Best practices for using soil geographic databases

*David G. Rossiter**

May 8, 2026

Contents

* ISRIC-World Soil Information, Wageningen (NL)

The ISRIC website "Soil Geographic Databases"[1] is a comprehensive and regularly-updated collection of links to all freely-downloadable primary soils information usable in a GIS either as 'points', lines, polygons or grids (a.k.a. 'rasters'). It also includes links to ISRIC's own data products[2], including the WoSIS (World Soil Information Service)[3] 'point' database and the SoilGrids[4] world grids.

When a resource is identified, the question is: does it meet the user's requirements? Is it fit for purpose?

This document is a guide to the steps that should be taken to understand a dataset's origin, purpose, structure, and information. It is organized as a set of questions to be answered by the prospective user. These are supplemented with examples.

1 Documentation

We hope that the dataset is thoroughly documented. If not, we may have to guess from secondary information.

1. Is there formal metadata? How complete is it?

   For example, the SSURGO/STATSGO2 polygon databases from the USA have an extensive metadata page[5].

2. Same for informal metadata (e.g., text files, README's)

3. Is there a report or journal article that explains the dataset's origins?

   Ideally, the article would be in a journal that specialize in datasets and have standards for their papers, e.g., Earth System Science Data[6], or in journal sections such as the "Data articles" of the European Journal of Soil Science [20].

   An example data paper is by Helfenstein et al. [10].

4. Are there any papers or reports that use this dataset? Do they describe in sufficient detail how the dataset was used?

   An example of the use of the cited paper, found with Scopus citation search, is the paper of Brouwer et al. [3].

   Thousands of papers have cited the various versions of SoilGrids. These can be found with a search on Web of Science, Scopus, or Google Scholar. For example, a Google Scholar search on 07-May-2026 for "SoilGrids" finds as the first hit the SoilGrids v2.0 paper [19], and has a link to 1,972 citations to date. These suggest the possible uses of that dataset.

[1] https://docs.isric.org/soil-geographic-databases/
[2] https://isric.org/explore/
[3] https://isric.org/explore/wosis/
[4] https://isric.org/explore/soilgrids/
[5] https://www.nrcs.usda.gov/resources/data-and-reports/ssurgo/stats2go-metadata
[6] https://www.earth-system-science-data.net/

2 Administration

Datasets are expensive to build and maintain. We hope there is a responsible organization that created and/or supports them.

2.1 Lineage

1. Who created the dataset and for what purpose?

   This can imply the appropriate uses of the dataset. For example, "World Soil Information Service (WoSIS) is one of ISRIC' s flagship products …[it] aims to: (1) Safeguard world soil profile data in its originally-received format (especially for soil legacy data), (2) Quality assess and standardise soil data; (3) Provide standardised soil data for digital soil mapping and a range of environmental applications." This emphasizes the custodial and quality control/standardization aim, so it is suitable for cross-boundary projects.

2. Is there evidence of the creator's skill?

   This can be from the institution's reputation and experience, and from evidence in the metadata, papers describing and using the dataset, and reports.

3. Is there any reason to suspect bias due to political, commercial, or scientific interests? If so, what might be the bias and how could it be identified?

2.2 Permissions

1. Who owns the dataset?
2. What is the license of the dataset, and what are its terms?

   For example, is it one of the Creative Commons licenses[7]?

3. Are there restrictions on the use of the dataset? Can these be loosened by an agreement with the dataset owner?

   For example, is it permitted to reproduce the dataset as supplementary material to a paper or report? If so, what acknowledgement must be given?

3 Geographic scope and georeferencing

1. What political and/or geographic area is covered by the dataset?
2. Is the dataset intended to be comprehensive, or does it only cover part of the space, either geographically or thematically?

   For example, does the dataset only represent soils under certain land uses, e.g., forest soils? If so, how is the restriction defined?

[7] https://creativecommons.org/cc-licenses/

Example: The Dutch soil survey of the 1960's and 70's did not survey in then-urban areas, because the survey was paid for by the Agriculture Ministry.

For example, does the dataset not cover parts of a geographic area due to conflicts, access, or border disputes? Is this explicit or must it be inferred?

Example: Datasets purporting to cover the national area of the People's Republic of China often shown Taiwan and most of Arunchal Pradesh state (India), claimed by the PRC as part of the Tibet region, yet no field observations can be made there. This can only be inferred by knowledge of the various territorial disputes – which we here do not attempt to judge.

Sometimes this is explicit, for example in a paper describing the soil colour of China grid [14], the authors explain "[the point source data] did not cover Taiwan, which however has soil-forming environments similar to the cross-strait mainland province Fujian", showing that the part of the predictive map covering Taiwan is purely by extrapolation.

3. What is the coördinate reference system (CRS)?

   Ideally, this is given as an EPSG code. If in text, try to find it in the EPSG database[8]. For older datasets you may need to infer from coördinates. See the "Grids & Datums" series by Clifford Mugnier in Photogrammetric Engineering & Remote Sensing[9] for extensive discussion of the systems used in each country.

   Example: A polygon map was found for an area of interest in Kenya [13]. It was published by the Department of Agriculture, Kenya. The CRS is not given. However, the topographic base map is named in the map margins, from which some detective work [15] reveals that this was developed in the East African War System transverse Mercator projection, belt I on the Arc 1950 datum (EPSG:6209). Although the datum is in the EPSG database, the obsolete projection is not, so it must be defined by hand as a CRS[10], using the information in the Grids & Datums article. Then the map can be scanned, georeferenced with control points converted to a project CRS, and integrated into a GIS.

4. What is the bounding box?

5. For **points**, what is the accuracy of georeferencing?

   This may be given explicitly as a PDOP (Position Dilution Of Precision), by reference to the technology used from which a precision can be estimated, or by an estimate.

   For example, WoSIS tags each observation location with an approximate positional uncertainty in four classes (< 100 m, 100 m $\rightarrow 1$ km, $1 \rightarrow$

[8] `epsg.org` or the unofficial `epsg.io`
[9] Search the PE&RS site `https://my.asprs.org/PERS`
[10] for example, using the `st_crs` function of the `sf` package in `R`

10 km, > 10 km), as inferred from the coordinates given in the source datasets [2, §3.4.1].

Whether this uncertainty is adequate for a proposed use depends on that application. If attempting to return to a site for monitoring, the accuracy has to be quite precise. If used for machine learning soil mapping, the site just needs to be within the resolution of the gridded product, or at least within a radius with similar environmental covariates.

Some products have an imposed geographic uncertainty, e.g., by random distance and direction displacement, due to privacy regulations. The magnitude of the possible displacement compared to the scale of the soil landscape can indicate whether these points can be used for mapping.

6. For **polygons**, what is the *nominal map scale* of the paper map on which this was based? Convert that to minimum area that can be shown, using the conservative Cornell formula: $(S/1000)^2/250$, where S is the *scale number*, i.e.,the denominator of the map scale ratio.

   Older products may have poor geodetic control. For example, in the USA it proved impossible to rectify surveys published on semi-controlled corrected photomosaics [7], so boundary lines had to be re-drawn by hand on proper topographic base maps, before digitizing. So scans of the original soil survey reports are impossible to directly georeference for use in a GIS.

   Example: $1:50\,000 \rightarrow 50\,000, 50\,000/1000 = 50, 50^2 = 2500, 2500/250 = 10$ so a map at nominal scale $1:50\,000$ can not show any area smaller than 10 ha. For $1:500\,000$ this works out to 1000 ha $= 10$ km^2.

7. For **polygons**, what is the base map and how precise is it? Does it conform to map accuracy standards corresponding to its scale? Were the polygons drawn directly on the base map or transferred from other sources, such as unrectified aerial photographs?

8. For **grids**, what is the grid resolution, in units of the original CRS? If this is a geographic CRS (e.g., EPSG 4326), what is the approximate grid size in meters in the area of interest?

   Some grids may use variable grid sizes to indicate acceptable thematic accuracy: the size of the grid cell needed to reduce the uncertainty to a specified level [17]. In this case, which areas are shown at which resolutions?

4 Time period

1. What is the time period covered by the dataset? What could have changed since then?

   The obvious example is datasets of dynamic soil properties, such as soil organic Carbon, but there are other possible changes over time.

Example: The Dutch soil survey of the 1960's and 70's. Much has changed since then: groundwater levels [e.g., 9], land use and water management [e.g., 12], and urbanization converting the previous agricultural soils, often with extensive transport of soil material and land shaping.

Once the date of the dataset is known, search for information on possible changes that could affect its currency.

2. For point observations of dynamic soil properties that vary by season, is the date of sampling given?

5 Thematic information

5.1 Point datasets

These are observations of (parts of) soil profiles at single locations. Of course, not zero-dimensional "point", but they have a single georeference.

1. What are the data items?
2. Are any data items manipulations of primary information? Are any imputed?

   For example, data items may be inferred or supplemented with values obtained by pedotransfer functions or models.
3. For each data item, what are the units of measure?
4. For data items from laboratory measurement, what was the method? Is there a citation to a published source, e.g., [1, 5, 18, 25]? Is there a specified precision? This may be explicit or implied from the referred methods.
5. For soil class names, which soil classification system, which edition, which level?

   This may need to be inferred from the date of publication (if known) and the names themselves. For example, in early editions of Soil Taxonomy there was an *Ochrepts* suborder, which was eliminated in favour of the subgroups based on soil moisture regime in 1998 (8th edition of the Keys), so names like *Udepts* and *Ustepts* replaced the *Ochrepts*.

   Some databases, e.g., WoSIS, explictly give the reference to the classification system and publication year, if known from the original data source.

5.2 Polygons

These are soil class maps, often with ancillary information.

1. What themes are shown in the map?

For example, the Dutch 1:50 000 maps show single or multiple soil classes in the Dutch system, and also the summer and winter groundwater levels, and extra symbols for special soil conditions outside of the classification system [22].

2. What are the legend categories? Is there more than one legend? Does the legend include information other than soil classes?

   For example, the soil maps of the Netherlands include summer and winter groundwater levels.

3. What soil classification system, and at which categorical level, was used in the legend? See above under Points.

4. Are the map units of single soil classes or multiple? Is there information on the multiple classes and inclusions, i.e., their estimated proportion and arrangement within the map unit?

   For example, in the USA terminology, *consociations* are dominated by a single soil type,*associations* have a regular pattern of dissimilar soils, recognizable in detail on the landscape, whereas *complexes* have dissimilar soils in a fine or unpredictable pattern [21, Ch. 4].

5.3 Grids

1. What are the grid layers?

2. Do they represent the location as a whole, particular depth slices, or aggregates over specified depths?

   For example, SOC stock represents a location and is measured in T ha^{-1} or similar, whereas SOC concentration is measured in dg kg^{-1} or similar averaged over a specifed depth slice.

3. What are their units of measurement?

4. For soil classes, which is the soil classification system and which categorical levels are used? See above under Points.

5. Are the values point estimates of the grid cell centre, intended to represent the entire cell, or means?

   Most machine-learning methods predict at point support, i.e., the centre of the grid. Geostatistical methods can also predict at point support, but can also predict block means.

6. For soil classes, is the value a single class or a set of classes?

7. Is there an uncertainty layer corresponding to each value layer? If so, how is uncertainty expressed and how was it computed?

6 Data processing

6.1 Point datasets

All the steps from data collection to final presentation in the dataset should have been documented.

1. What was the sampling plan?

   Examples: Opportunistic, purposive, probability (e.g., stratified random), optimized for mapping by geostatistics or machine learning, spatial coverage. See the on-line book by Brus [4]. This has major implications for the dataset's representativeness.

   For example, the Soil Series project for Chinese Soil Taxonomy[e.g., 11] placed observations at the most typical location for each series, according to the judgement of the surveyor. So central concepts are well-represented but not transitions.

2. What was the field protocol?
3. What was the lab protocol?
4. How were the datasets "cleaned"? Were "outliers" (unusual values) removed, and on what grounds? Does this change the representativeness of the dataset?
5. How was the dataset evaluated for completeness and correctness?
6. How was uncertainty or accuracy evaluated?

6.2 Polygons

1. How was the polygon map created?

   For example, concept mapping from landscape interpretation, followed by field checks? Or directly from field survey?

2. What is the base map?

6.3 Grids

1. Are there specifications which this map followed [e.g., 6]?
2. How were the values in the grid cells computed? What is the implication for the mapped values?

   For example, if the grid values were computed by kriging, the map will be spatially smooth.

3. What is the reported accuracy and precision? Is information given only over the whole map or per-grid cell?
4. If uncertainty is given per-grid cell, which areas are mapped with sufficient precision for the intended use?

5. What was the data source for the computations?

   These sources should then all be checked recursively. Typically, these are point datasets but can also be polygon maps or even other grids.

   Note: For machine-learning models using the SCORPAN approach, it may be interesting to know about the covariates and their importance, but here we are evaluating the product, not the method.

6.4 Compilations

Some datasets are compilations of several diverse datasets.

1. What were the sources of the datasets used in the compilation?
2. How were the diverse datasets harmonized?

   For example, if soil textural classes were measured partly by sedimentation and partly by laser diffraction, a conversion formula must be used [e.g., 8], and similarly for SOC [e.g., 24].

7 Access

It must be possible to include the dataset in one's own workflow.

7.1 Repository

1. Where is the dataset stored and by whom?
2. Is it permitted to mirror the dataset for easier access?

7.2 Mechanics

1. How is the dataset accessed?

   For example, SoilGrids has a FAQ page describing all the access methods[11].
2. What file formats are available?
3. Do these refer to known standards?

   For example, GeoTIFF [16], NetCDF [23].
4. Is there computer code provided to access the dataset?

References

[1] American Society of Agronomy. Methods of Soil Analysis. https://acsess.onlinelibrary.wiley.com/hub/books/methods-soils, 2022. 5

[11] https://docs.isric.org/globaldata/soilgrids/SoilGrids_faqs_02.html

[2] NH Batjes, L Calisto, and LM de Sousa. Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). *Earth System Science Data*, 16(10): 4735–4765, October 2024. ISSN 1866-3508. doi: 10.5194/essd-16-4735-2024. 4

[3] Ruben T. Brouwer, Kim C.I. van Etten, Perry G.B. de Louw, Jakob Wallinga, and Julian Helfenstein. Modeling crop suitability for rewetting landscapes in the Netherlands across present and future climate scenarios. *Agricultural Water Management*, 325, 2026. ISSN 0378-3774. doi: 10.1016/j.agwat.2026.110190. 1

[4] Dick Brus. Spatial Sampling with R, April 2023. URL https://dickbrus.github.io/SpatialSamplingwithR/. 7

[5] M. R. Carter and E. G. Gregorich. *Soil Sampling and Methods of Analysis, Second Edition*. CRC Press, 2nd edition, 2007. ISBN 978-0-8493-3586-0. 5

[6] Science Committee. Specifications: Tiered GlobalSoilMap.net products; Release 2.4 [07/12/2015]. Appendix C: Correlations of soil properties derived from different soil analytical methods. Technical report, GlobalSoilMap.net, 2015. 7

[7] T P D'Avelo and R L McLeese. Why are those lines placed where they are?: An investigation of soil map recompilation methods. *Soil Survey Horizons*, 39:119–126, 1998. 4

[8] Giovani Stefani Faé, Felipe Montes, Ekaterina Bazilevskaya, Rodrigo Masip Añó, and Armen R. Kemanian. Making soil particle size analysis by laser diffraction compatible with standard soil texture determination methods. *Soil Science Society of America Journal*, 83(4):1244–1252, July 2019. ISSN 0361-5995. doi: 10.2136/sssaj2018.10.0385. 8

[9] P. A. Finke. Updating the (1 : 50,000) Dutch groundwater table class map by statistical methods: An analysis of quality versus cost. *Geoderma*, 97(3-4):329–350, 2000. 5

[10] Anatol Helfenstein, Vera L. Mulder, Mirjam J. D. Hack-ten Broeke, Maarten van Doorn, Kees Teuling, Dennis J. J. Walvoort, and Gerard B. M. Heuvelink. BIS-4D: Mapping soil properties and their uncertainties at 25 m resolution in the Netherlands. *Earth System Science Data*, 16(6):2941–2970, June 2024. ISSN 1866-3508. doi: 10.5194/essd-16-2941-2024. 1

[11] B Huang and SG Lu. 中国土系志: 云南 *Soil Series of China: Yunnan (in Chinese)*. Science Press, Beijing.(In Chinese), 2020. 7

[12] Bas Kempen, Dick J. Brus, Gerard B. M. Heuvelink, and Jetse J. Stoorvogel. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma*, 151(3-4):311–326, 2009. doi: 10.1016/j.geoderma.2009.04.023. 5

[13] Kenya Department of Agriculture. Soil survey of the East Konyango area : parts of Survey of Kenya, G.S.G.S., 4786 series, sheets 129/II, 129/IV, 130/I & 120/III. Map & report, Government Printer, Nairobi, 1961. 3

[14] Feng Liu, David G. Rossiter, Gan-Lin Zhang, and De-Cheng Li. A soil colour map of China. *Geoderma*, 379:114556, December 2020. ISSN 0016-7061. doi: 10.1016/j.geoderma.2020.114556. 3

[15] C. Mugnier. Grids and datums: Republic of Kenya. *Photogrammetric Engineering & Remote Sensing*, 69:593–597, 2003. 3

[16] Open Geospatial Consortium. OGC GeoTIFF. https://www.ogc.org/standards/geotiff/, February 2023. 8

[17] J. Padarian and A. B. McBratney. QuadMap: Variable resolution maps to better represent spatial uncertainty. *Computers & Geosciences*, 181: 105480, December 2023. ISSN 0098-3004. doi: 10.1016/j.cageo.2023.105480. 4

[18] Marc Pansu and Jacques Gautheyrou. *Handbook of Soil Analysis: Mineralogical, Organic and Inorganic Methods*. 2006. ISBN 978-3-540-31210-9. doi: 10.1007/978-3-540-31211-6. 5

[19] Laura Poggio, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, June 2021. ISSN 2199-3971. doi: 10.5194/soil-7-217-2021. 1

[20] David G. Rossiter, Jenni Dungait, Vera Leatitia Mulder, and Gerard B. M. Heuvelink. Editorial A new article type: The 'Data article'. *European Journal of Soil Science*, 73(3):e13265, 2022. ISSN 1365-2389. doi: 10.1111/ejss.13265. 1

[21] Soil Science Division Staff. *Soil Survey Manual*. Number 18 in USDA Handbook. Government Printing Office, Washington, DC, 2017. 6

[22] G. G. L. Steur and W. Heijink. *Bodemkaart van Nederland : Algemene Begrippen En Indelingen : Schaal 1 : 50.000*. Stichting voor Bodemkartering (STIBOKA), Wageningen, 1980. ISBN 90-220-0754-5. 6

[23] University Corporation for Atmospheric Research. NetCDF. https://www.unidata.ucar.edu/software/netcdf, 2026. 8

[24] A. Wotherspoon, R. P. Voroney, N. V. Thevathasan, and A. M. Gordon. Comparison of three methods for measurement of soil organic carbon. *Communications in Soil Science and Plant Analysis*, 46(sup1): 362–374, February 2015. ISSN 0010-3624. doi: 10.1080/00103624.2014.989111. 8

[25] 张甘霖 and 龚子同. 土壤调查实验室分析方法 – *Soil Survey Laboratory Methods*. 科学出版社, 北京, 2012. ISBN 978-7-03-032979-0. 5