroroneous (Arrouays et al., 2017; Batjes et al., 2017). Large positional errors can have a negative impact on soil predictions (Samsonova et al., 2018). An alternative is to use existing polygon soil maps, e.g. via spatial disaggregation (Odgers et al., 2014). However, the cartographic scale, age, and purpose of these maps control the type, amount, and accuracy of the information that can be derived. Thus, following the example of Cooper et al. (2005), the febr also targets improving the geospatial data quality of published legacy soil data. Such an effort should ultimately increase the potential of soil data to be reused.

This paper aims at assessing the quality of the geospatial data of legacy soil observations used to feed the febr. It also aims at evaluating free geospatial data sources and expert knowledge as a means of fixing inconsistencies in geospatial data.

Materials and Methods

Open legacy soil data

The largest source of compiled open legacy soil data used to feed the febr is the Brazilian Soil Information System (BDSolos, www.bdsolos.cnptia.embrapa.br), designed to store accessible soil data produced as part of soil surveys and research projects undertaken by Embrapa and partner institutions since the 1960s. In Dec 2016, we downloaded all open legacy soil survey data in BDSolos: 9,119 soil observations from 223 soil surveys (Appendix 1). Since we were initially involved in studying the soil iron content across Brazil, only observations containing data on this soil property were retained (Figure 1). A total of 6,195 observations were selected, most of them from the states of Minas Gerais (883), Rio de Janeiro (680), Amazonas (453), Bahia (515), Pará (398), and Paraná (316).

Geospatial data quality

The quality of the geospatial data of the open legacy soil data was assessed with respect to the spatial coordinates and CRS, including the hemisphere (for geographic and projected CRS), zone (for projected CRS) and geodetic datum (Figure 1). The assessment implemented using the R software consisted of verifying, for each observation, if the geospatial data was available. When available, we checked whether the recorded values were within the expected limits – the boundary box of Brazil, country and state borders – and in accordance with existing standards – official soil description guidelines, and Brazilian cartographic legislation. Where this was not the case, the observation was flagged so as to indicate the occurrence of an inconsistency: missing or incorrect. All flagged observations were evaluated as to the type and source of the geospatial data inconsistency. Once we had an understanding of the inconsistency we tried to fix it by using a series of geospatial data sources.

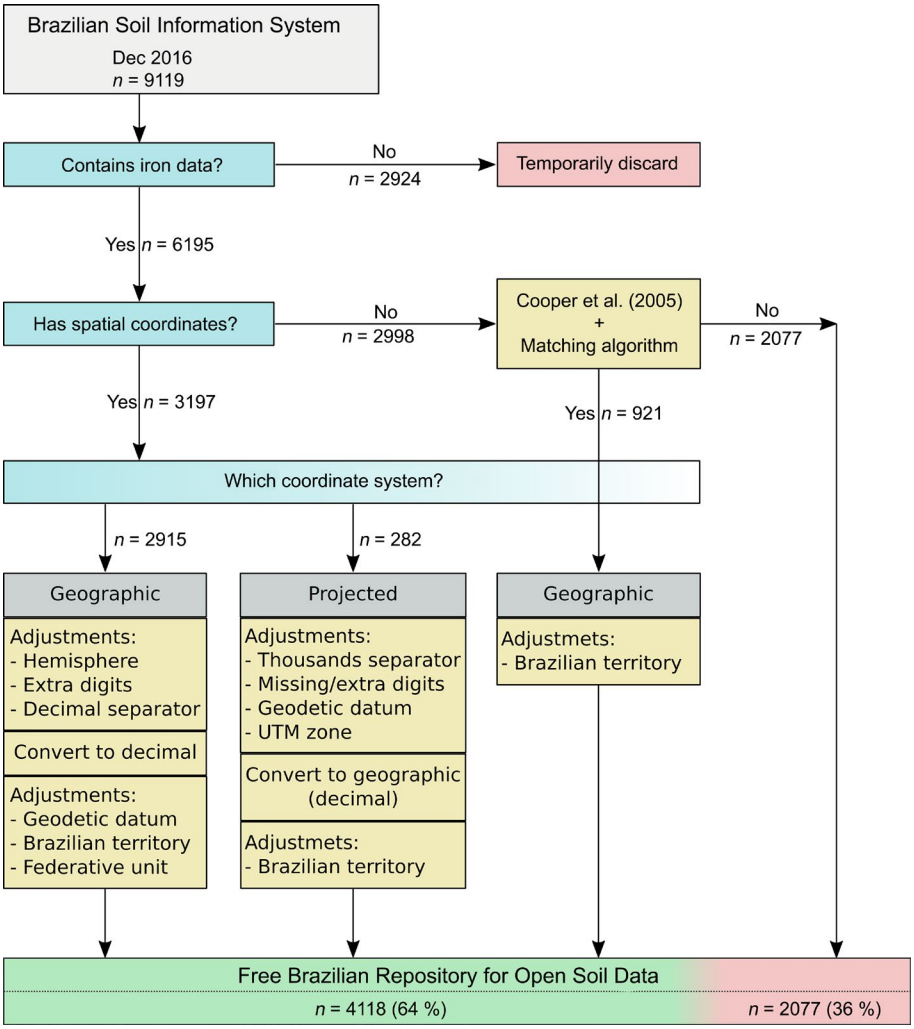

Figure 1 – Flowchart of the sequence of soil data quality check and data processing steps adopted in our study.

Fixing geospatial data inconsistencies

The primary source of geospatial data used to fix inconsistencies was the original survey reports from which soil data had been compiled. Official data of the administrative boundaries of states and municipalities of the Brazilian National Spatial Data Infrastructure (INDE, www.metadados.inde.gov.br) were used to support spatial visualization and understand geospatial survey descriptions. Where geospatial descriptions found in older survey reports were poorly detailed or difficult to understand we contacted soil survey experts with experience in the mapped region to ask for their help in locating the profile. If these resources were insufficient for fixing the inconsistency, then we used two alternative sources of geospatial data.

The first alternative source of geospatial data was the dataset compiled by Cooper et al. (2005). This dataset contains the coordinates of 5,086 observations recovered by reading from survey reports or visually estimating from soil maps. It also contains all necessary information for identifying observations and their survey report of origin. Thus, we devised a computer program in the R language to automatically find the observations in the dataset compiled by Cooper et al. (2005) matching the observations downloaded from BDSolos from which spatial coordinates were still missing (Figure 1). Among the matching criteria were survey type, publication year, volume, number, responsible institution, soil classification and profile identification number. When a match was found, we visually checked the soil classification and profile identification number in both datasets. If the match was correct, then the coordinates of the observation in the dataset compiled by Cooper et al. (2005) were attributed to the respective matching observation downloaded from BDSolos.

The second alternative source of geospatial data was a web mapping service (www.google.com.br/maps). This web mapping service provides easy and free access to fine resolution (about 1-2 m) satellite imagery, as well as road maps. It also offers a fast search engine that can be used to find cities, roads and water bodies, as well as tools for calculating road distances. Information on these easily identifiable geographical markers was present in most soil survey reports in the form of a textual description of the location of soil observations. By using this information to feed the search engine, we estimated the most likely coordinates of many of the remaining observations with incorrect geospatial data (Appendix 2). Soil survey experts were asked to evaluate the likelihood of the estimated coordinates.

Accuracy of geospatial data sources

The inconsistencies of the geospatial data of legacy soil observations were fixed using primary (survey reports) and alternative data sources. The latter included official (INDE) and unofficial sources (web mapping service). Thus, we assessed the positional accuracy of a number of the geospatial data sources so as to obtain an indication of the potential quality of the fix – expecting that the poorer the positional accuracy, the poorer the fix. The assessment included satellite images from the web mapping service, 1:25,000-scale topographic maps of the Brazilian Army, a 1:50,000-scale geological map, and a 1:100,000-scale soil map. Both soil and geological maps were produced based on the most recent topographic maps produced by the Brazilian Army on equivalent cartographic scales – see Samuel-Rosa et al. (2015) for details. Due to operational constraints, the assessment was carried out in an area of about 150 km² located in the southern border of the plateau of the Paraná Sedimentary Basin, in the state of Rio Grande do Sul (Figure 2A). A set of 14 ground control points (GCP) was used. Their locations were defined based on the existence of easily identifiable geographical markers across the sources of geospatial data.

Positional accuracy assessment consisted of comparing the x- and y-coordinates of GCPs (observed value, z) with the x- and y-coordinates of the respective geographical markers visually identified on the sources of

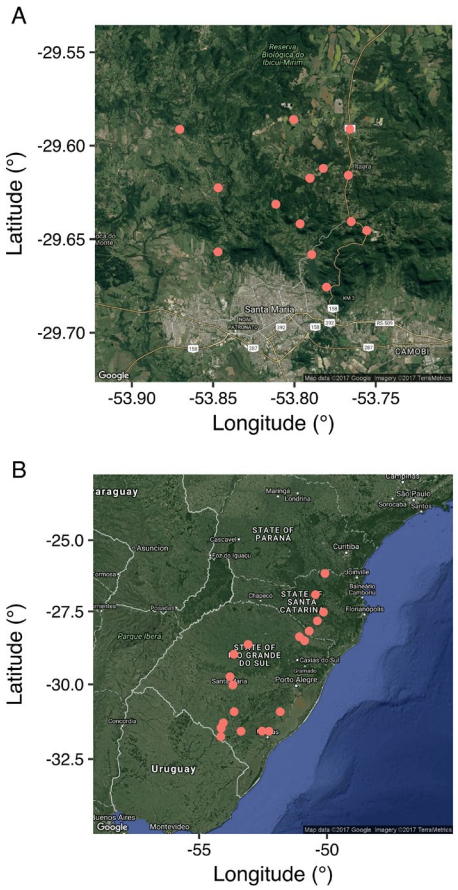

Figure 2 – (A) The 14 ground control points used for the positional validation of some alternative sources of geospatial data. (B) The 20 soil observations used for the positional validation of spatial coordinates estimated from textual descriptions using web mapping services.

geospatial data (predicted value, $\hat{z}$). The differences $e_{ij} = \hat{z}_{ij} - z_{ij}$, $i = 1, 2, \ldots, n$, j being either the y- or x-coordinate, and n the number of GCPs, were used to calculate the mean error (ME), Eq. (1), mean absolute error (MAE), Eq. (2), and root mean squared error (RMSE), Eq. (3).

$$ME_j = \frac{1}{n}\sum_{i=1}^{n} e_{ij} \quad (1)$$

$$MAE_j = \frac{1}{n}\sum_{i=1}^{n} \left|e_{ij}\right| \quad (2)$$

$$RMSE_j = \left(\frac{1}{n}\sum_{i=1}^{n} e_{ij}^2\right)^{0.5} \quad (3)$$

The error vector was computed to assess the absolute displacement of a geospatial data source in relation to its true, correct position on the Earth's surface. It was characterized with respect to its modulus (distance), $m_i = (e_{ix}^2 + e_{iy}^2)^{0.5}$, m being the Euclidean distance between observed and estimated observation location, and azimuth (direction), $\alpha_i = 180/\pi \cdot \text{atan2}(e_{ix}, e_{iy})$, the angle in relation to the North. If $e_{iy} < 0$, we added 180° to α_i; if $\alpha_i < 0$, we added 360° to α_i. The mean and root mean square modulus were computed as well as the mean azimuth.

Accuracy of estimated coordinates

Web mapping services are not considered formal data sources for fixing inconsistencies in the geospatial data of legacy soil observations. Thus, we assessed the accuracy of a number of the spatial coordinates estimated from the textual descriptions of the observation locations using the web mapping service. A set of 20 soil observations, covering the two southernmost Brazilian states (Figure 2B) (Curcio et al., 2000) was used for this purpose. The true spatial coordinates of these observations have not been published in the study report. After they had been estimated using the web mapping service, we contacted the soil survey experts that participated in the study to obtain a copy of the unpublished true spatial coordinates. The accuracy assessment consisted of calculating the deviation of the spatial coordinates estimated by using the web mapping service from the true ones observed in the field using the equations described in the previous section.

Results

Geospatial data inconsistencies

We found inconsistencies of varied types and magnitudes in the geospatial data of the open legacy survey soil data downloaded from BDSolos (Table 1). Some of these were easier to spot, such as the observations having the wrong value for the longitudinal or latitudinal hemisphere, or simply not having any data on the hemisphere and geodetic datum. Moreover, the observations with coordinates far outside the rectangle spanning the Brazilian territory. Aside from these, several observations were found to fall outside of Brazilian territory due to less evident inconsistencies (Figure 3A). The two largest groups of inconsistent observations were from the states of Amazonas and Rio de Janeiro. The first group consisted of observations made in the municipality of São Gabriel da Cachoeira in a single survey during the 1970s. The second group was composed of observations belonging to different soil surveys carried out in Rio de Janeiro. In general, the geospatial data of these misplaced observations was easily corrected by checking the original survey reports. But in certain cases the recorded geographic coordinates were in complete disagreement with the textual description of the observation location, possibly meaning that the coordinates were incorrect. Overall, we found mistakes, both in original survey reports and data downloaded from BDSolos, to be the main reason for the misallocation of observations. The involuntary substitution of symbols and numeric characters with similar appearance such as [−8°, −20°, −51°, −4°, −42°] being confused with [−3°, −28°, −61°, +4°, −40°] and vice-versa was the most common mistake. Where the CRS was also missing in survey reports we assumed it to be the SIRGAS 2000, currently the official geodetic datum of Brazil (Appendix 3).

We also found several observations lying outside the federative unit in which they were recorded to fall (Figure 3B). Most of them were close to state borders, suggesting the occurrence of small errors in the coordinates, possibly coming from the coarse-scale base maps used during the survey. Another possible source for this inconsistency are the recent changes in the borders of

Table 1 – Geospatial data inconsistencies found in the open legacy soil data in Brazil.

Geospatial data	Inconsistency found
Geodetic datum	- no data, except in recent studies;
Geographical hemisphere	- value 'East' for the longitudinal hemisphere; - value 'North' for the latitudinal hemisphere in state other than Amazonas, Amapá, Pará, or Roraima; - no data, mostly in Rio de Janeiro; - opposite latitudinal hemisphere, mostly in northern states;
Spatial coordinates	- latitude larger than 34° S, mostly in Acre; - longitude larger than 74° W, mostly in Acre; - longitude smaller than 28° W; - minutes and seconds > 60; - easting < 1000 m and/or northing < 10 000 m; - outside the Brazilian territory, mostly in Amazonas and Rio de Janeiro; - outside the recorded federative unit;

certain states. In other cases, errors in the coordinates were somewhat large, generally arising from the same kind of mistakes that we identified in observations falling outside the Brazilian territory. The exceptions are two groups of observations from the state of Tocantins, both of them having been obtained before 1989, the year in which Tocantins was officially separated from the state of Goiás. Given these circumstances, it is reasonable to expect that these observations would be registered in BDSolos as being from Goiás although they now fall in Tocantins. This was the case of the first group of observations, located in the south of Tocantins.

The second group, located in the north of Tocantins, was registered in BDSolos as being from Pará. For a reason of which we remain ignorant, soil survey authors annotated the word 'Pará' in field soil descriptions that were actually made in Tocantins, thus suggesting that instead, the observations had been made in the state of Pará. This annotation was often accompanied by the name of a city in Pará, generally Castanhal or Anhangá (São Francisco do Pará), instead of Araguatins, or a city nearby in Tocantins, where the observations were actually made (Figure 2B). For example, the description of Profile 80 (the translation is ours): "5 km SW of Xambicá, along the Araguaia River; Latitude 6°30′ S; Longitude 48°38′ W; Municipality of Anhangá, Pará" (Camargo et al., 1975). Note that the coordinates are correct, placing the observation along the Araguaia River in the municipality of Xambioá, Tocantins. For a number of observations, when the data was entered in BDSolos, the coordinates were deleted or replaced, possibly because the additional information making reference to a city in the state of Pará created an equivocated mistrust about their veracity. Only by consulting soil surveys experts we were able to understand this inconsistency and find the proper solution, i.e. ignore the word 'Pará'.

Recovering missing spatial coordinates

Visiting original survey reports, using alternative sources of geospatial data and consulting soil survey experts enabled us to reach a total of 3,197 observations possessing geospatial data without apparent gross inconsistencies. This represented only 52 % of the total number of observations being processed. The states that concentrated most of the remaining 2,998 observations (48 %) with missing geospatial data were Bahia, Minas Gerais, Paraná, Pará, São Paulo, and Rio Grande do Sul, with 440, 360, 307, 199, 180 and 176 observations, respectively (Figure 3C). The spatial distribution of the georeferenced observations in these six states suggests that the compilation of open legacy soil survey data in Brazil has been determined by local demands for soil information.

The nation-wise dataset compiled by Cooper et al. (2005), used in conjunction with our matching algorithm, enabled us to recover the coordinates of another 921 observations. Six of these observations had to undergo further processing because they fell outside Brazilian territory (Figure 1). This amounted to 64 % of the observations being processed having geospatial data without apparent gross inconsistencies (Figure 3D). Considerable gains in spatial coverage were achieved in several states, mainly Paraná, São Paulo, Minas Gerais, Bahia, Ceará, Rio Grande do Norte, Pernambuco, Paraíba, and Espírito Santo. The region covered by these states is known to have concentrated the majority of the more detailed soil survey efforts in Brazil during the last century.

Despite the clear improvements, large areas remain with very poor spatial coverage in many states. First, this is because the coordinates of 2,077 observations are unknown to us. The dataset compiled by Cooper et al. (2005) is known to contain most of these coordinates, but our matching algorithm was inefficient in finding them. A closer look at both datasets revealed this was due to many observations being identified under different names, codes or references. This is likely because there is no consensus or an official protocol on how to compile and organize open legacy soil data in Brazil. The second reason for the observed poor coverage is the fact that we ignored 2,924 observations from BDSolos because they did not contain any soil iron data (Figure 1). These observations likely present geospatial data inconsistencies similar to those described so far. Thus, it is reasonable to expect that, by following the workflow of the present study, we could reach a total of approximately 5,989 consistently georeferenced soil observations covering the Brazilian territory.

Accuracy of geospatial data sources

The accuracy assessment performed in a mountainous area of southern Brazil (Figure 2A) showed that fine resolution satellite images (1-2 m) available on the web mapping service have a higher positional accuracy than other more traditional, official sources of geospatial data (Table 2). According to Presidential Decree No. 89 817/1984, high-quality Brazilian map standards require that at least 90 % of the GCPs have positional errors smaller than 13 m, and an overall positional error (i.e. RMSE) less than 8 m on the cartographic scale of 1:25,000. In the case of the 1:25,000 topographic maps that we evaluated, the RMSE was eight times greater than the limit established by the legislation. Accordingly, the positional accuracy was poorer for the coarser scale soil and geological maps. In comparison, the RMSE of the tested web mapping service complies with Brazilian legislation and is similar to that generally reported in the field by popular Global Navigation Satellite System (GNSS) receivers.

Accuracy of estimated coordinates

The assessment of the accuracy of the estimated spatial coordinates of the 20 soil observations of Curcio et al. (2000) (estimated using the textual descriptions of their locations and a web mapping service) showed that

Table 2 – Positional accuracy of sources of geospatial data. Statistics are the mean, mean absolute, and root mean squared error (*e*) in the x- and y-coordinates, the mean error vector module and azimuth, and the root mean squared error vector module.

Statistic	e_x (m)	e_y (m)	Error vector (m)	Azimuth (°)
Google Maps satellite imagery at a spatial resolution of 1-2 m				
Mean	-1	3	6	184
Mean absolute	3	5	-	-
Root mean squared	4	8	8	-
Topographic maps on a cartographic scale of 1:25,000				
Mean	50	27	63	63
Mean absolute	50	32	-	-
Root mean squared	56	34	65	-
Geologic map on a cartographic scales of 1:50,000				
Mean	10	-102	140	169
Mean absolute	43	125	-	-
Root mean squared	59	134	147	-
Soil map at a cartographic scale of 1:100,000				
Mean	30	-36	105	128
Mean absolute	58	64	-	-
Root mean squared	85	76	114	-

absolute positional estimation errors can be as small as 29 m (Table 3). This is equivalent to the precision of coordinates measured in arc-seconds (~ 30.8 m) as reported in semi-detailed soil survey reports and is more accurate than that observed in 1:25,000 topographic maps (> 60 m). However, in general, the length of the error vector was somewhat variable, concentrating between about 100 and 500 m. In comparison, this was the level of positional accuracy observed for geologic and soil maps with cartographic scales of 1:50,000 and 1:100,000, respectively (Table 2). Poorer results were observed for five out of the twenty soil observations tested. Three of them had an error vector of about 1 km – which is still more accurate than coordinates measured in arc-minutes (~1.85 km). Such errors are due to the fact that observation location descriptions commonly refer to road distances using the kilometer as the measurement unit, setting the practical quantification limit to ~1 km – plus six meters, the average positional error of the web mapping service (Table 2). For

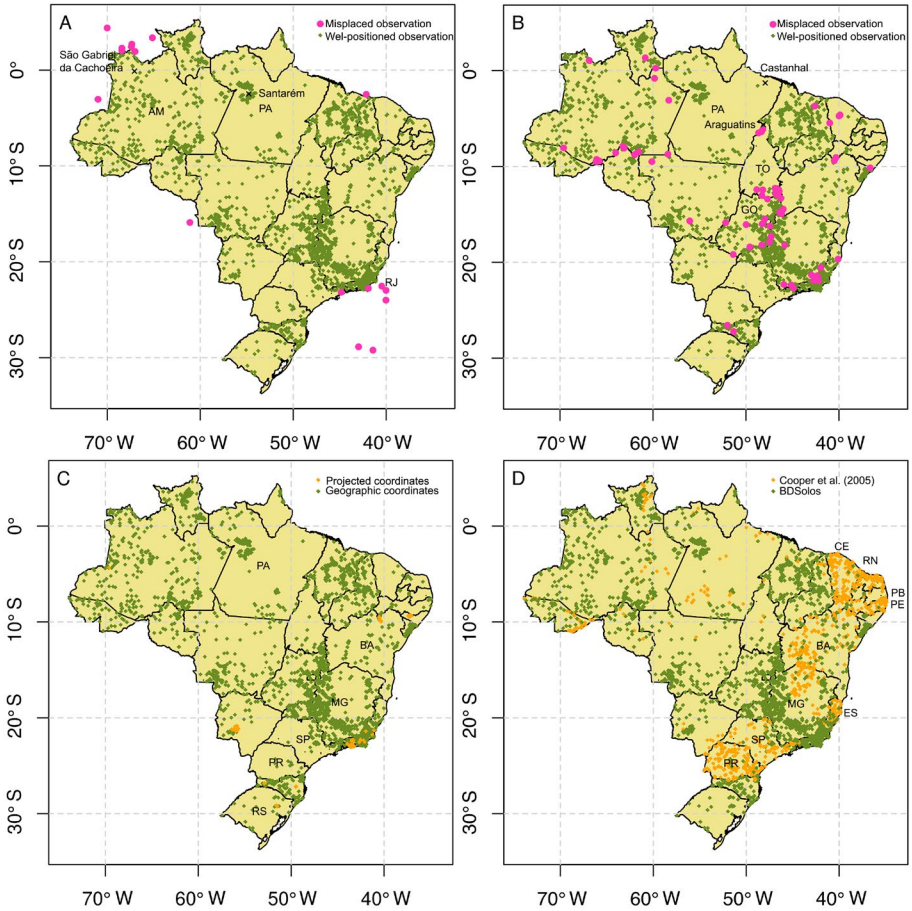

Figure 3 – Four selected moments of our study. (A) Geographic coordinates of observations lying outside of Brazil were corrected. (B) Geographic coordinates of observations lying outside of the recorded state were corrected. (C) Observations with geographic coordinates were grouped with those containing projected coordinates. (D) Geographic coordinates compiled by Cooper et al. (2005) were attributed to the remaining observations. Name of selected states: AM = Amazonas, BA = Bahia, CE = Ceará, ES = Espírito Santo, GO = Goiás, PA = Pará, PB = Paraíba, PE = Pernambuco, PR = Paraná, RJ = Rio de Janeiro, RN = Rio Grande do Norte, RS = Rio Grande do Sul, SP = São Paulo, TO = Tocantins.

Table 3 – Absolute error of the estimated spatial coordinates of twenty soil observations of Curcio et al. (2000). The estimation consisted of using textual descriptions of the observation locations to find the most likely coordinates in a web mapping service.

Profile	x-coord	y-coord	Error vector
	m		
01	3,752	2,045	4,273
02	266	431	506
03	240	460	519
04	150	55	159
05	195	538	572
06	25	12	28
07	13	104	105
08	87	429	437
09	65	116	133
10	197	93	218
11	180	28	182
12	56	996	998
13	52	47	70
14	1,050	516	1,170
15	35	38	52
16	7,738	7,672	10,897
17	22	18	29
18	40	58	71
19	1,142	235	1,166
20	32	2	32

example, Profile 12 (the translation is ours): "Tupanciretã (RS), route Santa Maria - Cruz Alta, 7 km after the access road to Tupanciretã". When more accurate or richer details are present in the description – such as using the meter as a measurement unit or describing a geographical marker – then the positional accuracy of the estimated coordinates will likely be considerably higher. For instance, Profile 06 (the translation is ours): "Bagé (RS), route Bagé - Aceguá, 15.3 km from the entrance gate of Bagé". These results show that, on average, coordinates estimated using the web mapping service are more accurate than coordinates reported in coarser scale soil survey reports and maps. This means that the coordinates of observations from these surveys could be fine-tuned so as to increase their positional accuracy by using free web mapping services – provided there is a detailed description of their locations.

Finally, we observed unacceptably large positional errors (> 1 km) for the estimated coordinates of Profile 01 (4 km) and Profile 16 (10 km) (Table 3). However, we found that, instead of the estimation procedure, the source of the errors was the measured field coordinates and the textual description of the observation location in BDSolos, respectively. While the description of Profile 01 reports that it was observed in a highway cutting, its field coordinates refer to a site about 4 km southeast of that highway in the middle of what seems to be a paddy field. By consulting soil survey experts that participated in the study, we verified that the textual description was correct. The estimated coordinates were fine-tuned with the help of these soil surveys experts, who then recommended that the estimates be used to replace the coordinates measured in the field. This means that the final positional error of the estimated coordinates likely is considerably lower than 4 km, possibly between ~100 and ~500 m, the most common range of values observed for the remaining 18 observations (Table 3).

For Profile 16, the geospatial description found in the study report states that the observation was made at about 5 km from the Pelotas River, in a soil pit near the road Vacaria-Lages. This description is somewhat vague, not making clear on which side of the Pelotas River the observation was made. When the data was entered in BDSolos, it was registered that the observation was made south of the Pelotas River, in the municipality of Vacaria, in the state of Rio Grande do Sul. However, its field coordinates refer to a location near the same road but on the other side of the Pelotas River, in the municipality of Capão Alto, in the state of Santa Catarina. Soil survey experts that participated in the study confirmed that the coordinates recorded in the field were correct. Thus, as for Profile 01, the positional error was not due to the estimation procedure using the web mapping service, but to the inaccuracy of the textual description of the observation location.

Discussion

The usefulness of open legacy soil data for spatial applications such as DSM depends on the accuracy of the geospatial data. This paper showed that, in the case of the Brazilian open legacy soil survey data, geospatial data inconsistencies of several types and magnitudes come from various sources, including typing mistakes in base maps and survey reports. However, we have also shown that it is possible to recover missing coordinates and correct gross errors in geospatial data by using various free geospatial data sources and consulting soil surveys experts. The latter can help understanding vague geospatial descriptions and fine-tune the estimated coordinates. The tested web mapping service was very useful for fixing geospatial data inconsistencies, not only due to its many operational advantages – fine resolution satellite imagery and up-to-date road maps, fast search engine equipped with a distance calculator, user-friendly interface with intuitive controls –, but also to its much higher positional accuracy as compared to more traditional, official geospatial data sources. Overall, the positional accuracy of the coordinates estimated using the web mapping service depends almost entirely on the precision and level of detail of the textual description of the observation location found in the survey report. In the worst case, the estimated spatial coordinates can be as or more accurate than coordinates measured in arc-minutes as found in coarse scale survey reports and maps.

The evaluation of the geospatial data inconsistencies showed that base maps and survey reports cannot be entirely trusted. This is because of the presence of typing mistakes in recorded coordinates (relatively common), mismatched information in geospatial descriptions (less common), and errors in coordinates recorded in the field (uncommon). Only through the active participation of soil survey experts with a certain level of familiarity with the soil data source could such inconsistencies be properly eliminated. For this reason, we think that only trained personnel, with experience in soil science and/or geography, have the necessary skills to effectively improve the quality of geospatial data of open legacy soil data. By trained personnel we mean experts in soil science and/or geography or, alternatively, students under the supervision of soil scientists and/or geographers. Despite this, the participation of non-specialists in soil research has gained popularity (Rossiter et al., 2015; Ringrose-Voase et al., 2017). We remain skeptical of its usefulness for the purpose of correcting geospatial data inconsistencies. For the time being, we believe that a joint, collaborative and lasting geospatial data rescue effort among Brazilian soil experts and institutions should be fostered (Table 4) – as it has already been done elsewhere as shown by Arrouays et al. (2017) and Batjes et al. (2017).

Fostering public and collaborative open legacy soil data compilation and cleaning projects is an efficient way to avoid duplicated efforts and thus save public resources. Despite this, we recognize that our own present effort is a partial duplication of previous works, e.g. Chagas et al. (2004), Cooper et al. (2005), Ottoni et al. (2014) and Simões et al. (2015). However, this duplication is justified by the fact that several factors – divergence in goals, technological and methodological choices, institutional policies – limit the continuation and improvement of previous work by others. A proof of this is the fact that previous initiatives were unable to solve the problem of permanently safeguarding all kinds of soil data in Brazil. This is the reason why we have recently joined forces with other Brazilian soil scientists to develop the febr, where all data processed in the present study have already been made publicly available. With the proper computer algorithm, this data can be immediately used to update the geospatial data of observations in BDSolos (Table 4). This avoids the future duplication of efforts and promotes the reuse of open legacy soil data in Brazil.

We recognize that work still has to be done to improve the quality of the existing open legacy soil survey data for large scale spatial applications in Brazil – Table 4 shows a number of broad recommendations. Upon completion of the present study, there were less than five thousand observations in febr ready to be used for the DSM planned in the PronaSolos. According to recent experiences, reasonably accurate large scale DSM models can be calibrated using only a few hundred observations (Padarian et al., 2017). The main requirement is that soil observations cover the main environmental factors controlling the large scale soil variation. Perhaps the set of observations in febr already meets this requirement. However, we think that with the active collaboration between soil scientists through febr – specially soil surveys experts – and by using a free web mapping service to improve geospatial data, considerable improvements can be achieved in the very short term.

Finally, it remains to be evaluated if the positional accuracy of estimated coordinates influences the performance of large scale DSM models, for example, carrying out uncertainty propagation analyses (Nol et al., 2010). For the time being, if the spatial resolution of covariates and spatial predictions is ≥ 1000 m, such as used by Hengl et al. (2014), then increasing the accuracy of coordinates would only marginally improve DSM perfor-

Table 4 – General recommendations for soil scientists, maintainers of the Brazilian Soil Information System (BDSolos), and soil scientists compiling open legacy soil data.

Target	Recommendations
Field soil scientists	- Describe the location of the observations by making explicit reference to well known identifiable geographic markers such as states, cities, roads and rivers, preferably indicating the distance from these markers, avoiding using only a farm or business name, AND - Recording the spatial coordinates using a geographic coordinate system, with latitude and longitude values in the degree-minute-second format, the seconds annotated with at least two decimal digits, and SIRGAS 2000 as geodetic datum. For example: Latitude 15°48'02.41" S Longitude 47°51'40.64" W SIRGAS, 2000.
BDSolos maintainers	- Develop a computer algorithm to harvest the data in febr that we have already processed to update the corresponding soil observations contained in BDSolos, AND - A data entry interface to be fed with comma-separated-value files thus enabling soil scientists to directly upload soil data that is already structured in bidimensional tables.
BDSolos maintainers and soil scientists compiling legacy soil data	- Use bidimensional tables, as available in popular spreadsheet software such as MS Office Excel, LibreOffice Calc, to simplify the soil data typing process and enable gross inconsistencies to be easily caught by eye, AND - Double blind data entry as data quality control strategy, by which two people in different locations, working offline, enter the same data at different times without communicating with each other, OR - Concomitant data entry as data quality control strategy, by which multiple people in different locations, working online, enter data and cross-check the data entered by others, possibly at the same time and communicating with each other, AND - Online mapping services, such as Google Maps and OpenStreetMap, to evaluate the consistency of existing spatial coordinates of soil observations, as well as to find the most probable spatial coordinates when these are missing based on the description of the location of soil observations.

mance. The influence would likely be greater when making spatial predictions at finer spatial resolutions, such as those intended for PronaSolos (≤ 100 m). In this case, using coarser resolution covariates could be considered given that these commonly show a stronger correlation with soil variables (Samuel-Rosa et al., 2015). Alternatively, when calibrating DSM models, soil observations could receive weights that are inversely proportional to their positional accuracy. If unknown, a gross estimate of the latter could be approximated based on the geospatial data source – cartographic scale or spatial resolution plus an estimate of its positional accuracy – and the description of the observation location – precision and level of detail. The results obtained in this study could be used as a starting point for these approximations. However, the reader should be aware that it is not clear how the errors coming from geospatial data sources and observation location descriptions combine together and affect one another. Further studies are needed in this area. The effectiveness of these strategies could then be checked by using a sound validation method and evaluating the uncertainty of DSM predictions (Brus et al., 2011). Highly influential observations – due to positional inaccuracies – could simply be dropped from DSM models or, preferentially, be subjected to review by soil survey experts.

Acknowledgments

We are grateful to two anonymous reviewers for helpful comments on the original version of the paper. The following colleagues helped in different phases of data collection and/or processing: Dr. Pablo Miguel (Universidade Federal de Pelotas - UFPel) and MSc. Luis Fernando Chimelo Ruiz (Universidade Federal do Rio Grande do Sul - UFRGS). We also thank the development/maintenance teams and module/package authors of the free and open source software (FOSS), operational system (OS), and geospatial data platforms that were used to develop our study. The first author was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Ministério da Educação do Brasil, through the Programa Nacional de Pós Doutorado (PNPD/CAPES). The second and third authors were supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Ministério da Ciência, Tecnologia, Inovações e Comunicações (310247/2015-2 and 142492/2014-0, respectively).

Authors' Contributions

Conceptualization: Samuel-Rosa, A. Data acquisition: Samuel-Rosa, A.; Dalmolin, R.S.D.; Moura-Bueno, J.M.; Teixeira, W.G.; Alba, J.M.F. Data analysis: Samuel-Rosa, A. Design of methodology: Samuel-Rosa, A. Software development: Samuel-Rosa, A. Writing and editing: Samuel-Rosa, A.; Dalmolin, R.S.D.; Moura-Bueno, J.M.; Teixeira, W.G.; Alba, J.M.F.

References

Arrouays, D.; Leenaars, J.G.; Forges, A.C.R.; Adhikari, K.; Ballabio, C.; Greve, M.; Grundy, M.; Guerrero, E.; Hempel, J.; Hengl, T.; Heuvelink, G.; Batjes, N.; Carvalho, E.; Hartemink, A.; Hewitt, A.; Hong, S.-Y.; Krasilnikov, P.; Lagacherie, P.; Lelyk, G.; Libohova, Z.; Lilly, A.; McBratney, A.; McKenzie, N.; Vasquez, G.M.; Mulder, V.L.; Minasny, B.; Montanarella, L.; Odeh, I.; Padarian, J.; Poggio, L.; Roudier, P.; Saby, N.; Savin, I.; Searle, R.; Solbovoy, V.; Thompson, J.; Smith, S.; Sulaeman, Y.; Vintila, R.; Rossel, R.V.; Wilson, P.; Zhang, G.-L.; Swerts, M.; Oorts, K.; Karklins, A.; Feng, L.; Navarro, A.R.I.; Levin, A.; Laktionova, T.; DellAcqua, M.; Suvannang, N.; Ruam, W.; Prasad, J.; Patil, N.; Husnjak, S.; Pásztor, L.; Okx, J.; Hallett, S.; Keay, C.; Farewell, T.; Lilja, H.; Juilleret, J.; Marx, S.; Takata, Y.; Kazuyuki, Y.; Mansuy, N.; Panagos, P.; Liedekerke, M.V.; Skalsky, R.; Sobocka, J.; Kobza, J.; Eftekhari, K.; Alavipanah, S.K.; Moussadek, R.; Badraoui, M.; Silva, M.D.; Paterson, G.; da Conceição Gonçalves, M.; Theocharopoulos, S.; Yemefack, M.; Tedou, S.; Vrscaj, B.; Grob, U.; Kozák, J.; Boruvka, L.; Dobos, E.; Taboada, M.; Moretti, L.; Rodriguez, D. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14: 1-19. DOI:10.1016/j.grj.2017.06.001

Batjes, N.H.; Ribeiro, E.; van Oostrum, A.; Leenaars, J.; Hengl, T.; Jesus, J.M.D. 2017. WoSIS: providing standardised soil profile data for the world. Earth System Science Data 9: 1-14. DOI:10.5194/essd-9-1-2017

Brus, D.J.; Kempen, B.; Heuvelink, G.B.M. 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62: 394-407. DOI:10.1111/j.1365-2389.2011.01364.x

Camargo, M.N.; Freitas, F.G.; Beek, K.J.; Garland, L.E.; Ramalho Filho, A.; Tomasi, J.M.G.; Castello, D.S. 1975. Schematic Soil Map of the North, Mid-North and Mid-West Regions of Brazil. = Mapa Esquemático dos Solos das Regiões Norte, Meio-Norte e Centro-Oeste do Brasil. Embrapa Solos, Rio de Janeiro, RJ, Brazil (in Portuguese).

Chagas, C.S.; Carvalho Junior, W.; Bhering, S.B.; Tanaka, A.K.; Baca, J.F.M. 2004. Organization and structure of the Brazilian soil information system (SigSolos - version 1.0). Revista Brasileira de Ciência do Solo 28: 865-876 (in Portuguese, with abstract in English). DOI:10.1590/S0100-06832004000500009

Cooper, M.; Mendes, L.M.S.; Silva, W.L.C.; Sparovek, G. 2005. A national soil profile database for Brazil available to international scientists. Soil Science Society of America Journal 69: 649-652. DOI:10.2136/sssaj2004.0140

Curcio, G.R.; Carvalho, A.P.; Bognola, I.A.; Gomes, I.A.; Rossi, M.; Coelho, M.R.; Santos, R.D. 2000. Meeting of Correlation, Classification and Application of Soil Surveys: Tour Guide for Soil Studies in the States of Rio Grande do Sul, Santa Catarina and Paraná. = Reunião de Correlação, Classificação e Aplicação de Levantamentos de Solos: Guia de Excursão de Estudos de Solos nos Estados do Rio Grande do Sul, Santa Catarina e Paraná. Embrapa Florestas, Colombo, PR, Brazil (in Portuguese).

Hengl, T.; Jesus, J.M.; MacMillan, R.A.; Batjes, N.H.; Heuvelink, G.B.M.; Ribeiro, E.; Samuel-Rosa, A.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; Gonzalez, M.R. 2014. SoilGrids1km: global soil information based on automated mapping. PLoS ONE 9: e105992. DOI:10.1371/journal.pone.0105992

Nol, L.; Heuvelink, G.B.M.; Veldkamp, A.; de Vries, W.; Kros, J. 2010. Uncertainty propagation analysis of an N_2O emission model at the plot and landscape scale. Geoderma 159: 9-23. DOI:10.1016/j.geoderma.2010.06.009

Odgers, N.P.; Sun, W.; McBratney, A.B.; Minasny, B.; Clifford, D. 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214-215: 91-100. DOI:10.1016/j.geoderma.2013.09.024

Ottoni, M.V.; Lopes-Assad, M.L.R.C.; Pachepsky, Y.; Rotunno Filho, O.C. 2014. A hydrophysical database to develop pedotransfer functions for Brazilian soils: challenges and perspectives. p. 467-494. In: Teixeira, W.G.; Ceddia, M.B.; Ottoni, M.V.; Donnagema, G.K., eds. Application of soil physics in environmental analyses. Springer International, Basel, Switzerland.

Padarian, J.; Minasny, B.; McBratney, A. 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Regional 9: 17-28. DOI: 10.1016/j.geodrs.2016.12.001

Polidoro, J.C.; Mendonça-Santos, M.L.; Lumbreras, J.F.; Coelho, M.R.; Carvalho Filho, A.; Motta, P.E.F.; Carvalho Junior, W.; Araújo Filho, J.C.; Curcio, G.R.; Correia, J.R.; Martins, É.S.; Spera, S.T.; Oliveira, S.R.M.; Bolfe, E.L.; Manzatto, C.V.; Tôsto, S.G.; Venturieri, A.; Sá, I.B.; Oliveira, V.Á.; Shinzato, E.; Anjos, L.H.C.; Valladares, G.S.; Ribeiro, J.L.; Medeiros, P.S.C.; Moreira, F.M.S.; Silva, L.S.L.; Sequinatto, L.; Aglio, M.L.D.; Dart, R.O. 2016. Brazil's National Soil Program (PronaSolos) = Programa Nacional de Solos do Brasil (PronaSolos). Embrapa Solos, Rio de Janeiro, RJ, Brazil. (in Portuguese).

Ringrose-Voase, A.J.; Grealish, G.J.; Thomas, M.; Wong, M.; Glover, M.; Mercado, A.; Nilo, G.; Dowling, T. 2017. Four Pillars of digital land resource mapping to address information and capacity shortages in developing countries. Geoderma. DOI: 10.1016/j.geoderma.2017.10.014

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

Samsonova, V.P.; Meshalkina, J.L.; Blagoveschensky, Y.N.; Yaroslavtsev, A.M.; Stoorvogel, J.J. 2018. The role of positional errors while interpolating soil organic carbon contents using satellite imagery. Precision Agriculture. DOI: 10.1007/s11119-018-9575-4

Samuel-Rosa, A.; Heuvelink, G.B.M.; Vasques, G.M.; Anjos, L.H.C. 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244: 214-227. DOI: 10.1016/j.geoderma.2014.12.017

Simões, M.G.; Oliveira, S.R.M.; Ferraz, R.P.D.; Santos, H.G.; Manzatto, C.V. 2015. Democratization of information on Brazilian soils: geoportal and soil database accessible via web. Cadernos de Ciência e Tecnologia 32: 55-69 (in Portuguese, with abstract in English).

Appendix

Appendix 1 – Download all open soil data in BDSolos

These are the steps to download all of the existing open soil data in BDSolos:

1. Access https://www.bdsolos.cnptia.embrapa.br/consulta_publica.html

2. Read the terms of use and, if you agree, press *Concordo*.

3. In the next window (*Etapa 1: Seleção de Atributos*), select all works (*Trabalhos*) and all attributes of sampling points (*Pontos de Amostragem*) and horizons (*Horizontes*), and press *Ir para a etapa 2*.

4. In the next window (*ETAPA 2: Seleção de Filtros*), filter the works by their identification (*Identificação*) and, in the pop-up window, set the criteria for the publication year (*Ano de Publicação*) to less than the next year (< 2018), press OK and then *Visualizar Resultados*.

5. In the next window (*ETAPA 3: Seleção de Resultados*), select all works (*Todos*), choose the CSV format for download, and press *Visualizar Resultados Selecionados*.

Appendix 2 – Estimating spatial coordinates using Google Maps

Consider the following description of the location of a soil observation extracted from Curcio et al. (2000) (the translation is ours): Profile 12 (the translation is ours): "Tupanciretã (RS), route Santa Maria - Cruz Alta, 7 km after the access road to Tupanciretã". These are the steps to estimate the most likely spatial coordinates of this observation:

1. Access www.google.com.br/maps/

2. In the search box, type 'Cruz Alta - RS' and press *Enter* – the map will zoom into the municipality of Cruz Alta.

3. Press *Directions* and, in the new search box, type 'Santa Maria - RS' and press *Enter* – the map will zoom into the road Santa Maria - Cruz Alta.

4. Drag the marker of the city of Santa Maria to the entrance of access road to Tupanciretã in highway BR 158.

5. Drag the marker of the city of Cruz Alta to about 7 km of the first marker on the road Santa Maria - Cruz Alta.

6. Enable the satellite view. Evaluate if the location of the second marker approximately matches the description of the location of the observation. Make adjustments if needed.

7. Right-click on the second marker. Select *What's here?* in the pop-up window.

8. Copy the spatial coordinates that appear in the pop-up window – these are the most likely coordinates of the observation given the textual description of its location.

Appendix 3 – Codes of coordinate reference systems used in Brazil

Table A3.1 shows the codes of the Geodetic Parameter Dataset of the European Petroleum Survey Group (EPSG) for coordinate reference systems based on the Universal Transverse Mercator (UTM) conformal projection and geodetic datums most commonly used in Brazil. In comparison, geographic coordinates systems that use Córrego Alegre, SAD69, SIRGAS 2000, or WGS84 as geodetic data are coded EPSG:4225, EPSG:4618, EPSG:4674, and EPSG:4326, respectively.

Table A3.1 – Codes of the coordinate reference systems most commonly used in Brazil.

UTM zone and hemisphere[1]	Geodetic data			
	Córrego Alegre[2]	SAD69	WGS 84	SIRGAS 2000
18N	-	29168	32618	31972
18S	-	29188	32718	31978
19N	-	29169	32619	31973
19S	-	29189	32719	31979
20N	-	29170	32620	31974
20S	-	29190	32720	31980
21S	22521	29191	32721	31981
22S	22522	29192	32722	31982
23S	22523	29193	32723	31983
24S	22524	29194	32724	31984
25S	22525	29195	32725	31985

[1]S = South hemisphere, N = North hemisphere. [2]The geodetic datum Córrego Alegre covers only the UTM zones from 21 to 25 in the South Hemisphere, '-' meaning 'not available'.