SOIL Discuss., author comment AC7
https://doi.org/10.5194/soil-2021-80-AC7, 2022

Reply on RC2

David G. Rossiter et al.

Author comment on "How well does Digital Soil Mapping represent soil geography? An investigation from the USA" by David G. Rossiter et al., SOIL Discuss., https://doi.org/10.5194/soil-2021-80-AC7, 2022

RC2: 'Comment on soil-2021-80', Referee #2, Bradley Miller, 10 Feb 2022

We thank Prof. Miller for his careful review, including some fundamental questions to be addressed, as well as confusions to be cleared up.

1. Reply to summary and general comment:

1.1 "The spatial analysis techniques presented in this manuscript introduce important evaluations that need to be considered when comparing soil map products. However, the manuscript would benefit from a more thoughtful consideration of the meaning of the evaluations chosen. The metrics selected offer a variety of calculations, but it seems possible that they may be reflecting some of the same differences between the maps. Even after attempting to mute the effect of resolution, some patterns in the results remain. For example, some smoothing should probably be expected from models that – at their core – rely on regression fitting. So there appears to be some opportunity here to go beyond just reporting differences in the metrics. The metrics are intended to measure map characteristics other than relative smoothness. Can all the differences detected by the metrics be attributed to relative smoothness, and if not, what other map characteristics might these metrics be picking up on? I advocate for this because if others are to be convinced to apply these evaluations elsewhere, they will likely want to have some sense of how the results should be interpreted and/or what should be learned from them."

Thank you for this key observation. We admit to being unclear on why certain metrics show certain similiarities and differences, in some cases contrary to our expectations. We intend to add a section to the Conclusions. We will also review the description of each metric in \S3, where we have explained why the metric is used and what it might show. Note that we had already done this, especially for \S3.4.2 "Landscape metrics", but we will expand and clarify.

1.2 "The spatial analysis of map properties promoted in the present manuscript reminds me of the books “Soil and Landscape Analysis” by Hole and Campbell (1985) and “Pattern of the Soil Cover” by Fridland (1977). Although dated in their focus on analyzing polygons, these books suggested several ways additional information could be evaluated from the spatial patterns in the map. Then again, the methods proposed here essentially go back to a polygon type of analysis since they require the map to be classified."

Agreed. Those books were in fact the inspiration for this study. In a revision we intend to give some background to pattern analysis in soil survey, including these books, as part of the paragraph at L80ff. and perhaps again referred to in the Discussion.
Proposed text, after discussing how polygon soil survey is performed: "It has long been recognized that the soil cover forms patterns at various scales (Fridland 1974, Hole and Campbell 1985), and the traditional soil mapper attempts to find those patterns that are expressed at the map design scale." We prefer the more accessible 1974 Geoderma paper from Fridland to his book, which in any case is referred to in the paper.

1.3 "The way in which the mapping methods for the different products are described is sometimes unsupported and potentially not fully relevant to the spatial analysis applied. The distinctions between machine-learning models (categorized as PSM) and traditional soil mapping seem questionable. If the authors want to keep these assertions, then citations or better explanations for why they think they are true need to be added. A building concern in the description of traditional soil survey is a potential failure to recognize what the two approaches being compared have in common, which is important for understanding the differences in the respective products."

We think that there is a major difference between DSM, which proceeds from observations and covariates to a map, with traditional survey, which depends on expert interpretation of the soil-landscape, supported by purposive observations to support/reject/modify the surveyor's mental model. We bring out this contrast in the first paragraph of the paper.

These differences are indeed relevant to the analysis. Traditional methods stratify the landscape into polygons based on expert judgement and landscape analysis, whereas DSM methods do not stratify as such, they predict per-grid cell, using covariates that are meant to represent soil forming factors.

As for the mapping methods of the different products, we don't see how our descriptions are "unsupported". We cite the original papers for all three DSM products, and the description of the NRCS products is directly from their descriptions by the NRCS, and by the two co-authors who have worked in the NRCS soil survey.

It is difficult to more fully answer this comment without reference to specific texts that the reviewer thinks are deficient. However, we will review our explanations for clarity.

1.4 "To improve the manuscript, I would encourage the authors to consider how the strategies of the different mapping approaches may be connected to the results of the spatial evaluation metrics applied. As an example, if we were to recognize traditional soil survey as a predictive map product, then the covariates would largely be what the mapper could see in the available airphoto. Although the airphoto bases used for traditional soil survey leave a lot to be desired compared to modern covariates, one can see more detail of landform shapes than with a 100m digital terrain data set. Although resolution is already partially alluded to in the text, this kind of context may help sort through what may be driving some of the differences detected with the various map evaluation metrics."

This is a good point. We did not discuss traditional approaches in any detail, as our aim was to compare their results with DSM products, and then within these. The sentence about traditional methods mention above does suggest what the surveyor is looking at. We will expand on this in the discussion.

1.5 "While I offer considerable critique on the characterization of traditional soil survey and push for some consideration about what the selected metrics really describe about the maps being evaluated, I applaud the spatial analysis approach to evaluating different map products. Thinking about how to evaluate maps beyond the prediction of points is an important contribution to the advancement of soil mapping as a science."

We appreciate the reviewer's encouragement, and hope to have dealt with the above main issue, and the general points and details below. Indeed "how to evaluate maps beyond the prediction of points" was our main motivation for this work.

General Comments

2. "Please consider being more consistent with terminology and abbreviations. For example, traditional soil survey versus conventional soil survey and map scale versus design scale. They appear to be used interchangeably, but the change in term makes the reader wonder if there is a difference being implied and if so, what that difference may be. Regarding terminology, I disagree with several of the terms used in the present manuscript but attempt to use the terms most frequently used by the authors in the present manuscript to facilitate communication in this review of the manuscript."

Indeed we were sloppy here.
(1) We will use "traditional" survey throughout, and explained what we mean by that. Immediately after the first mention of DSM we intend to add "This [DSM] is in contrast to what we here call ``traditional'' soil survey, in which the soil surveyor develops a mental model of the soil geography (Hudson 1992) by interpreting the landscape with the aid of airphotos, purposive transects, and detailed profile descriptions at locations thought to represent the central concepts of the soil classes present in the study area" and here citing the Soil Survey Manual. We then will replace the "conventional soil maps" with "polygon maps made by traditional soil survey methods".
(2) "I disagree with several of the terms used in the present manuscript" we would like to know which and why.

3. "Information presented in sections 1 through 2.3 sometimes circles back on itself. Consider reorganizing to avoid repeating some information. I think this would also help readers understand and compare the processes by which these maps are made along with the resulting map characteristics."

Indeed there is some redundancy here, which we intend to edit out.

Section 1 has no information on the specific products, only a discussion of the overall problem and approach. Then in Section 2 we broke down the differences between products into (1) nature of the maps to compare, (2) the primary data on which they are based, (3) the environmental covariates used in DSM. We intend to remove the heading of (1), as it follows directly from the section heading. (2) can also be folded into the earlier description of the NRCS products. This section will be tightened and some information moved to \S2.4 "Mapping Methods". This should make the differences between products easier to understand.

4."Most figure captions are more like titles. Please elaborate in the figure captions to guide the reader in what to look for in the figures."

Our practice has been to point out the most important features of a figure in the text, at the point the figure is referenced. In some cases this text would be too long to fit into a caption. However we will expand the figure texts to make them independently understandable. For some figures the discussion in the text can not easily be summarized for a figure caption. In these cases we will add "See text for explanation".

Specific Comments

L53-55 – "The acknowledgment of the Scull et al (2003) paper using the term ‘predictive soil mapping’ is appreciated. However, I’m concerned that the use of this term could lead to confusion over what differentiates traditional soil survey from the new approaches that utilize computational algorithms to produce soil maps. The potential issue here is the perception that traditional soil survey is somehow not predictive. Of course, it is not possible to observe soil everywhere, which requires some level of spatial prediction for any kind of soil map with exhaustive coverage."

Yes, this was pointed out by RC1 and some CC, and is correct. "PSM" will be replaced by "DSM" throughout. Scull will still be cited and the previous term "PSM" acknowledged.

L62 – "How fair is it to say that machine-learning models can be implemented with fewer locations visited when the machine-learning models presented for comparison in this manuscript rely on a database of observations collected by the activity of traditional soil mapping? In statistical theory, it makes sense that some optimized sampling design should be able to capture the needed information. But has this been the experience of soil mapping?"

Here we are not saying that the maps presented here depend on fewer points. In this general introduction about the attractions of DSM, one of them is indeed that with clever sampling in covariate space -- if the relation with covariates is strong (as stated in this sentence) -- fewer points can be used than if the standards for sampling density at various map scales are followed. However, as the reviewer points out, this has not, to our knowledge, been studied as a direct comparison between traditional and digital soil mapping. This is partly because traditional maps are quite difficult to evaluate quantitatively, unlike DSM which at least can do some cross-validation.

So we will soften the statement to:

"Further, it may be that fewer locations can be visited in order to develop reliable models, as compared to traditional survey techniques. If the relation with covariates is strong, and locations representative of the entire covariate feature space are included in the training set, it may be possible to map large areas from relatively few field observations."

We also will refer here (as in L190) to the "homosoil" concept: identical environmental conditions (as represented by covariates) should result in the same soils.

L64 – "Please explain how PSM has a greater ability to map inaccessible areas than traditional methods."

DSM can extrapolate based on similarity in covariates, again the "homosoil" concept, with the caution (as we write) "if the available training data cover the covariate space of the inaccessible area." This is essential the same idea as the "Konzeptbodenkarten" (conceptual soil maps) long used in German practice to map an area from covariates (in the past, polygon maps of various themes such as geology and relief) prior to fieldwork. As an early and thorough example see Grenzius, R.: Konzeptbodenkarten für den städlischen Raum, Z. Pflanzen. Bodenk., 156, 209–212, https://doi.org/10.1002/jpln.19931560304, 1993, where maps of topography, historical land use, biotypes, flooding, climate, etc. are used to make a concept map. This could be called a "pre-map". But in the case of concept maps there is not yet soil information, this will come from an appropriately stratified survey. In the case of DSM extrapolation, the soil information comes from the training area and is extrapolated based on the covariates and machine-learning model.

We could add an explanation here, but this introduction is not meant to be a full exposition of DSM.

L73 – "The assumption that mapping scale drives resolution is largely a concept held over from paper maps. In many disciplines (e.g., geology) we are seeing newer generations of maps adding detail without changing the extent of the map. Adding those details (making the resolution finer) has become functionally possible because the producer and the user can ‘zoom’ in and out of the map in a GIS. This technological development renders the question of mapping scale nearly mute. However, the remaining question is if there is sufficient information to support the finer resolution."

We disagree. Mapping scale and resolution are concepts from two different concepts of space (polygon and grid) but are related by the spatial detail of information. Scale is not extent, it is the ratio of the physical map to real world, e.g., 1:50 000. The paper map can of course be reproduced with a larger piece of paper but this does not add information. So the design scale of a polygon product, based on the concept of minimum legible delineation (Forbes, T. R., Rossiter, D., and Van Wambeke, A.: Guidelines for evaluating the adequacy of soil resource inventories, Cornell University Department of Agronomy, Ithaca, NY, 1982) is related to the density of information. Tom Hengl (among others) has related the two concepts, see Hengl, T.: Finding the right pixel size, 32, 1283–1298, https://doi.org/10.1016/j.cageo.2005.11.008, 2006.

We disagree that the ability to zoom in renders the question of scale moot ("having little or no practical relevance", often because in law the situation that gave rise to a proceding is no longer relevant). If the information density is not sufficient there is nothing more to see. We discuss the selection of a proper pixel size in the context of DSM in the SOIL paper presenting SoilGrids v2.0 (Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021)

L80 – "This paragraph appears to present a non sequitur. The argument is made that point evaluations of PSM do not consider the spatial pattern of predictions. However, it is not clear to me how the subsequent information presented about traditional soil survey methods shows how that approach does more to evaluate spatial patterns of predictions."

This observation is correct. We were contrasting with the traditional approach, and implicitly suggesting that map users would compare the pattern of polygons (map unit delineations) with the landscape, and also that the pattern of polygons is immediately visible to the map user, as opposed to the pixelated results of DSM. We intend to re-write this paragraph as:

"A more serious issue is that point evaluations of DSM products do not consider the spatial pattern of predictions. By contrast, traditional soil surveys produce polygon maps of relatively homogeneous pedons, with the boundary lines placed at inflection points of maximum change between soil bodies (Lagacherie 1996). These maps explicitly show the surveyor's interpretation of the soil landscape as developed from a mental model of the soil-forming processes, and which when viewed as a whole show the pattern of the soil cover. It has long been recognized that the soil cover forms patterns at various scales (Fridland 1974, Hole and Campbell 1985}, so that the traditional soil mapper attempts to find those patterns expressed at the map design scale. Since DSM predictions are on a grid cell basis, there is no concept of relatively homogeneous natural soil bodies nor inflection points between them. However, it might be expected that if the valus of the DSM covariates representing the soil-forming factors also cluster in a similar pattern to the soil cover, the DSM predictions would also cluster and approximate map units from traditional survey, despite being predicted separately per grid cell. The question is thus to what degree DSM products represent the actual soil landscape spatial pattern and, more importantly, the underlying pedogenetic and geomorphic processes."

We think this explains the difference between traditional and DSM methods of pattern formation.

L83-88 – "This section appears to be building an assumption that traditional soil survey does not include any kind of model that uses input data to make predictions. Although the authors point out real shortcomings in the “paradigm” of traditional soil survey described by Hudson (1992), they have left out how that approach uses ‘mental models’. This omission obscures what machine-learning and traditional soil survey methods have in common."

See the previous comment (L80); we think the proposed re-write brings out this point.

L91-95 – "The cartographic reasons that traditional soil survey is constrained to the levels of detail it has is a useful explanation here. However, for translating polygons delineated in USA soil survey maps, there is a more direct approach. The USA soil survey program has a strict protocol for minimum delineation areas. The USA’s “Field Book for Describing and Sampling Soils” specifies minimum-size delineations of 0.6 ha for 1:12,000, 1.0 ha for 1:15,840, and 2.3 ha for 1:24,000. Now, we should note that those are minimum delineation sizes, and the mean delineation size will be larger than that. The mean delineation size can vary by landscape and by the style of the mapper."

It's clear that these are taken directly from the Forbes et al. criteria, probably because Richard Arnold was at Cornell when they were being developed, and then went on to head the NRCS Soil Survey. We will mention that these are in the Field Book and cite it.

L126 – "The use of STATSGO2 is interesting here. With the possible exception of some areas where only an order 5 map has been made, STATSGO is a purposefully generalized map product that is aggregated by expert knowledge. At first, I questioned if it made sense to include STATSGO in this evaluation, but then STATSGO was not evaluated. Considering STATSGO is not evaluated in this manuscript, it does not seem relevant."

STATSGO is used to fill in areas without more detailed survey. This is explained in the gNATSGO description and in L130.

L128 –" “State” should not be capitalized."

Will be corrected, also in several other locations where this is not part of a proper noun.

L162-169 –" This is a nice, succinct description of the state of SSURGO."

Thank you.

L175-183 – "I think it may be misleading to state that SG2 does not use any information from SSURGO when SG2 uses profiles from the NRCS pedon database. Yes, SG2 is not using SSURGO itself, but they are both using a set of training points that they have in common for the USA area. This overlap in source information is even more so for the SPCG, which makes use of additional pedons that were produced from the activities of the USA Soil Survey. In the case of the RaCA dataset, it is new enough that it probably has not strongly influenced the SSURGO map. Nevertheless, the role of these training points in all the map products should be explained clearly. Specifically, I disagree with the idea that SSURGO is independent of the data points managed by the NRCS."

This was poorly-worded. We meant that SSURGO polygons were not used. Indeed SG2 uses NRCS profiles, which were also (indirectly) used to build SSURGO. We will make it clear that SSURGO is a map product, and does not include the points.
From

https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627: "SSURGO datasets consist of map data, tabular data, and information about how the maps and tables were created." This only applies to the polygons. So we will add "any information derived from SSURGO polygons ..." We also will add that SPCG uses SSURGO map units to derive parent material and drainage classes.

L199 – "Add space after the period."

Will do. How did that slip through the editing process?

L247-252 – "Libohova et al. (2014) explored the validity of these ranges to represent uncertainty. That evaluation seems relevant here."

Thank you for reminding us to include this reference: Libohova, Z., Wills, S., and Odgers, N. P.: Legacy data quality and uncertainty estimation for United States GlobalSoilMap products, in: GlobalSoilMap: Basis of the Global Spatial Soil Information System, Boca Raton, 63–68, 2014.

L267 – "Consider a rubric here to define how the expert judgement will be evaluating the maps. This will help the reader understand the value system being applied in this evaluation and communicate a more structured approach to how the qualitative comparison will be made."

Each expert will see things differently, but there are some commonalities. We will explain a general approach ("rubric"). Proposed text:

"The DSM product can be evaluated at selected known points, typically from field observation of test areas: is the ``correct'' soil type or property predicted? and if not, is the error a reasonable approximation? More interesting for our purposes are patterns in the DSM product. These can be compared to patterns used in the mental model of traditional soil survey, for example, toposequences and sequences of parent material outcrops. In both cases (points and patterns) the evaluator may be able to infer which DSM covariates would be needed to improve the map."

L283 – "Where does variability come from for any single point in these maps. Won't there be a single value for a grid cell, or is there something else being brought in here? Is this using the uncertainty ranges? In any case, please explain clearly to help the reader know the basis for the proportional nugget."

There is one value per grid cell. But when these are plotted on an empirical variogram the smooth function fit to the increasing separations in geographic space vs. these in feature space often does not intercept the semivariance axis at 0, when the separation is 0. This is attributed to many causes, e.g., lab error or variability within the grid cell (which is represented only by its centre). This is a common phenomenon in variography for soil data. L283-4 "The proportional nugget shows the inherent variability at a point, at a scale shorter than the grid spacing" is correct, we will modfify the sentence to explain the possible causes of this. Because of smoothing there is generally low or no nugget variance in DSM products. We intend to change this sentence to "The proportional nugget shows the variability at the prediction point at the centre of a grid cell, at a scale shorter than the grid spacing."

L383-L388 – "This paragraph drifts into results by beginning the evaluation. Recommend keeping the description of methods separate from the results found by implementing them."

Agreed. We had this figure here to illustrate the method, but indeed these are results, and will be moved to a first subsection of the "Regional spatial patterns" results, with some more explanation that is only a preliminary evaluation.

Table 1 – "Please be consistent in abbreviations."

Will be changed to match the use of abbreviations in the text.

L408 – "Change ‘distributions’ to ‘distribution’?"

Yes.

L429 – "Add missing ‘s’ after ‘PSP’"

Yes

Figure 3 – "Consider including the r values in the boxes."

Table 2 – "Again, please be consistent in abbreviations, both for matching abbreviations used in the text and previous tables."

Will be changed to match the use of abbreviations in the text. The entire document will be checked for consistency.

L446 – "add ‘s’ at end of ‘indicate’"

Figure 4 – "The ‘SoilGrids’ and ‘SPCG100’ maps show large areas of pH lower than shown in the ‘gNATSGO’ in the high elevation portions of the east and south areas. Could this be a case of extrapolation in the feature space? If so, how might this be reflected in the evaluation metrics presented in the is manuscript?"

Random forests do not extrapolate beyond their training values, so this can not be the cause. In this case it is likely that the higher elevations in the Catskills (to the southeast of this study area), which have similar forest land cover, have biased these predictions. In fact, the NASIS database (incorporated into WoSIS) does have several points in these higher hills, and these show the lower pH (around 5.0) of the topsoils, even in forest, as opposed to the pH 4.5 predictions from SG2 and SPCG. A local model covering only this tile might well have results that are more consistent with gNATSGO.

L447-448 – "The authors suggest that the smoothing effect was caused by the spatial continuity of the covariates, which seems reasonable considering the resolution. It seems to me that any kind of fitted model is, almost by design, going to smooth out some patterns in the training data. Would the authors mind commenting on this possible additional factor?"

Good point. We will comment on this. Re-thinking the original comment, not all covariates are smooth (e.g., vegetation indices can vary greatly pixel-to-pixel if land use varies at the scale of the grid cell. There is more we will say about these results. Proposed text (which will be reconsidered and expanded):

"gNATSGO has the shortest effective range. This indicates fine-scale structure at 250 m resolution, which is of the same order as the minimum legible delineation (MLD) as a grid cell (see Introduction). The mappers who defined the boundaries between soil classes (and thus representative property values) were able divide the landscape at this high spatial frequency, if appropriate to the soil pattern. The DSM products have longer ranges, likely due to the longer-range spatial continuity in many of the covariates. PSP has a longer range and lower sill than the gNATSGO from which it is derived, due to the harmonization inherent in the DSMART algorithm. It has the highest proportional nugget, due to DSMART randomly assigning pixels within a gNATSGO map unit to its constituents, so that neighbouring pixels may be contrasting at the shortest separation. The very low proportional nuggets of the other products are due to the coarse resolution."

Figure 10 – "Did the ‘SPCG100USA’ semivariogram actually reach a sill?"

In an exponential variogram model the sill is never reached; by convention 95% of the fitted sill parameter is considered the effective range. The total sill is 11.82 (Table 2), at the presented cutoff the variogram reaches about 11, which is 93% of the total sill. So not quite 95% but close enough. It's clear from the nice fit to the exponential model form that the variogram is asymptotic to a sill and not increasing without bounds.

L462-463 – "These sentences are a little unclear; please consider rewording and/or expanding upon the explanation."

Yes, the homogeneity and completeness are not intuitive. They were explained in \S3.4.1 "V-measure", which will be referenced here. However the caption and explanation were in fact incorrect, we missed that the sense of these maps is reversed from that of the accompanying table.

Also while reviewing this section, we decided that it would be informative to also show the relation between SG2 and SPCG, because these are much closer in method, and in fact show much more similar patterns, which is revealed by these maps. We will discuss these differences, pointing out the reason for them.

The proposed revised text for the first comparison is:

"Figure XX shows the inhomogeneity and incompleteness of the SG2 pH class map (the second map for the V-measure), with respect to the gNATSGO pH class map (the reference map). These values are the inverse of the composite values of Table XX: the very low values in the table correspond to high values in the figure. In the homogeneity map, the blue polygons are the most homogeneous areas of the SG2 map, i.e., where an SG2 polygon has the most homogeneous set of gNATSGO classified values and thus comes closest to the reference. In the completeness map, the blue polygons are the most complete areas of the SG2 map, i.e., where the gNATSGO reference map has the most homogeneous set of SG2 classified values. The two maps have no areas with similar patterns."

The second comparison will be similarly discussed and then contrasted with the first comparison.

In addition \S3.4.1 will be expanded and better explained.

L468-469 – "If the difference between gNATSGO and PSP is going to be called out, maybe it is worth mentioning that the difference between gNATSGO and SPCG is even more. Some discussion about why the difference between gNATSGO and PSP captured the authors' attention may also be warranted."

L473-474 – "Just to be clear, which PSM product is being referred to here?"

PSP. This was not clear. "This" will be replacd by "PSP".

L478 – "This is the first mention of silt concentration! This switch makes for a mismatch between the methods described and the results presented."

Indeed. This section is meant to show how PSP, which claims to disaggreated gSSURGO, is not always (usually?) successful. In this study area the topsoil pH does not differ too much in local landscapes (although it does have a wide range across a 1x1 degree tile), so is not a suitable example of what we want to show. We explain this as "Here we use silt concentration, as it reveals stronger qualitative discrepancies than pH in this test area." We will add paranthetically "than pH (used in the previous sections of this paper)..."

L480-482 – "This content would be better suited in the figure caption."

Correct. Will be moved.