How well does Predictive Soil Mapping represent soil geography? An investigation from the USA

David G. Rossiter[1,2], Laura Poggio[1], Dylan Beaudette[3], and Zamir Libohova[4]

[1]ISRIC-World Soil Information, Postbus 353, Wageningen 6700 AJ, NL
[2]Section of Soil & Crop Sciences, New York State College of Agriculture and Life Sciences, 233 Emerson Hall, Cornell University, Ithaca NY 14853 USA
[3]USDA-NRCS Soil Survey Division, 19777 Greenely Rd., Sonora, CA, 95370 USA
[4]USDA-ARS, Dale Bumpers Small Farms Research Center, 6883 South State Hwy 23, Booneville, AR 72927 USA

Correspondence: David G. Rossiter (david.rossiter@isric.org)

Abstract. We present methods to evaluate the spatial patterns of the geographic distribution of soil properties in the USA, as shown in gridded maps produced by Predictive Soil Mapping (PSM) at global (SoilGrids v2), national (Soil Properties and Class 100m Grids of the USA), and regional (POLARIS soil properties) scales, and compare them to spatial patterns known from detailed field surveys (gSSURGO). The methods are illustrated with an example: topsoil pH for an area in central

5 New York State. A companion report examines other areas, soil properties, and depth slices. A set of R Markdown scripts is referenced so that readers can apply the analysis for areas of their interest. For the test case we discover and discuss substantial discrepancies between PSM products, as well as large differences between the PSM products and legacy field surveys. These differences are in whole-map statistics, visually-identifiable landscape features, level of detail, range and strength of spatial autocorrelation, landscape metrics (Shannon diversity and evenness, shape, aggregation, mean fractal dimension, co-

10 occurence vectors), and spatial patterns of property maps classified by histogram equalization. Histograms and variogram analysis revealed the smoothing effect of machine-learning models. Property class maps made by histogram equalization were substantially different, but there was no consistent trend in their landscape metrics. The model using only national points and covariates was not better than the global model, and in some cases introduced artefacts from a lithology covariate. Uncertainty (5–95% confidence intervals) provided by SoilGrids and POLARIS were unrealistically wide compared to gSSURGO low and

15 high estimated values and show substantially different spatial patterns. We discuss the potential use of the PSM products as a (partial) replacement for field-based soil surveys.

Contents

1 Introduction

Predictive Soil Mapping (PSM), also commonly referred to as Digital Soil Mapping (DSM), has been defined as "the development of a numerical or statistical model of the relationship among environmental variables and soil properties, which is then applied to a geographic data base to create a predictive map" (Scull et al., 2003). Since the seminal paper of McBratney et al. (2003), recently reviewed by Minasny and McBratney (2016), PSM has been widely-applied from the field to global levels.

A principal attraction of PSM is that it produces consistent, geometrically-correct and reproducible gridded maps over large areas, given training data ("point" observations of soil classes, properties or conditions), a set of environmental covariates covering the entire area to be mapped at some fixed grid resolution, and a set of algorithms implemented in computer code. This removes the need for expertise in discovering and interpreting the soil-landscape relations, also known as the "paradigm" of soil survey (Hudson, 1992), which is vital for traditional soil survey and difficult to acquire and harmonize among surveyors.

Further, 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, large areas can be mapped from relatively few field observations. Maps made by PSM can include areas that are not accessible to field mappers because of permissions or difficult access, if the available training data cover the covariate space of the inaccessible area. However, PSM requires sufficient sampling density to cover the full covariate space, since most PSM methods do not extrapolate, and in any case extrapolation is inadvisable.

It is not claimed that PSM is inherently superior to traditional survey, but well-known problems of traditional survey are avoided: multiple survey projects over time with inconsistent standards and mapping concepts, inconsistency among mappers, difficulties in objectively identifying boundaries, and indeed the need to identify boundaries. However, traditional soil surveyors and users of their maps are often critical of PSM products, and may not understand how they were made and how they should be used (Arrouays et al., 2020). In the USA there is increasing awareness of, and interest in, PSM products. Here the most important point of contention has to do with PSM resolution (pixel size), which implies a mapping scale, compared to the scale at which differences can be reliably interpreted for user needs. Criticism of PSM products is proportional to the degree to which their implied spatial precision and accuracy is over-sold.

The success of PSM in reproducing known "point" observations is typically reported by evaluation ("validation") statistics based on data splitting (independent evaluation) or by cross-validation. These evaluations are almost never based on random sampling (Brus et al., 2011), and since the source point datasets are almost always biased towards certain land uses, access constraints or landscape locations, these evaluations carry forward these biases and must be interpreted with caution.

A more serious issue is that point evaluations of PSM products do not consider the spatial pattern of predictions. By contrast, soil surveys based on extensive field observation of the soil landscape (sometimes called the "traditional" soil survey method) produce polygon maps of soil map units (SMU) composed of one or more soil type units (STU), e.g., soil series, which then provide modal or representative profiles with measurements or estimates of soil properties, usually by genetic horizon. These explicitly show the soil landscape, as interpreted by the soil surveyor and as managed by the land user. Since PSM maps are not derived from a paradigm such as the surveyor's explicit stratification of the soil landscape, their spatial pattern depends

on the input data (training points and covariates) and the PSM algorithm applied. The question is thus to what degree PSM products represent the actual soil landscape spatial pattern and, more importantly, the underlying pedogenetic and geomorphic processes.

PSM maps are most commonly produced at grid cell resolutions from 1 km to 30 m, and even to 10 m for precision
90 agriculture applications. Environmental covariates are available at these resolution, so that PSM products at high resolutions can show fine details that can not be presented at the design scale of conventional maps. These have minimum legible delineations (MLD) of 0.25 cm^2 (Vink, 1975) or 0.40 cm^2 (Forbes et al., 1982) on the published map, multiplied by the scale factor. For example, a polygon map at 1:24 000, typical of USA conventional soil survey, can represent spatial patterns of 1.44 (Vink) to 2.3 (Forbes) ha minimum-size polygons. These correspond to single grid cell resolutions of 240 to 384 m, coarser than
95 higher-resolution PSM products from (30 to 100 m). But the question remains whether this implied fine detail represents true differences or artefacts of the mapping process – in other words, should the PSM map unit trust the apparent differences between adjacent grid cells, or are some or most of these differences due to artefacts ("noise") of the PSM process? Further, there is the question of how well the medium-resolution products (e.g., 250 m) represent the soil landscape at regional extent.

The objective of this study is to present methods with which to evaluate the landscape and detailed level spatial patterns of
100 PSM maps. These maps may have been developed for global, national, or regional spatial extents. These patterns are compared with digital soil maps based on polygon maps produced by field survey and expert soil-landscape analysis. We chose the USA as a study area because of the availability of field-based soil surveys at 1:12 000 to 1:24 000 design scale, linked to detailed descriptions of modal soil profiles, available as a seamless digital product. These comparisons may be useful in the context of current plans (Thompson et al., 2020) for updating and completing the USA soil survey using PSM methods and
105 GlobalSoilMap (GSM) specifications (Arrouays et al., 2014). They should also be useful for developing realistic expectations for what PSM can and cannot deliver (Arrouays et al., 2020).

To evaluate PSM methods we apply them to selected test areas and soil properties, and comment on the results. This paper introduces the methods and data sources, and includes an illustrative example (one area, one soil property, one depth slice), with a minimum of discussion of soil geography. A companion ISRIC Report (Rossiter et al., 2021) presents four case studies
110 in diverse soil geographic contexts, each with different soil properties and depth slices. We encourage readers to apply the methods to their own study areas within the USA and to their soil properties of interest, to evaluate the utility of the several PSM products. For this, we provide our analysis scripts as R Markdown documents (R Studio, 2020); see "Code availability" at the end of this paper.

2 Data sources

115 The source maps are of three kinds: (1) digital products based on field survey without any statistical modelling; (2) PSM products based on field survey products and enhanced by statistical modelling using environmental covariates; and (3) PSM products based on statistical modelling using training points and environmental covariates. This latter is the most common PSM method worldwide, especially for areas without extensive field surveys. These differ in their primary data, their environmental

covariates and geographic scope, their mapping methods, the resolution of their predictions, and their uncertainty assessment.
120 These have implications for interpreting their relative success. We summarize these below; see the Supplementary Materials
and corresponding papers for details.

The first source is represented by the reference product from the Natural Resources Conservation Service (NRCS) of the
United States Department of Agriculture (USDA), based on extensive field survey and considered to be the most accurate
information, despite the occasional presence of artefacts from the overall mapping program, as explained later in this section.
125 The National Soil Geographic Database **gNATSGO** is a composite of the Soil Survey Geographic Database (SSURGO, mostly
1:24,000 scale), and State Soil Geographic Database (STATSGO2, 1:250,000 scale), and the very detailed Raster Soil Survey
Database (RSS), according to the most detailed product available for all areas of the USA (NRCS Soils, 2020a). For each
State or equivalent political unit, the SSURGO and STATSGO2 polygon maps of SMU produced by field survey have been
rasterized to a grid, each cell keyed to a a soil map unit (SMU) (NRCS Soils, 2020b). Grid cells link to the best available
130 (i.e., greatest detail STATSGO-SSURGO-RSS) SMU. The digital products are delivered at 30 m resolution for the 48 States
of the continental USA and the District of Columbia (abbreviated CONUS), i.e., excluding Alaska, Hawaii and the Island
Territories. The SMU are mappable landscape elements, at the survey design scale. These usually have multiple component
STU, with reported estimated proportion and sometimes the landscape relations within the SMU. However the locations of
the STU within the SMU are not mapped due to the design scale. The STU are linked to database tables of representative or
135 synthetic soil profiles, with field and laboratory measurements of multiple soil properties, as well as interpretations for soil use.
To obtain values for soil properties in a grid cell, properties of the components of the corresponding SMU are combined by
area-weighted averaging. To obtain values at coarser resolution, grid cells are averaged by upscaling.

The second source is represented by **POLARIS soil properties** (Chaney et al., 2019) (further **PSP**), the result of harmonizing
diverse SSURGO and STATSGO2 polygon data with the DSMART algorithm (Odgers et al., 2014) to produce a probabilistic
140 raster soil class or component map (30 m grid resolution) and then extracting property information from gSSURGO, aggregated
by component name. Despite the source data, this is not an NRCS product.

There are two products representing the third source, one for the world and one for the continental USA only. This allows us
to compare globally- and nationally-consistent products. The global product is **SoilGrids v2.0** (further **SG2**) (ISRIC - World
Soil Information, 2020; Poggio et al., 2021), a further development of SoilGrids1km (Hengl et al., 2014) and SoilGrids250m
145 (Hengl et al., 2017). This uses a global point dataset and environmental covariates that cover the entire world (except the
high Arctic and Antarctica), and models that are globally-consistent. The continental product is the **Soil Properties and Class
100m Grids of the United States** (further **SPCG**) (Ramcharan et al., 2018), which followed the methodology of Hengl et al.
(2017) with the addition of USA-specific covariates, notably parent material and drainage classes extracted from SSURGO or
STATSGO2, and only used the CONUS extent of environmental covariates in model building.

150 **2.1 Primary data**

A major difference between products is the extent to which primary data relies on field soil survey.

At one extreme are the hand-drawn polygons of SSURGO, the basis of the gridded products gSSURGO and gNATSGO. These polygons are available from the NRCS as vector GIS layers (Natural Resources Conservation Service, 2019), and in a convenient format on a geographic background as SoilWeb (California Soil Resource Lab, 2020). These are a representation
155 of those delineated by the field surveyors on stereo-pairs or ortho-photos and subsequently converted to vector digital format. Soil surveys conducted in the last 15 years were compiled using on-screen digitization in a GIS. These polygons are organized in map units (SMU) with one or more components (STU), usually named for a soil series but more specific than the parent soil series concept. Taxa above the soil series (family or subgroup) are commonly used in soil surveys of national forest land or wilderness areas. Soil series are the lowest level of Soil Taxonomy (Soil Survey Division Staff, 2014) and are described
160 in the Official Series Descriptions (OSD), as modal profiles with a set of ranges for the observed morphology and laboratory measurements.

The component STU in a mapped SMU vary in the observed field properties from the OSD modal description, but usually fit within soil series range. The observed field properties of soil component units are utilized for developing a set of interpretations for SSURGO polygon map units. Further, since field surveys were carried out over a long time period, series names and
165 mapping concepts may differ between adjacent survey areas (a.k.a. survey "vintage"). At boundaries between survey areas, polygon lines at survey limits have been matched during digitizing (D'Avelo and McLeese, 1998). SSURGO SMU delineations and linked tabular data represent a progressive data collection and correlation effort spanning nearly 100 years. As such, there exist many soil survey vintages, each a snapshot in time, tied to specific land-use assumptions and technological limitations. Systematic, continuous updates to the entire SSURGO database have been made since 2013.

170 However, the transfer from unrectified photos to topographic base and the edge matching between survey areas has not always been flawless, and in addition polygons may have been mis-drawn on the original survey (Supplementary Fig. 1). Thus we can not take gSSURGO as a completely reliable reference.

PSP does not use any point observations; rather, it samples pseudo-points from gSSURGO and uses these as training points for the DSMART disaggregation algorithm (see below).

175 At the other extreme is SG2, which does not use any information derived from SSURGO or STATSGO. Its training points are extracted from the freely-shareable World Soil Information Service (WoSIS) point dataset from ISRIC-World Soil Information (Batjes et al., 2020). These include all profiles in the NRCS pedon database. The WoSIS points are augmented by several datasets that can not be published externally due to restrictions by the original data providers to ISRIC, but which can be used in mapping. In total $\approx 240\,000$ profiles were used in model building.

180 SPCG is similar to SG2 in that it is primarily based on point observations, but it has a richer source of these than SG2: the NCSS Characterization Database (34 183 pedons comprising 213 499 horizons), the National Soil Information System (NASIS), and the Rapid Carbon Assessment (RaCA) dataset (31 215 pedons); this latter only for organic C, total N, and bulk density.

2.2 Environmental Covariates and geographic scope

185 The three PSM products (SG2, PSP, and SPCG) use a large number of gridded GIS coverages as environmental covariates in their predictive models. These represent soil-forming factors, and include climate, ecology, geology, land use/cover, terrain, vegetation and hydrography; see Supplementary Information for details.

PSP also uses coarse-resolution (≈ 2 km) estimates of U, Th, and K γ-ray decay products. The model is trained in overlapping tiles, thus each tile is derived from a local model.

190 SG2 only uses environmental covariates available over the whole world. The model is trained for the whole world, not per-country or region, thus it is a global model. This corresponds to the "homosoil" concept (Mallavan et al., 2010): identical environmental conditions anywhere in the world should result in the same soils. The obvious question is whether or not the additional information from outside the CONUS leads to an improved model for this region.

SPCG also uses SSURGO and STATSGO2 polygons to derive parent material (87) and drainage (4) classes as environmental covariates. The model is trained for the CONUS, thus it is a regional model.

2.3 Mapping methods

gSSURGO is based on field survey, mostly on unrectified airphoto bases. The many individual survey areas have been partially homogenized during a process of digitization and recompilation onto topographic or orthophoto bases during the 1990's (D'Avelo and McLeese, 1998) and are provided as the polygon SSURGO map.Field methods are described in successive

200 editions of the Soil Survey Manual (Soil Survey Division Staff, 2017) and the field book for describing and sampling soils (Schoeneberger et al., 2012). Mapping is based on conceptual models of soil-landscape relations developed in each survey area (Hudson, 1992), confirmed by purposive auger observations and a small number of full profile descriptions to characterize map unit composition. SSURGO is progressively updated by field inspection and correlation, as problems are identified by soil surveyors or map users. The SSURGO polygons are rasterized to gSSURGO. Since SSURGO is compiled from diverse field

205 surveys over many years, in some areas there are artefacts of that survey process (Supplementary Fig. 2).

PSP uses the DSMART disaggregation algorithm (Odgers et al., 2014) to predict the most probable component (STU), along with their probability of occurrence, at each 30 m resolution grid cell, and from the modal soil properties of the component, a probability-weighted aggregation. In this context "disaggregation" is the process of attempting to take a coarser-resolution gridded or smaller-scale polygon product which is known to have multiple STU, and identify the locations at a finer grid

210 resolution where these components would be found, should the original survey have been at larger scale. This depends on fine-scale covariates that, in theory, relate to the STU within an SMU. It attempts to deal with the problems caused by multiple surveys over time, inconsistencies among mappers, and poor georeference of SMU boundaries by sampling out of mapped SMU polygons according to declared proportions of map unit components (STU) and using these as pseudo-observations to train PSM models of STU occurrence. PSP provides a fine-scale map equivalent to $\approx$1:3 000 design scale, i.e., from 16 to 64

215 times finer resolution than the original 1:12 000 to 1:24 000 surveys included in SSURGO. An obvious question is whether it

is possible to map at this resolution from the SSURGO source, even with the fine-resolution covariates used by DSMART. See Chaney et al. (2019) for details.

The other two methods are representative of the dominant PSM method as implemented, with some differences in detail, in many countries and for many properties (e.g., Reddy et al., 2021; Liu et al., 2020; Araujo-Carrillo et al., 2021). SG2 uses random forests implemented in the `ranger` R package, with prior covariate selection by recursive feature elimination and model tuning by cross-validation of model hyperparameters (number of covariates at each tree split, number of trees in the forest). See Poggio et al. (2021) for details. SPCG is an extension of the original SoilGrids approach, but uses an ensemble of two tree-based machine learning methods: random forests (as in the original SoilGrids) and gradient boosting. See Ramcharan et al. (2018) for details.

2.4 Resolution, depths and coordinate reference systems

About 90% of gSSURGO and gNATSGO is derived from polygon maps with a design scale (1:12 000 to 1:24 000, depending on the original survey), which corresponds to MLD 1.44 to 2.3 ha (1:24 000) or 0.38 to 0.575 ha (1:12 000) polygons, depending on the definition of MLD (see above). These correspond to single grid cell resolutions of 240 to 384 m (1:24 000) or 60 to 96 m (1:12 000). gSSURGO is delivered as gridded coverages at 30 m horizontal resolution on an Albers Equal Area projection covering the CONUS, with standard parallels at 29.5° and 45.5° N and the central meridian at -96° E on the NAD83 datum, which uses the GRS80 ellipsoid. gNATSGO is a 90 m resolution generalization of this. Property information is provided per horizon or layer, each with depth limits. Thus to produce a prediction for a depth slice these must be aggregated by depth-weighted average by thickness across the depth slice. PSP predicts at 1 arc-second of longitude and latitude resolution, i.e., 0.0002777778° on the WGS84 datum, equivalent to ≈ 32 m latitude, and proportionally smaller longitude depending on latitude. Depth slices are the standards specified by GlobalSoilMap. SPCG predicts at 100 m resolution for seven point depths (0, 5, 15, 30, 60, 100 and 200 cm) in the same projection as gSSURGO. Predictions are means of a depth slice. SG2 predicts at 250 m resolution for the standard depth slices specified by GlobalSoilMap on an equal-area Interrupted Goode Homolosine (IGH) projection on the WGS84 datum (Moreira de Sousa et al., 2019). Depth slice predictions are in fact point predictions at the centre of the depth slice, considered to represent that interval. The Supplementary Information explains how these products are accessed and made compatible for comparison at regional and local scales.

2.5 Uncertainty assessment

SG2 and PSP predict the 5% and 95% quantiles of the distribution of predictions using Quantile Regression Forests (QRF) (Meinshausen, 2006). These limits are specified by the GlobalSoilMap consortium (Arrouays et al., 2014), defined as "the 90% Prediction Interval (PI) which reports the range of values within which the true value is expected to occur 9 times out of 10 …there is no assumption that this prediction interval is necessarily symmetric around the predicted value." (Science Committee, 2012). gSSURGO provides "representative", "upper" and "lower" limit values of each property of a STU, per horizon or layer. The National Soil Survey Handbook, §618.2 (United States Department of Agriculture, Natural Resources Conservation Service, n.d.) explains that the representative value approximates the median, but that the quantiles corresponding

to the low and high values can be adjusted to the percentiles which best show the spread of the property within an STU. If
there are sufficient laboratory data of sampled profiles of the STU in the National Soil Information System (NASIS) (Natural
Resources Conservation Service, n.d.), these are used as the basis for establishing the range. In all cases expert opinion is used
to adjust these to represent the range that a map user can expect to find in the field. Thus these are not directly comparable
to the results of QRF, but do give some idea of how the field mappers, supported by laboratory observations, conceive of the
spread of a property. Note that none of these assessments imply a parametric probability distribution.

As pointed out by Arrouays et al. (2020), "[t]he user community requires training in, and experience with, the new digital
soil map products, especially about the use of uncertainties". It would be hoped that the uncertainties computed by different
methods would be similar, but as will be seen here, that is not the case.

3   Evaluation methods

We compared PSM products at regional (nominal 250 m grid cells) and local (nominal 30 m grid cells) levels. We evaluated
both qualitatively, i.e., by visual inspection followed by expert interpretation, and numerically, over a $1 \times 1°$ tile. Spatial
patterns were evaluated over a $0.2 \times 0.2°$ subtile. To compare maps at the regional resolution, the higher-resolution maps were
aggregated to the lower resolution by weighted averaging (resampling) of the high-resolution pixels within one low-resolution
pixel. To compare maps at the local resolution, we only included the two products (gSSURGO and PSP) provided at that
resolution, along with the global product (SG2) as reference, this latter downscaled by increasing the grid resolution without
any attempt to disaggregate within the larger grid cell, over a $0.15 \times 0.15°$ subtile.

3.1   Qualitative methods

Qualitative methods for comparing maps rely on expert judgement to identify known soil-geographic patterns and evaluate to
what extent they are represented on the gridded maps. The maps are displayed side-by-side along with a map of their pairwise
differences. Areas of disagreement are identified and discussed.

3.2   Numerical methods – whole map

Numerical methods for comparing gridded maps as a whole include (1) MD: Mean difference, i.e., the disagreement between
maps; (2) RMSD: root mean squared difference; (3) RMSD adjusted for MD, i.e., the RMSD after accounting for any bias.
These take the first-listed map as reference and the second as the map to evaluate. They can be normalized by the number
of grid cells or total area. In addition, all maps can be compared by their Pearson (linear) correlations. These methods are of
limited interpretive value – they do characterize the entire map, especially for overall bias (MD), but do not explain where the
spatial discrepancies occur.

3.3 Numerical methods – spatial continuity

Soil properties are usually spatially-correlated: we expect similar values of properties in nearby grid cells. The degree of local spatial continuity can be assessed by the variogram computed over local neighbourhoods of the gridded map. We computed and modelled the variogram within a local neighbourhood and automatically fit with an exponential model, using the `fit.variogram` function of the `gstat` R package (Pebesma, 2004). Spatial structure is characterized by the range, proportional nugget and structural sill of the fitted variogram model. The range shows the radius over which the selected property has spatial correlation. The proportional nugget shows the inherent variability at a point, at a scale shorter than the grid spacing. The structural sill shows the overall variability within the range.

3.4 Numerical methods – patterns

Numerical methods for comparing patterns include: (1) the "V-measure" method (§3.4.1) (Nowosad and Stepinski, 2018) implemented in the `sabre` "Spatial Association Between REgionalizations" R package (Nowosad, 2020); (2) landscape-level metrics (§3.4.2) (Uuemaa et al., 2013) as used in ecology and derived from the FRAGSTATS computer program (McGarigal et al., 2012), implemented in the landscapemetrics R package (Hesselbarth et al., 2019). These include Shannon diversity and evenness, landscape shape index, and fractal dimension.

These methods must be applied to classified maps, so the continuous soil property maps must first be classified into ranges before analysis. Clearly, different choices of class limits and widths will result in different measures. A somewhat objective method to choose classes is histogram equalization. The analyst determines the number of classes, and equal numbers of grid cells are in each class. To compare maps, the combined values of all maps are used to construct the histogram. For the "V-measure" they must also be polygonized.

These metrics do not compare the values of the classes, only their spatial pattern.

3.4.1 V-measure

The V-measure method evaluates the spatial association between two regionalizations, i.e., partitions of a map into classes, called *regions* in the first map and *zones* in the second map. These are intersected to produce *segments* of the combined map. The segments, labelled with both zone and region, are used to compute two metrics, called *homogeneity* and *completeness*, with respect to the base (first-listed) regionalization. *Homogeneity* of the second map is a measure of the variance of the regions within a zone, normalized by the variance of the regions in the entire domain of the first map. If the regions within zones have less variance than within the entire second map, the second map is to some extent homogeneous with respect to the first map. A perfectly homogeneous partition (second map) (with value 1) is when each zone of the first map is entirely within a single region of the second map. A perfectly inhomogeneous partition (with value 0) is when each zone has the same composition of regions as the entire domain of the first map, i.e., the second map's partition (to be tested) is essentially random with respect to the first map's regionalization.

Completeness of the second map is the inverse of homogeneity: it assesses the variance of the zones within a region, normalized by the variance of the zones in the entire domain of the second map. This shows how well the partition of the second
310   map fits inside the regionalization.

These two together are combined into a single measure, the V-measure, as the harmonic mean of homogeneity h and completeness c (Equation 1). This has a range between 0 (no spatial association between the maps) and 1 (perfect association). Obviously, we prefer high association between maps produced by PSM and a reference map. We can also assess the agreement of the patterns produced by different PSM methods.

$$V = \frac{h \times c}{h + c} \tag{1}$$

3.4.2   Landscape metrics

The landscape metrics applicable to soil maps (as opposed to, e.g., vegetation maps) have diverse interpretations. We compare the metrics of two maps to see if they have a similar concept of the (classified) soil landscape. The `landscapemetrics` package can compute many indices; we choose a few that will best show the landscape-level difference between maps. We do
320   not consider metrics of individual patches, except as they contribute to landscape-level metrics. The algorithms for these can be read from the code repository (Hesselbarth, 2021); we present the formulas and interpretations in the following text.

- The **Shannon Diversity Index** `shdi` (Equation 2), where p_i is the proportion of pixels of class $i = (1 \ldots N)$, characterizing the landscape diversity in two senses. More classes and/or a more even distribution of proportions lead to a higher landscape diversity. This does not account for spatial contiguity, it just considers the class of each pixel, irrespective of
325   position.

$$D = -\sum_{i=1}^{N} p_i \ln p_i \tag{2}$$

- The **Shannon Evenness Index** `shei` (Equation 3) is a normalization of Shannon Diversity by the maximum diversity possible for the given number of classes (N). It varies from 0 (completely uneven distribution - low landscape diversity) to 1 (all proportions are equal - high landscape diversity). It does not depend on the number of classes. Again, this does
330   not account for spatial contiguity.

$$E = \frac{D}{\ln N} \tag{3}$$

- The **Landscape Shape Index** `lsi` (Equation 4), where A is the total area of the landscape and E' is the total length of edges, including the boundary, quantifies the internal boundary complexity of a landscape tile, with a value of 1 when

the landscape consists of a single square patch, increasing without limit as the length of edges within the landscape

335 increases.

$$\text{LSI} = \frac{0.25E'}{\sqrt{A}} \qquad (4)$$

– The **Landscape Aggregation Index** `lai` (Equation 5), where g_{ii} is the number of like adjacencies, $max - g_{ii}$ is the classwise maximum possible number of like adjacencies of class i (i.e., if all pixels in the class were in one cluster), and P_i is the proportion of landscape comprised of class i, to weight the index by class prevalence. Thus `lai` equals

340 the number of like adjacencies divided by the theoretical maximum possible number of like adjacencies, summed over each class and over the entire landscape. It ranges from 0 for maximally disaggregated to 100 for maximally aggregated landscapes.

$$\text{AI} = \left[\sum_{i=1}^{m}\left(\frac{g_{ii}}{max - g_{ii}}\right)P_i\right](100) \qquad (5)$$

– The **Mean Fractal Dimension** `frac_mn` characterizes the complexity of the landscape as the mean of the fractal

345 dimension of all patches in the landscape. It approaches 1 if all patches are square, and 2 if all patches are irregular. It is scale-independent. The patch-level fractal dimensions (Equation 6) are computed from the patch perimeters p_{ij} in linear units and areas a_{ij} in square units.

$$\text{FRAC} = \frac{2 * \ln *(0.25 * p_{ij})}{\ln a_{ij}} \qquad (6)$$

– The **Co-occurrence vector** `cove` proposed by Nowosad (2021) summarizes the entire adjacency structure of the map

350 and can be used to compare map structures. This is a normalized form of the co-occurrence matrix, which counts all the pairs of the adjacent cells for each category in a local landscape, in the form of a cross-classification matrix. This vector can be considered as a probability vector for the co-occurence of different classes. Co-occurrence vectors of different categorical maps can then be compared by computing the distance between them. Many distance measures are possible; we choose the Jensen-Shannon distance (Equation 7), which computes the entropy H of each probability vector

355 v_i and entropy of their average, and from these the distance in entropy space between them. Increasing values indicate increasing dissimilarity in the adjacency patterns. The computation of `cove` is implemented in the `motif` R package, and the Jensen-Shannon distance in the `philentropy` R package.

$$\text{JSD}(v_1, v_2) = H\left(\frac{v_1 + v_2}{2}\right) - \frac{1}{2}[H(v_1) + H(v_2)] \qquad (7)$$

3.5 Regional patterns

360 Regional patterns are at the scale of regional trends such as lithologic units, elevation zones in mountains, and repeating patterns (e.g., basin-and-range, ridge-and-valley). A "region" in this context is on the order of $1 \times 1°$ to $5 \times 5°$. An appropriate resolution for this scale is $(250\,m)^2$, as used in SG2.

For this evaluation we produced difference maps of the PSM products (SG2, SPCG, PSP) vs. the gNATSGO gridded maps. All except SG2 were upscaled to the SG2 resolution ($250\,m$). The gNATSGO maps are assumed to be the most correct. We 365 then comment on the differences and speculate on the causes, based on our knowledge of the PSM procedures used to make each product and the nature of the soil landscape.

We selected an example $1 \times 1°$ test area, based on its well-known soil-forming factors and environments and our familiarity with the soil geography. For the pattern analysis within this area we selected a $0.20 \times 0.20°$ subtile and projected the maps to the UTM18N grid on the WGS84 datum (ESPG code 32618).

370 3.6 Local patterns

Local patterns are at the scale of geomorphic features such as hillslope catenas, fluvial terraces, outwash fans, valley trains and drumlin fields. "Local" in this context is on the order of 10×10 km. An appropriate resolution for this scale is $\approx 30 \times 30$ m. Only gSSURGO and PSP map at this level; we included the other two products to see the effect of (unjustified) downscaling.

For this evaluation we used the same test areas as for the regional patterns, but examined smaller areas with distinctive 375 soil-landscape relations. These were evaluated by two methods.

3.6.1 Visual method

We produced ground overlays of key soil properties at selected depth slices with corresponding KML specifications, and displayed these in Google Earth as semi-transparent overlays, using the original resolution of each product, projected into WGS84 geographic coordinates as required by Google Earth. These were then compared with gSSURGO maps streamed 380 within Google Earth by SoilWeb Earth (California Soil Resource Lab, 2020). This shows the mapped polygons, labelled with their map unit, and linked to the map unit description, which in turn is linked to the Official Series Descriptions (OSD) (NRCS Soils, 2020c) with complete description of the soil properties modal values and ranges.

Figure 1 shows SSURGO map units draped over a ground overlay of pH (0–5 cm) from SG2, produced by the SoilWeb streaming coverage in Google EarthPro, with a point query showing the SSURGO map unit composition. The map unit is 385 described by its constituent soil series and their estimated proportions. Each series can then be queried for its Official Series Description (OSD) (NRCS Soils, 2020c), which gives a typical profile, a range of properties, and a link to lab data for the series. Clearly the pattern of properties follows the map unit delineations. This figure also shows the result of gridding the SSURGO polygons on a 30 m grid: grid cells overlap polygon boundaries.
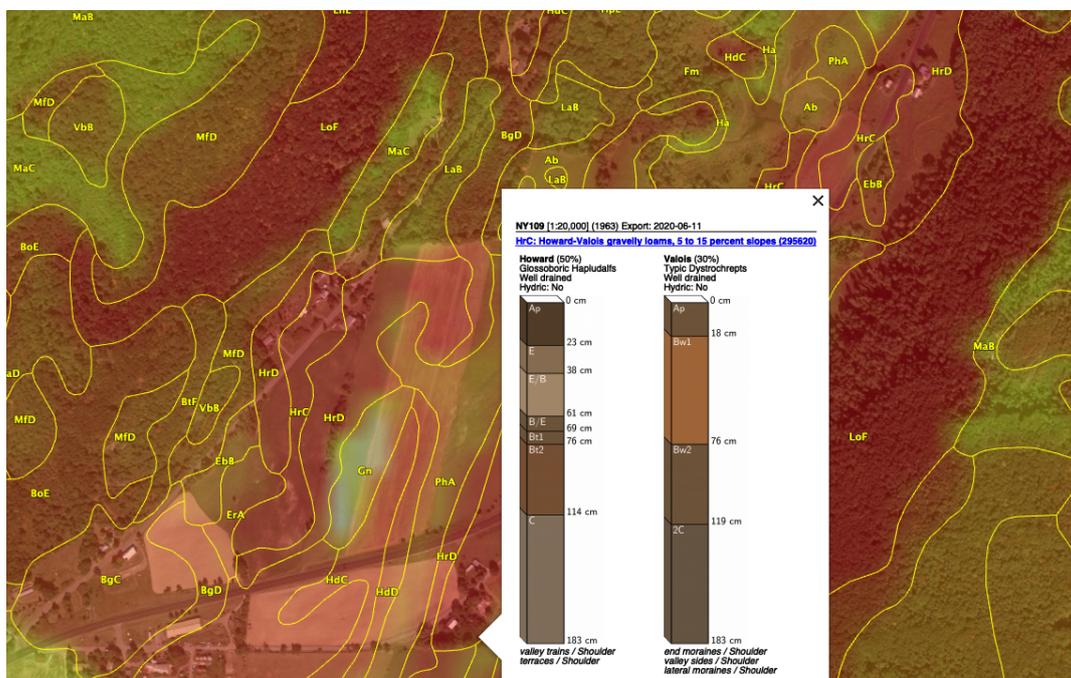
Figure 1. SoilWeb view of SSURGO map units and a ground overlay of pH, 0–5 cm, predicted by SG2. Colours are low (red) to yellow (high) pH. Point inquiry at $-76°38'05"\,\mathrm{E}, 42°19'53"\,\mathrm{N}$

3.6.2   Quantitative method

390   This follows the procedures of the regional assessment, except that V-measures are not computed, due to the very fine pattern of classified polygons.

4   Example area and soil property

To illustrate the method, we selected one area familiar to the first author, and an important soil property with strong spatial variability and pattern, namely pH in the 0–5 layer. We selected this property because in our experience this is often well-

395   modelled by PSM methods. For example, SG2 had global cross-validation statistics of 0.78 pH median RMSD and a Model Efficiency Coefficient (MEC, the R^2 of the 1:1 line actual vs. observed) of 0.67 (Poggio et al., 2021). We select the topmost depth slices because it is most represented by many environmental covariates, especially land cover as well as those derived from remote sensing. Thus the example shown here may be the best case, where all mapping methods should provide similar results.

400   The example area is in central New York State, bounding box (-77 – -76° E), (42–43° N); the subtile for pattern evaluation was (-76.8 – -76.6° E), (42.2–42.4° N), centred at Cayuta NY. See the Supplementary Information for an explanation of the regional soil geomorphology.

5 Regional spatial patterns

5.1 Regional maps

405 Table 1 shows the statistical differences between gNATSGO (reference) and the PSM products. All PSM products under-predict topsoil pH with respect to gNATSGO, by about 0.38–0.48 pH units. The RMSD is substantial also, on the order of 0.49–0.67 pH units, somewhat less than this when corrected for bias.

PSM_product	MD	RMSD	RMSD.Adjusted
SoilGrids250	3.778	6.107	4.798
POLARIS	3.841	4.915	3.067
SPCG100	4.798	6.687	4.659

Table 1. Statistical differences between gNATSGO and PSM products, pHx10, 0–5 cm

Figure 2 shows whole-map histograms. PSP has a bimodal distributions, and predicts few pH values around pH 5.8. This is quite strange since this value is well-represented in the gNATSGO map. The other distributions are fairly symmetric, although

410 SG2 and SPCG are more even than gNATSGO, which is strongly concentrated near pH 6.2. This shows the smoothing effect of the machine learning models.

Figure 3 shows the pairwise Pearson correlations between the products. The products are overall well-correlated. SG2 and SPCG are very closely correlated, since they use similar mapping methods, despite the additional covariates used by SPCG. PSP and gNATSGO are also closely-correlated. These correlations do not account for bias. They do however show that the

415 maps are similar in their overall pattern as evaluated per-grid cell.

Figure 4 shows gNATSGO (reference) along with the predictions of pH of the PSP products. Figure 5 shows these as difference maps. These figures reveal substantial differences between products. The most obvious difference is in the detail of the spatial pattern. Despite having been upscaled to regional resolution, gNATSGO shows finer detail than the other products, especially PSP.

420 These figures also show the spatial distribution of the bias. Compared to gNATSGO, SG2 and SPCG underestimate pH in the higher hills in the NE portion of the map, and in the glacio-lacustrine sediments along the lakeshores. SG2 misses the soils derived from Onondaga limestone glacial till towards the southern end of the till plain. SG2 has no information on parent material and uses global models. SPCG has very similar differences, despite using SSURGO-derived parent material as a covariate.

425 PSP predictions are closer to gNATSGO than are those of SG2, which is not surprising since PSP also uses gSSURGO as its primary information source. This product has removed some of the fine variation of gNATSGO. However the disaggregation by DSMART results in quite some discrepancies with gNATSGO. In particular, the Homer-Tully outwash valley (northeast side of map) is under-predicted by one pH unit, and the surrounding hills over-predicted by almost as much. Many of the valley trains (southern side of map, running towards the Susquehanna River) are under-predicted. This is likely due to PSP' soil series
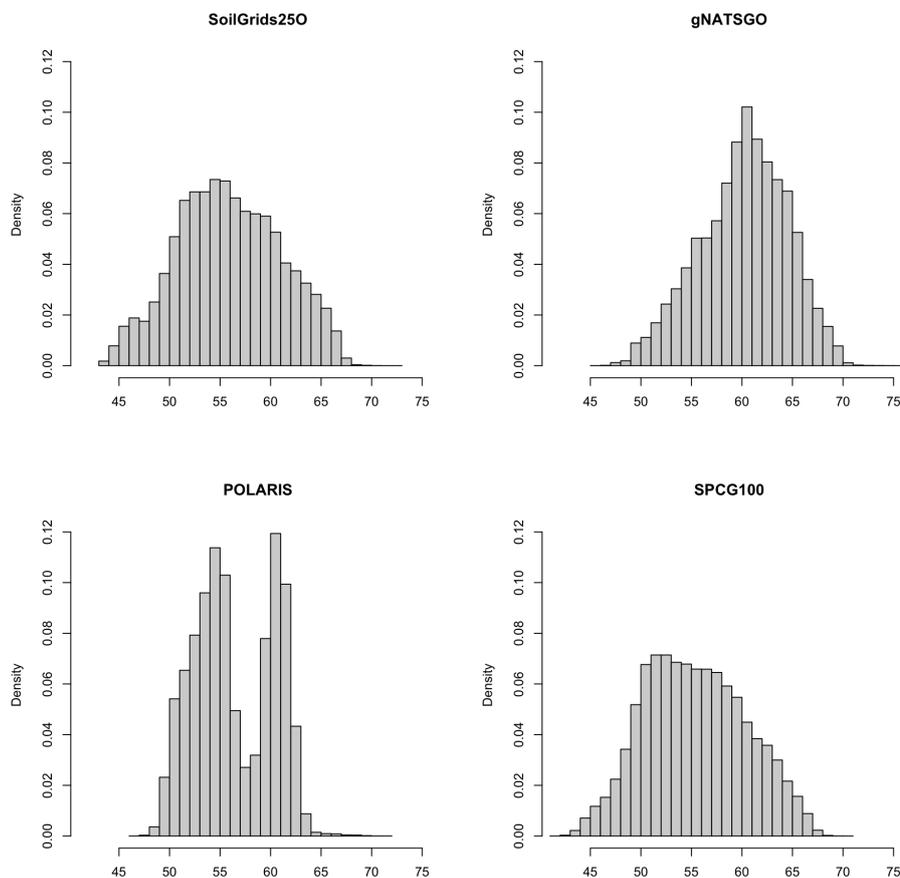
Figure 2. Histograms of pHx10, 0–5 cm

predictions, which are based on estimated map unit composition and random selection of series locations within map units for PSM calibration.

5.2 Uncertainty

The 5%, 50%, and 95% prediction quantile maps are shown in Fig. 6 (SG2) and 7 (PSP). The "low", "representative" and "high" values from gNATSGO are shown in Fig. 8 Each figure has its own stretch. Figure 9 shows the inter-quartile range 5–95% (IQR) for the two PSM products, along with the low-high range for gNATSGO.

SG2 has a fairly consistent IQR, mostly from about 2.5 to 3.5 pH, whereas PSP has a much wider range of uncertainties, mostly from about 1.5 to 4.5 pH, and shows much more spatial pattern. PSP has the widest ranges on the steep valley sides and especially in the Seneca Army Depot at the north inter-lake area, and the lowest on the broad till plains and through valleys. These are wide ranges, and although an honest reflection of the PSM models, should give pause to map users. This suggests
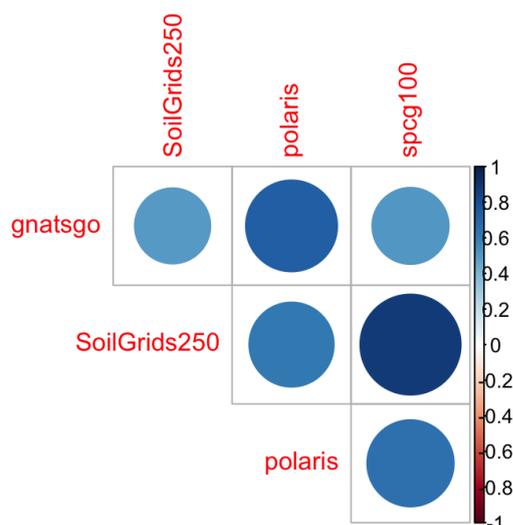
Figure 3. Pearson correlations between all products, pH, 0–5 cm

440   that the GlobalSoilMap specifications for uncertainty (Arrouays et al., 2014) are unduly pessimistic. Sources for uncertainty assessment (SG2: training points and global covariates, PSP: mapped soil series and national covariates) and the different machine learning methods lead to greatly different estimates of prediction uncertainty. gNATSGO has narrower ranges than the two PSM products and by design does not include unrealistic values.

5.3   Local spatial autocorrelation

445   The local variograms and their fitted exponential models are shown in Fig. 10. Table 2 shows their statistics.

Product	Effective range	Structural Sill	Proportional Nugget
gNATSGO	1929.00	10.30	0.00
SG2	3951.00	12.83	0.02
SPCG100USA	6924.00	11.81	0.01
POLARIS	4674.00	6.44	0.07

Table 2. Fitted variogram parameters, pH 0–5 cm. Effective range in m; structural sill in $(\mathrm{pH}x10)^2$, proportional nugget on $[0\ldots 1]$

gNATSGO has the shortest effective range. This indicate fine-scale structure at 250 m resolution. The PSM products have longer ranges, showing that these models do not capture well fine-scale variation. These show a smoothing effect, likely due
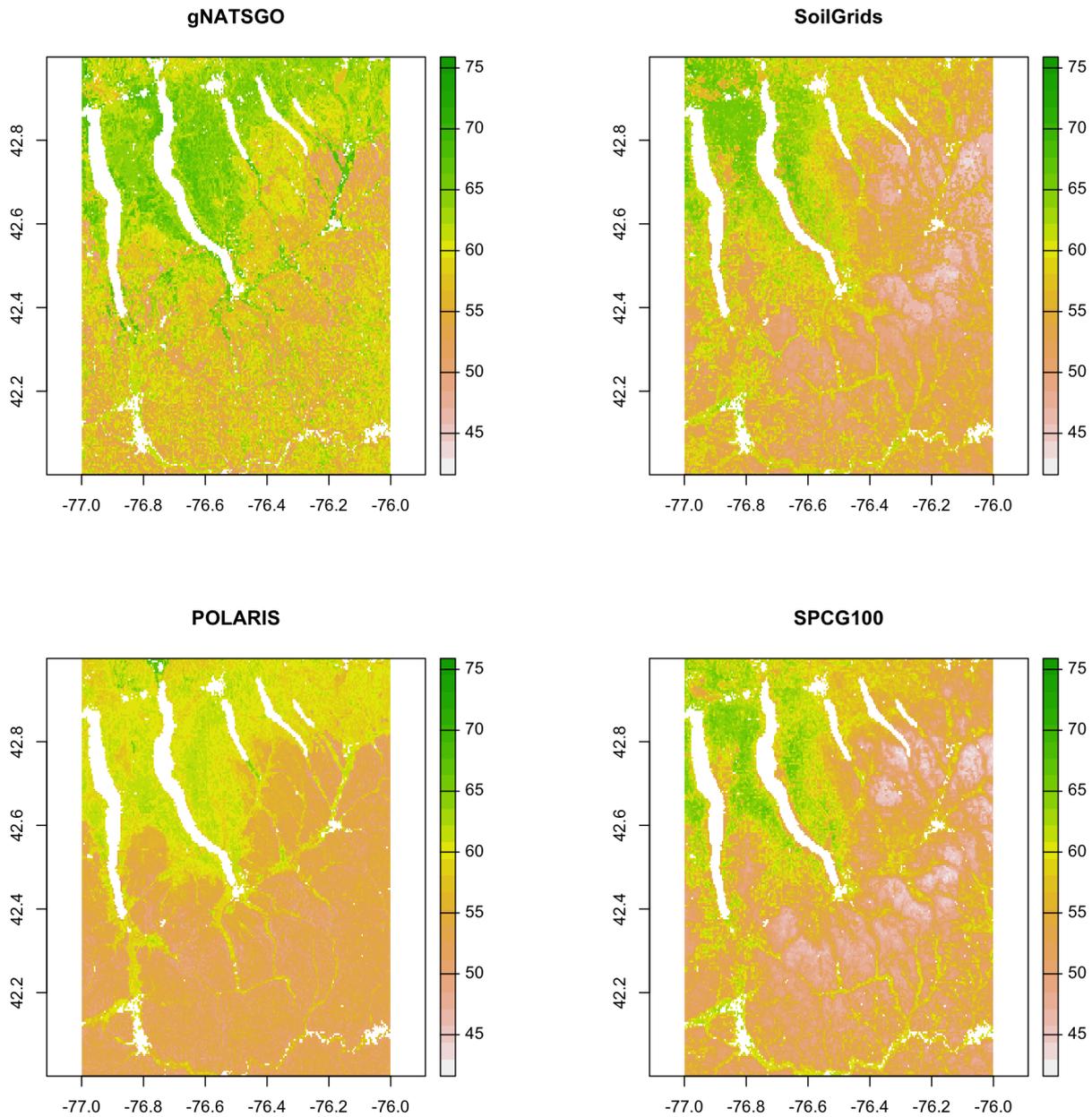
Figure 4. Topsoil (0–5 cm) pHx10, according to gNATSGO and PSM products
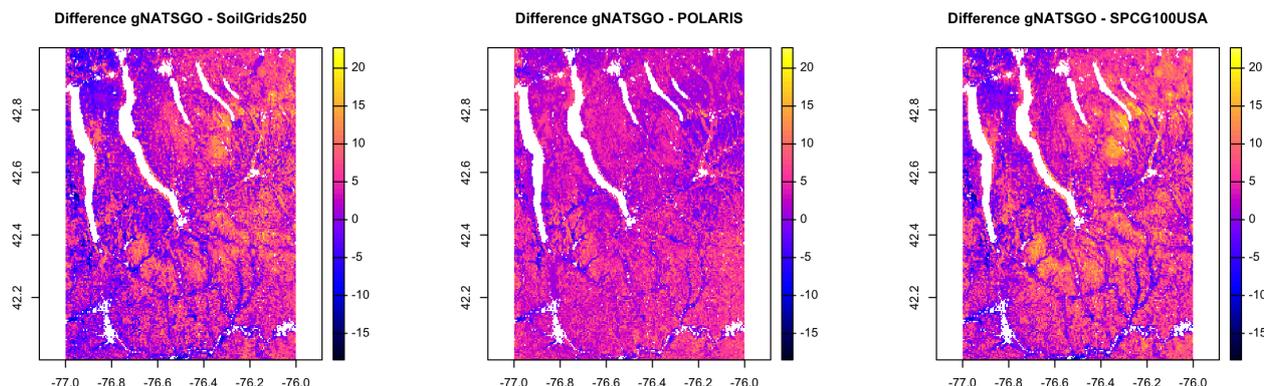
Figure 5. Difference between gNATSGO and PSM products, pHx10, 0–5 cm
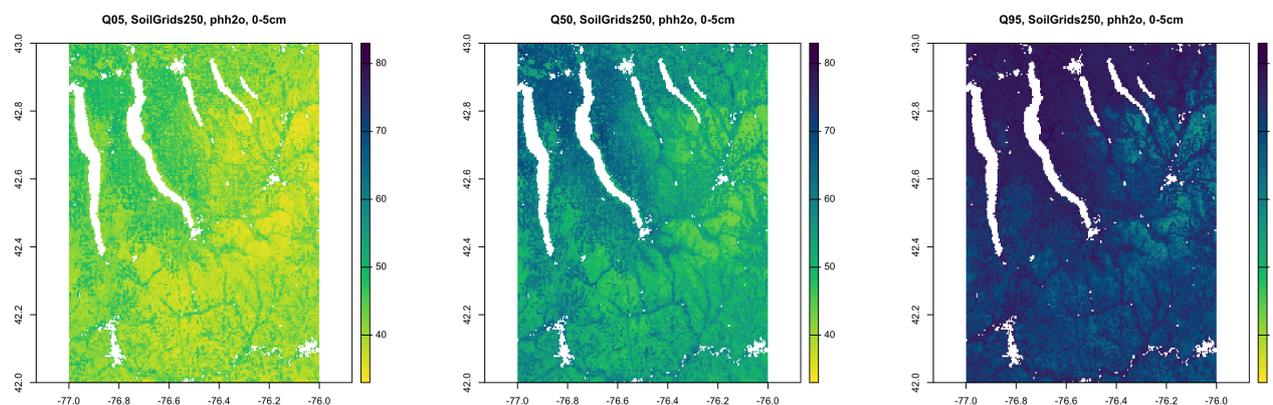
Figure 6. Quantiles of the prediction, SG2, pHx10, 0–5 cm

to spatial continuity in the covariates. PSP has a long range and low sill, due to the harmonization from DSMART. The low proportional nuggets are due to the relatively coarse resolution.

450 ## 5.4 Classification

Figure 11 shows the topsoil pH classified into eight histogram-equalized classes in the 0.2 x 0.2° sub-tile. The higher pH are shown in green, the lower in red. Class limits are approximately 5.01, 5.14, 5.27, 5.40, 5.54, 5.71, and 6.02 pH, with the extreme values of 4.52 and 6.96 pH. The maps show obvious spatial differences in class distribution. gNATSGO shows more areas in the highest pH class than the PSP products. The products based on gSSURGO, i.e., gNATSGO and PSP, show a finer
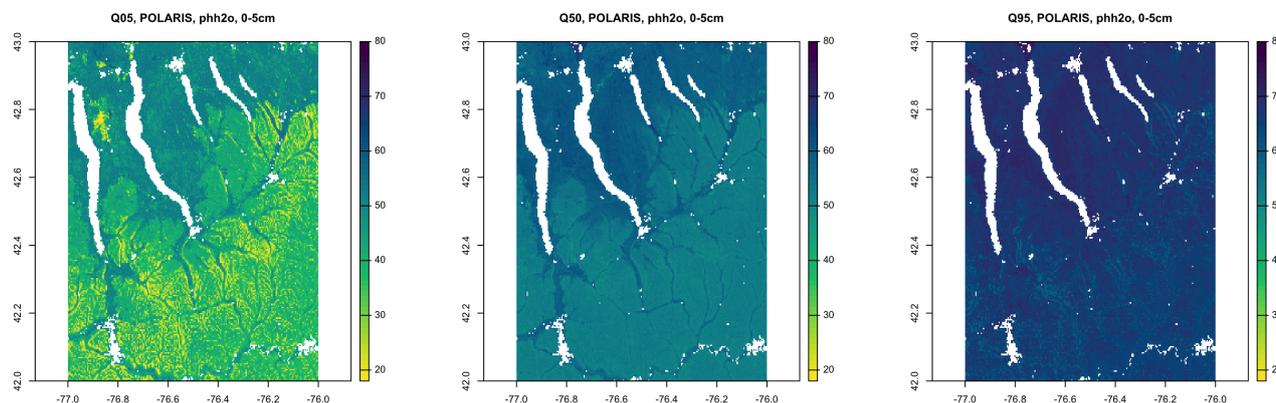
Figure 7. Quantiles of the prediction, PSP, pHx10, 0–5 cm
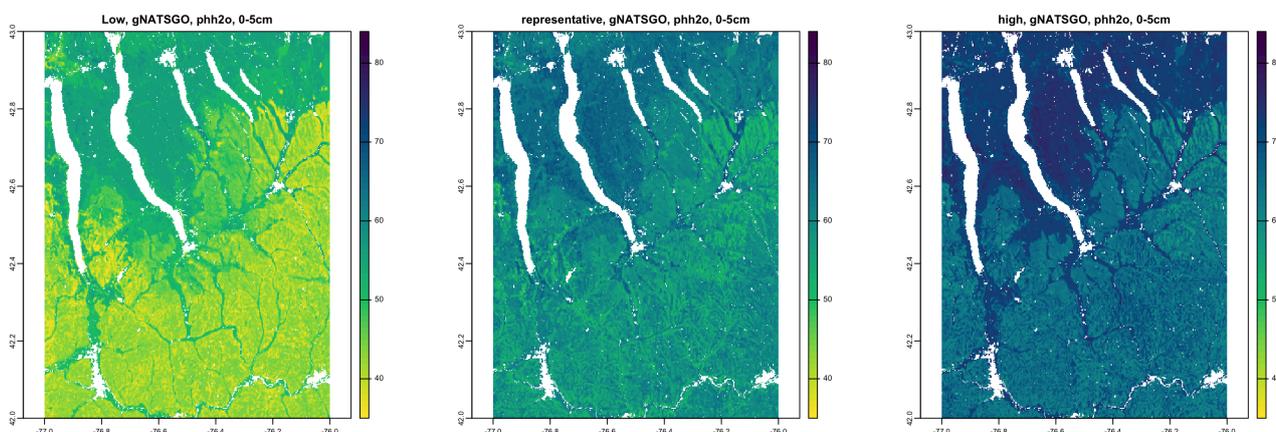
Figure 8. Low, representative, high values from gNATSGO, pHx10, 0–5 cm

spatial pattern than the purely PSM products, i.e., SG2 and SPCG. But SPCG shows large homogeneous areas of the lowest class, covering the highest hills, whereas SG2 presents a more nuanced view.

5.5 V-measure

Table 3 shows the statistics from several V-measure comparisons, based on the histogram-equalized class maps. Only SG2 and SPCG have somewhat comparable patterns. gNATSGO is considerably different from all other products, due to its detailed spatial pattern based on field survey.
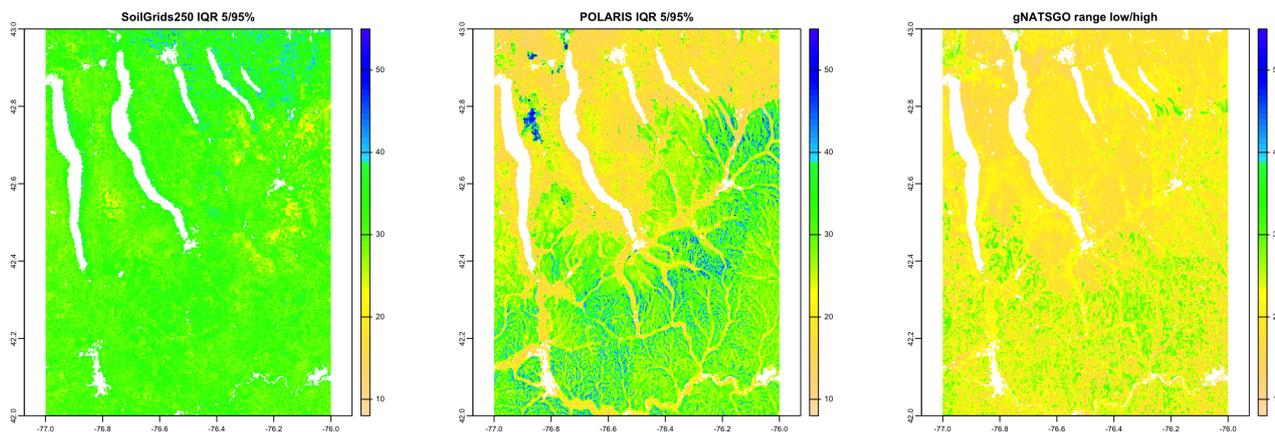
Figure 9. Inter-quantile ranges 0.05–0.95, pHx10, 0–5 cm
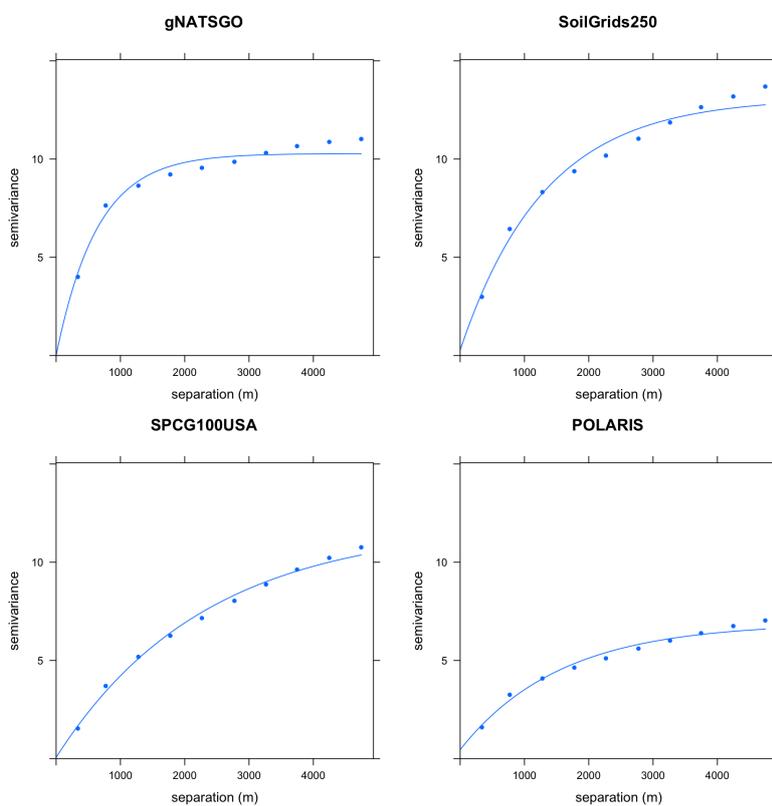
Figure 10. Fitted variograms, pH 0–5 cm. Semivariance units $(\mathrm{pH}x10)^2$
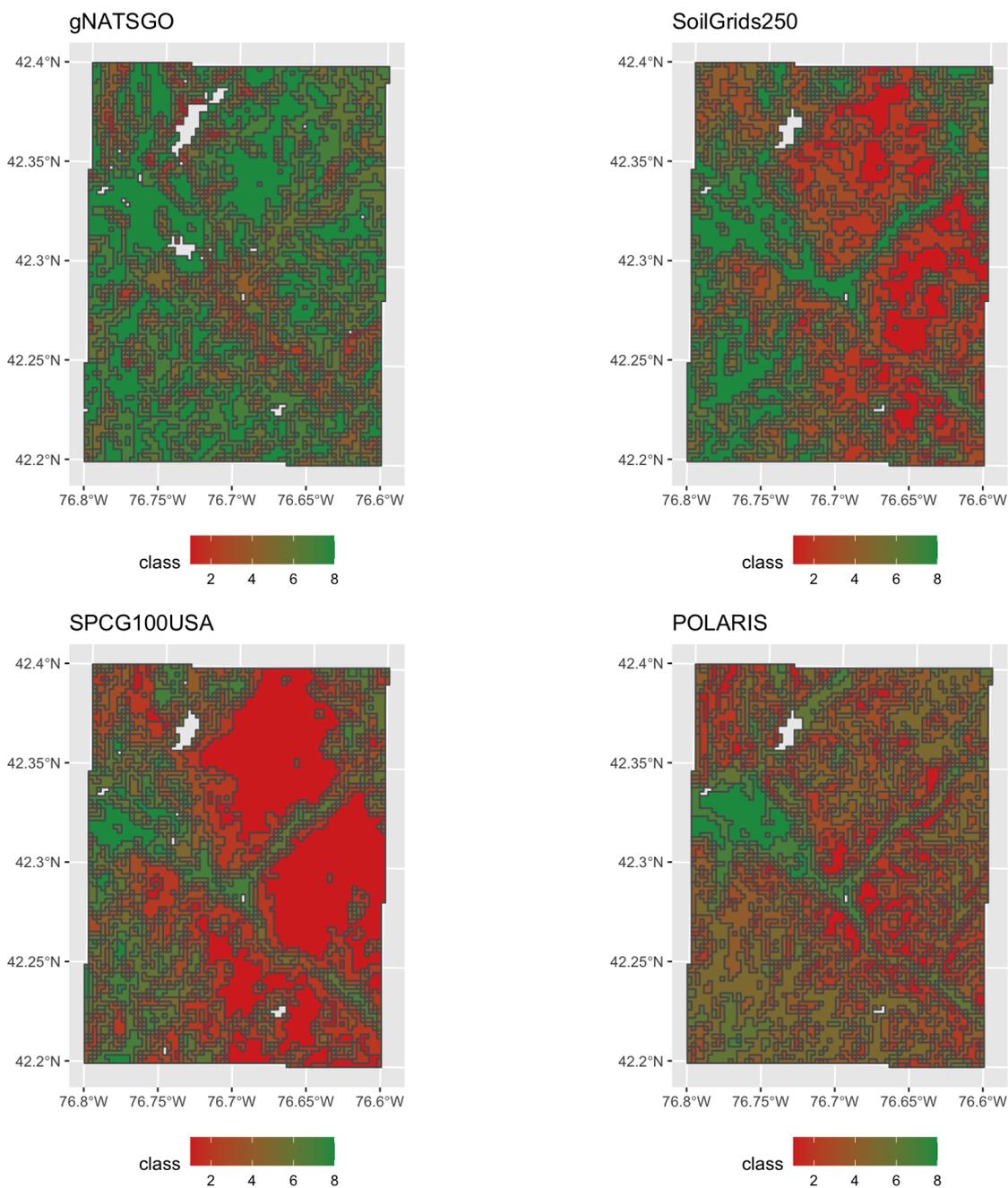
Figure 11. pH classes, 0–5 cm, central NY, detail

PSM_products	V_measure	Homogeneity	Completeness
gNATSGO vs. SG2	0.0131	0.0146	0.0118
gNATSGO vs. SPCG100USA	0.0259	0.0276	0.0244
gNATSGO vs. POLARIS	0.0839	0.0896	0.0789
SG2 va. SPCG100USA	0.3341	0.32	0.3496

Table 3. V-measure statistics, pHx10 0–5 cm
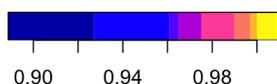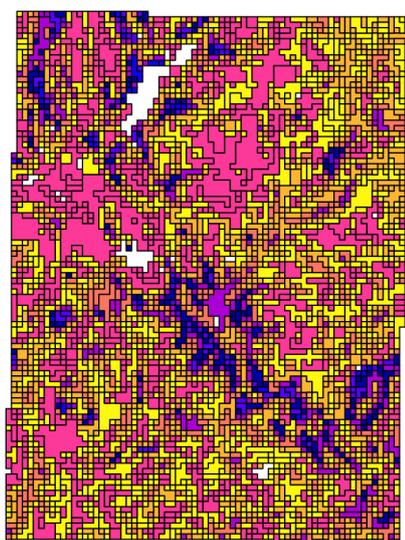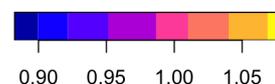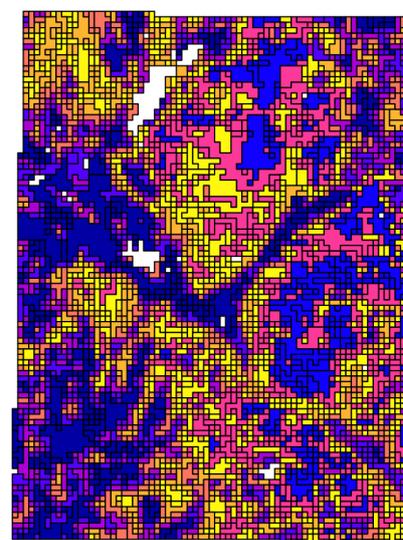
Figure 12. Homogeneity (left) and Completeness (right) of the SG2 pH class map, with respect to gNATSGO pH class map, 0–5 cm

Figure 12 shows the computed homogeneity and completeness of the SG2 pH class map, with respect to the gNATSGO pH class map. In the yellow areas of the homogeneity map one, one gNATSGO predicted class is contained in the SG2 region; in the blue areas many are. Overall the agreement is fairly good.

5.6 Landscape metrics

465   Table 4 shows the statistics from the landscape metrics calculations. The mean fractal dimensions are almost identical. There is quite some range of aggregations, with SPCG most aggregated, i.e., least complex. Otherwise the results are inconsistent; all we can say is that the map patterns vary considerably among products. Table 5 shows the Jensen-Shannon distance beween

product	ai	frac_mn	lsi	shdi	shei
gNATSGO	48.17	1.034	22.605	1.666	0.801
SoilGrids	50.631	1.034	21.78	2.062	0.991
SPCG100USA	58.489	1.041	18.551	1.887	0.907
POLARIS	47.002	1.04	23.241	1.897	0.912

Table 4. Landscape metrics statistics, pH 0–5 cm. `frac_mn`: Mean Fractal Dimension; `lsi`: Landscape Shape Index; `shdi`: Shannon Diversity; `shei`: Shannon Evenness; `ai`: Aggregation Index

	gNATSGO	SoilGrids	SPCG100USA	POLARIS
gNATSGO	0.000	0.148	0.281	0.261
SoilGrids	0.148	0.000	0.067	0.085
SPCG100USA	0.281	0.067	0.000	0.111
POLARIS	0.261	0.085	0.111	0.000

Table 5. Jensen-Shannon distance beween co-occurence vectors

co-occurence vectors of the four products. The co-occurence patterns of SG2 is quite similar to that of the other products, whereas gNATSGO is quite different than PSP.

6 Local spatial patterns

The main interest here is to see how well PSM methods at relatively fine resolution reproduce known relations at the local geomorphic level, e.g., hillslopes, transects across valleys with multiple terrace levels, and within farms. It has been claimed that PSM at 30 m resolution is sufficient for management of, or even within, individual farm fields. This is the only PSM product which predicts at this resolution.

We examine this first qualitatively, i.e., by visual inspection, and then quantitatively, mostly following the methods of the regional assessment.

6.1 Qualitative assessment

Here we use silt concentration, as it reveals stronger qualitative discrepancies than pH in this test area. Figure 13 shows the silt concentration of the 0–5 cm layer for (top) the gridded SSURGO overlain on the original polygons from which it was derived, and (bottom) the disaggregated PSP grid cells in a hilly landscape near Caroline, NY. Red colours are low silt, in this window alluvial fans (the `C*` map units). Pale grey colours are organic soils (the `Hk`, `Hl` map units). Light colours are high-silt surface soils (the `L*`, `V*`, `B*`, `M*` map units), from thin glacial till developed on shale and mudstone bedrock.

The gSSURGO product follows the SSURGO lines exactly. Some of the sharp boundary lines do correspond with abrupt transitions on the ground, for example where the steep hillsides are buried by fan alluvium. But others are not, for example

Figure 13. Ground overlay from gSSURGO (top) and PSP (bottom), silt % 0–5 cm, with SSURGO polygons from SoilWeb. Centre of image
$-76°16'25"$ E, $42°22'53"$ N; view azimuth $247°$

485 on the hilltops. These differences are because the predicted silt concentrations are taken from the official series descriptions. PSP follows the map unit lines fairly well, but is much finer-grained; each 30 m pixel is separately predicted. This results in some smoothing of the abrupt boundary lines from gSSURGO on the hilltops. However within some SSURGO map units PSP predicts quite some differences in topsoil silt concentration. These are map units with contrasting components, which PSP attempts to disaggregate according to their correlation with covariates. For the most part these do not seem to be related to

490 terrain or land use.

For example, Fig. 14 shows detail of the Holly-Papakting map unit within this PSP window. This map unit has two contrasting soils in similar proportions: a mineral alluvial soil (Holly series) and an organic soil (Papakting series); the second has much lower silt concentration. It is difficult to see the reason for the pattern within this map unit. PSP has placed the component series in their proper proportions but not according to any landscape feature.
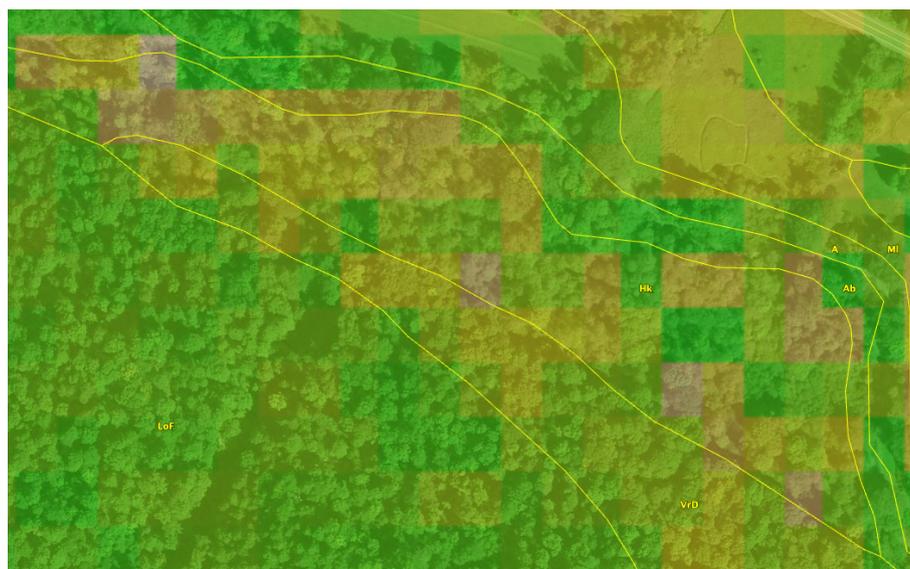
Figure 14. Ground overlay from PSP in the Holly-Papakting map unit, silt % 0–5 cm. Centre of image $-76°16'03"\,\mathrm{E}, 42°22'30"\,N$

6.2 Quantitative assessment

495

To see the fine differences at this high resolution, we consider a $0.15 \times 0.15°$ subtile with lower-right corner $-76.30°\mathrm{E}, 42.45°\mathrm{N}$ and evaluate pH, as in the regional assessment (§5).

Table 6 shows the statistical differences between gSSURGO (reference) and the PSM products, along with the predictions of pH. Figure 15 shows the pairwise Pearson correlations between the maps. These results are comparable to those for the full

500 tile at regional resolution: both SG2 and PSP under-predict pH by about 0.35–0.45 pH. Correlations are fairly strong between PSP and gSSURGO, and between SG2 and PSP, but weak between SG2 and gSSURGO.

PSM_product	MD	RMSD	RMSD.Adjusted
SoilGrids250	4.436	6.758	5.097
POLARIS	3.462	5.625	4.433

Table 6. Statistical differences between gSSURGO and PSM products, pHx10, 0–5 cm. Centre of map $-76°30'30"$ E, $42°52'30"$ N
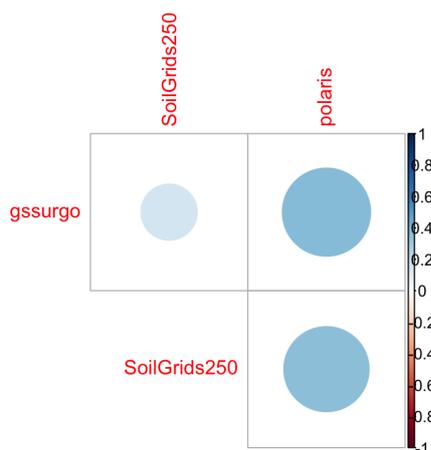
Figure 15. Pearson correlations between local products, pH, 0–5 cm

Figure 16 shows gSSURGO (reference) along with the predictions of pH by the PSP products. Figure 17 shows these as difference maps. Clearly, gSSURGO has overall higher values than the other two products, and despite the fine resolution, has in general large areas of identical values. The differentiation between map units follows sharp boundaries even within a single
505 landscape (e.g., the plateau towards the S of the map), and this is likely an artefact of relying on the representative profiles in the official series descriptions for property values. PSP has a finer pattern, due to disaggregation, and shows a more realistic local pattern by smoothing out the sharp boundaries between map units within a landscape. PSP shows large areas of low pH. SG2 does not follow well the landscape lines, especially the sharp boundaries between uplands and valleys, and predicts very low pH (≈ 4.5) on the plateau. It is difficult to recognize local landscape units in this global product.

510 **6.2.1 Class maps**

Figure 18 shows the topsoil pH classified into eight histogram-equalized classes. Class limits in this area are approximately 5.30, 5.44, 5.55, 5.61, 5.74, 5.89, and 6.15 pH, with the extreme values of 4.44 and 7.00 pH. SG2 clearly is less detailed than the other two products. PSP shows a fine pattern, not closely related to the fine pattern of gSSURGO. As previously noted, gSSURGO is consistently about one pH class higher than the other products.
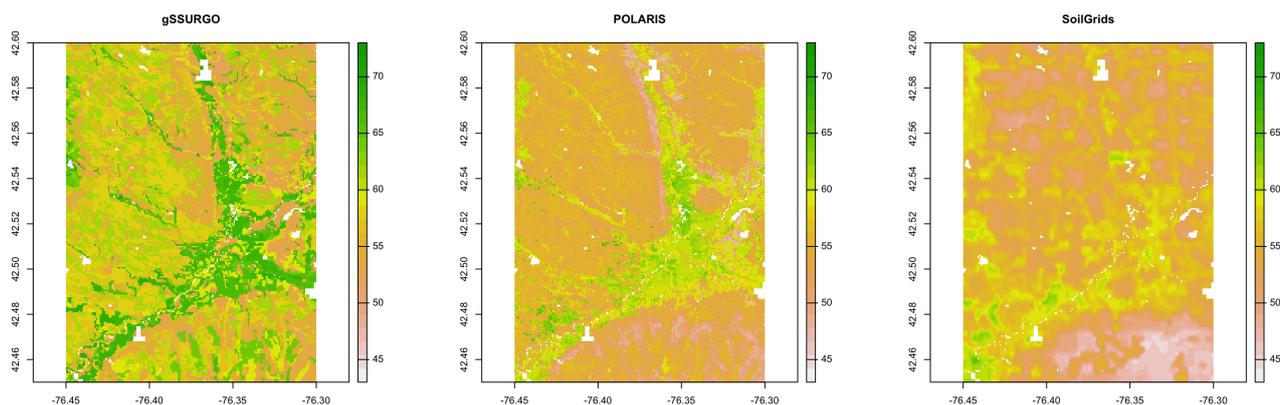
Figure 16. Topsoil (0–5 cm) pHx10, according to gSSURGO and PSM products
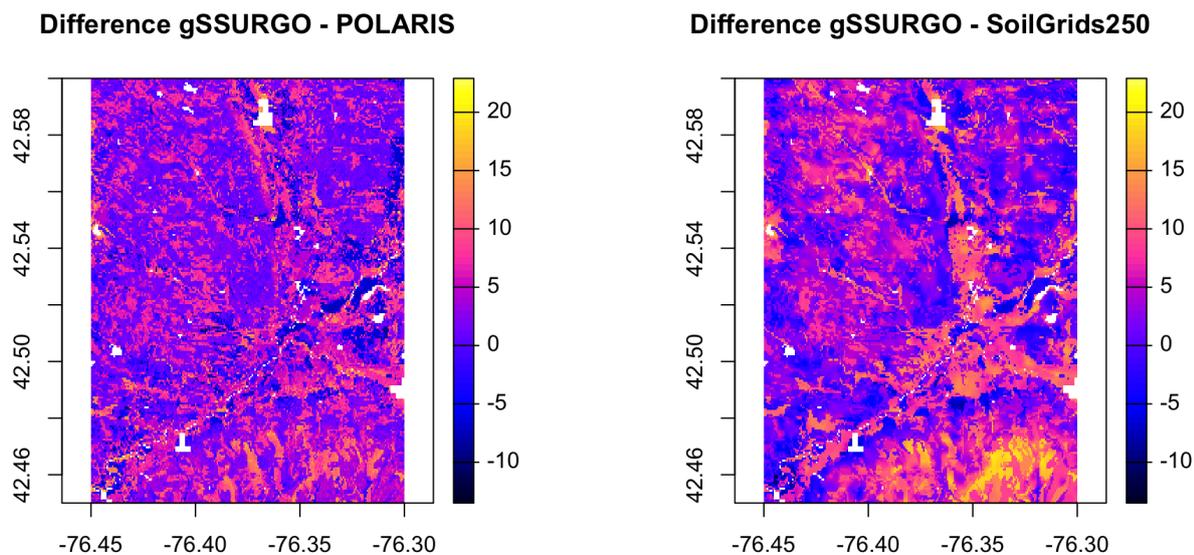
Figure 17. Difference between gSSURGO and PSM products, pHx10, 0–5 cm
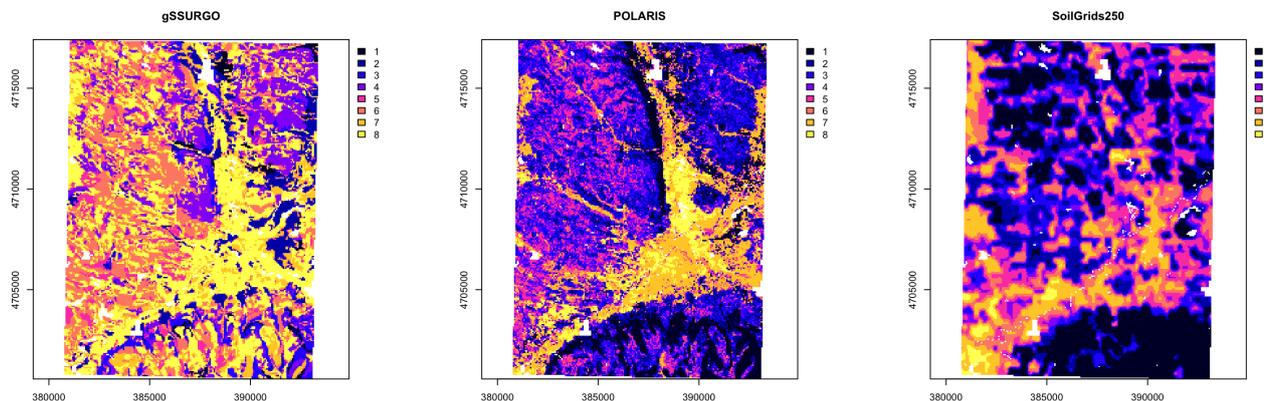
Figure 18. pH classes, 0–5 cm. Coordinates are UTM 18N meters
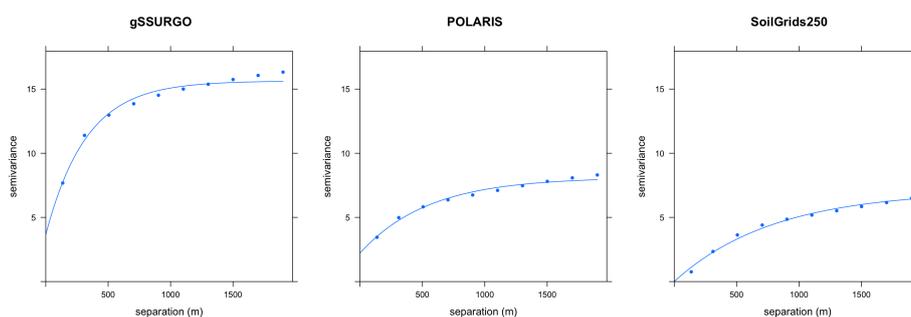
Figure 19. Fitted variograms, pH 0–5 cm. Semivariance units $(\mathrm{pH}x10)^2$

Product	Effective range	Structural Sill	Proportional Nugget
gSSURGO	975.00	11.97	0.23
SG2	2394.00	7.08	0.00
POLARIS	1638.00	5.91	0.27

Table 7. Fitted variogram parameters, pH 0–5 cm. Effective range in m; structural sill in $(\mathrm{pH}x10)^2$, proportional nugget on $[0\ldots1]$

6.2.2 Local spatial autocorrelation

The local variograms and their fitted exponential models are shown in Fig. 19. Table 7 shows their statistics. gSSURGO has the shortest effective range and highest sill. PSP has a longer range and low sill, due to the harmonization from DSMART that removes some of the overall variability. SG2 has no nugget variance, a low sill, and long range, consistent with its regional scale.

product	ai	frac_mn	lsi	shdi	shei
gSSURGO	73.293	1.050	72.465	1.837	0.883
SoilGrids	87.650	1.106	34.969	1.941	0.934
POLARIS	56.376	1.045	116.477	2.006	0.965

Table 8. Landscape metrics statistics (local), pH 0–5 cm. `frac_mn`: Mean Fractal Dimension; `lsi`: Landscape Shape Index; `shdi`: Shannon Diversity; `shei`: Shannon Evenness; `ai`: Aggregation Index

	gSSURGO	SoilGrids	POLARIS
gSSURGO	0.000	0.223	0.172
SoilGrids	0.223	0.000	0.112
POLARIS	0.172	0.112	0.000

Table 9. Jensen-Shannon distance beween co-occurence vectors (local)

6.2.3 Landscape metrics

Table 8 shows the statistics from the landscape metrics calculations. The mean fractal dimensions are almost identical. SG2 is much more aggregated, i.e., least complex, than gSSURGO or PSP. PSP has a higher landscape shape and Shannon diversity than the other products. Table 5 shows the Jensen-Shannon distance beween co-occurence vectors of the four products. The co-occurence patterns of SG2 is somewhat similar to that PSP but quite different from gSSURGO.

7 Conclusions

The presented methods are well-able to expose differences between PSM mapping methods, and between these and field survey. It is clear from the "best case" example presented in this paper that different PSM methods, with different training points, covariates, and algorithms, can produce quite different predictive soil maps. Comparing maps with point-wise evaluation from (almost always biased) field observations gives an incomplete picture of how the different methods represent the soil landscape, which is after all what dictates how the soil is used and managed. The main findings from the example case are:

1. Although the regional products (250 m resolution) are well-correlated, the PSM products are biased, under-predicting topsoil pH by about 0.38–0.48 pH units. They also differ substantially, with a RMSD adjusted for bias on the order of 0.31–0.48 pH. This is based on representative pH values of the mapped soil series, not on measured values.

2. The PSM products differ substantially among themselves and with the reference product in their local spatial pattern, as revealed by empirical variograms. gNATSGO has a short effective range, but this is smoothed to a range 2 to 3.5 times as long by PSM.

3. Classification by histogram equalization reveals major differences in the spatial patterns of the produced class maps, as evaluated both by visual inspection and landscape metrics.

4. Despite using USA-specific covariates (parent material, drainage classes) derived from gSSURGO and covariates limited in geographic scope to the USA, the predictive map made by SPCG is not substantially different from that made by SG2, likely due to the similar modelling method.

5. The estimates of uncertainty provided by SG2 and PSP are substantially different, both in width of the uncertainty interval (much narrower in SG2) and in spatial pattern. The confidence intervals seem unrealistically wide compared to the expert-derived high-low value range provided by gSSURGO.

6. At the local level (30 m resolution) the disaggregation provided by PSP does not appear to correspond to landscape positions associated with STU components. PSP obscures the fine-scale details of the local spatial pattern, and SG is substantially more general, due to its resolution.

These results will differ in different soil geographic regions, for different soil properties, and for different depth slices, as shown in the companion Case Studies report.

Our methods have two limitations due to the decision that they be applicable, using the supplied computer code, to any area within the USA. The first is the use of histogram equalization for the class maps which are then evaluated for the class pattern. For specific areas and properties it would be preferable to use established class limits relevant for land use, for example limits from soil survey interpretation tables. The second is the choice of the exponential model for automatic variogram fitting, as well as the somewhat arbitrary choice of empirical variogram cutoff and bin width. For each area, property and depth slice variograms could be computed and fit according to the analysts' prior knowledge.

Despite the discrepencies between PSM products and field survey, PSM can be a valuable tool for soil survey. It has the advantage of being reproducible and objective, given a set of training points, relevant environmental covariates, and a PSM method. Therefore, its output should be examined and compared with field-based maps to identify possible improvements, especially in areas with difficult field access or complex, difficult to interpret soil-landscape patterns. For unsurveyed areas PSM provides a useful pre-map for planning sampling and field survey. There is no substitute for actually examining the soil and landscape, but despite the issues revealed in this study, PSM can be an important aid to the soil surveyor.

Code availability. Source code as R Markdown documents are available at https://github.com/ncss-tech/compare-psm. These can be used to (1) import all products to compare, as well as some others not considered in this study; (2) create ground overlays and corresponding KML files for display in Google Earth; (3) compare SG2 and PSP for $1 \times 1°$ tiles; (4) compare SG2 with SPCG and gNATSGO for any rectangular tile; (5) compute landscape metrics and compare them between products for any subtile of these; (6) evaluate the success of PSP in disaggregating at 30 m resolution.

Author contributions. DGR conceptualized the approach, did most of the writing, wrote the R Markdown documents and performed the example case study. LP provided PSM expertise and detailed knowledge of SG2. DB and LZ provided USA-specific expertise, in particular about the NRCS and its products and services. All authors collaborated on the motivation, methods and conclusions.

570 *Competing interests.* There are no competing interests.

Acknowledgements. The contribution of Zamir Libohova was mostly accomplished during his tenure at USDA-NRCS-National Soil Survey Center, 100 Centennial Mall North, Room 152, Lincoln, NE 68508-3866 USA.

His contribution was partly supported by the LE STUDIUM Loire Valley Institute for Advanced Studies through its LE STUDIUM Research Consortium Programme.

References

Araujo-Carrillo, G. A., Varón-Ramírez, V. M., Jaramillo-Barrios, C. I., Estupiñan-Casallas, J. M., Silva-Arero, E. A., Gómez-Latorre, D. A., and Martínez-Maldonado, F. E.: IRAKA: The First Colombian Soil Information System with Digital Soil Mapping Products, CATENA, 196, 104 940, https://doi.org/10.1016/j.catena.2020.104940, 2021.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., d.L. Mendonca-Santos, M., Minasny, B., Montanarella, L., Odeh, I. O., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: GlobalSoilMap: Towards a Fine-Resolution Global Grid of Soil Properties, Advances in Agronomy, 125, 93–134, 2014.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A. C., Morgan, C. L., Roudier, P., Poggio, L., and Mulder, V. L.: Impressions of Digital Soil Maps: The Good, the Not so Good, and Making Them Ever Better, Geoderma Regional, 20, e00 255, https://doi.org/10.1016/j.geodrs.2020.e00255, 2020.

Batjes, N. H., Ribeiro, E., and van Oostrum, A.: Standardised Soil Profile Data to Support Global Mapping and Modelling (WoSIS Snapshot 2019), Earth System Science Data, 12, 299–320, https://doi.org/10.5194/essd-12-299-2020, 2020.

Brus, D., Kempen, B., and Heuvelink, G.: Sampling for Validation of Digital Soil Maps, European Journal of Soil Science, 62, 394–407, https://doi.org/10.1111/j.1365-2389.2011.01364.x, 2011.

California Soil Resource Lab: SoilWeb Apps, https://casoilresource.lawr.ucdavis.edu/soilweb-apps/, 2020.

Chaney, N., Minasny, B., Herman, J., Nauman, T., Brungard, C., Morgan, C., McBratney, A., Wood, E., and Yimam, Y.: POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties over the Contiguous United States, Water Resources Research, 55, 2916–2938, https://doi.org/10.1029/2018WR022797, 2019.

D'Avelo, T. P. and McLeese, R. L.: Why Are Those Lines Placed Where They Are?: An Investigation of Soil Map Recompilation Methods, Soil Survey Horizons, 39, 119–126, https://doi.org/10.2136/sh1998.4.0119, 1998.

Forbes, T., Rossiter, D., and Van Wambeke, A.: Guidelines for Evaluating the Adequacy of Soil Resource Inventories, Cornell University Department of Agronomy, Ithaca, NY, 1982.

Hengl, T., de 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., and Gonzalez, M. R.: SoilGrids1km — Global Soil Information Based on Automated Mapping, PLOS ONE, 9, e105 992, https://doi.org/10.1371/journal.pone.0105992, 2014.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global Gridded Soil Information Based on Machine Learning, PLOS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hesselbarth, M. H.: R-Spatialecology/Landscapemetrics, r-spatialecology, https://github.com/r-spatialecology/landscapemetrics, 2021.

Hesselbarth, M. H., Sciaini, M., With, K. A., Wiegand, K., and Nowosad, J.: Landscapemetrics: An Open-Source R Tool to Calculate Landscape Metrics, Ecography, 42, 1648–1657, https://doi.org/10.1111/ecog.04617, 2019.

Hudson, B. D.: The Soil Survey as Paradigm-Based Science, Soil Science Society of America Journal, 56, 836–841, https://doi.org/0.2136/sssaj1992.03615995005600030027x, 1992.

ISRIC - World Soil Information: SoilGrids – Global Gridded Soil Information, https://www.isric.org/explore/soilgrids, 2020.

Liu, F., Rossiter, D. G., Zhang, G.-L., and Li, D.-C.: A Soil Colour Map of China, Geoderma, 379, 114 556, https://doi.org/10.1016/j.geoderma.2020.114556, 2020.

Mallavan, B., Minasny, B., and McBratney, A.: Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe, in: Digital Soil Mapping, edited by Boettinger, J. L., Howell, D. W., Moore, A. C., Hartemink, A. E., and Kienast-Brown, S., pp. 137–150, Springer Netherlands, Dordrecht, http://www.springerlink.com/content/g227720377103484/, 2010.

615 McBratney, A. B., Mendon\c{c}a Santos, M. L., and Minasny, B.: On Digital Soil Mapping, Geoderma, 117, 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

McGarigal, K., Cushman, S. A., and Ene, E.: FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps, Tech. rep., University of Massachusetts, Amherst, MA, http://www.umass.edu/landeco/research/fragstats/fragstats.html, 2012.

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, 2006.

620 Minasny, B. and McBratney, A. B.: Digital Soil Mapping: A Brief History and Some Lessons, Geoderma, 264, Part B, 301–311, https://doi.org/10.1016/j.geoderma.2015.07.017, 2016.

Moreira de Sousa, L., Poggio, L., and Kempen, B.: Comparison of FOSS4G Supported Equal-Area Projections Using Discrete Distortion Indicatrices, ISPRS International Journal of Geo-Information, 8, 351, https://doi.org/10.3390/ijgi8080351, 2019.

Natural Resources Conservation Service: Web Soil Survey, https://websoilsurvey.nrcs.usda.gov/, 2019.

625 Natural Resources Conservation Service: National Soil Information System (NASIS), https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/tools/?cid=nrcs142p2_053552, n.d.

Nowosad, J.: sabre: Spatial Association Between Regionalizations, https://nowosad.github.io/sabre/, 2020.

Nowosad, J.: Motif: An Open-Source R Tool for Pattern-Based Spatial Analysis, Landscape Ecology, 36, 29–43, https://doi.org/10.1007/s10980-020-01135-0, 2021.

630 Nowosad, J. and Stepinski, T. F.: Spatial Association between Regionalizations Using the Information-Theoretical V-Measure, International Journal of Geographical Information Science, 32, 2386–2401, https://doi.org/10.1080/13658816.2018.1511794, 2018.

NRCS Soils: Gridded National Soil Survey Geographic Database (gNATSGO), https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcseprd1464625, 2020a.

NRCS Soils: Soils, https://nrcs.app.box.com/v/soils, 2020b.

635 NRCS Soils: Official Soil Series Descriptions, https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/class/?cid=nrcs142p2_053587, 2020c.

Odgers, N. P., McBratney, A. B., Minasny, B., Sun, W., and Clifford, D.: DSMART: An Algorithm to Spatially Disaggregate Soil Map Units, in: GlobalSoilMap: Basis of the Global Spatial Soil Information System, edited by Arrouays, D., McKenzie, N., Hempel, J., DeForges, A. C. R., and McBratney, A., pp. 261–266, CRC Press-Taylor & Francis Group, Boca Raton, 2014.

640 Pebesma, E. J.: Multivariable Geostatistics in S: The Gstat Package, Computers & Geosciences, 30, 683–691, https://doi.org/10.1016/j.cageo.2004.03.012, 2004.

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, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

R Studio: R Markdown, https://rmarkdown.rstudio.com/, 2020.

645 Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., and Thompson, J.: Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution, Soil Science Society of America Journal, 82, 186–201, https://doi.org/10.2136/sssaj2017.04.0122, 2018.

Reddy, N. N., Chakraborty, P., Roy, S., Singh, K., Minasny, B., McBratney, A. B., Biswas, A., and Das, B. S.: Legacy Data-Based National-Scale Digital Mapping of Key Soil Properties in India, Geoderma, 381, 114 684, https://doi.org/10.1016/j.geoderma.2020.114684, 2021.

650 Rossiter, D. G., Poggio, L., Beaudette, D., and Libohova, Z.: How Well Does Predictive Soil Mapping Represent Soil Geography? An Investigation from the USA. Case Studies., ISRIC Report (in review), ISRIC-World Soil Information, ISRIC-World Soil Information, https://doi.org/(to be assigned), 2021.

Schoeneberger, P. J., Wysocki, D. A., Benham, E. C., and Soil Survey Staff: Field Book for Describing and Sampling Soils, USDA Natural Resources Conservation Service, Lincoln, NE, 3.0 edn., http://soils.usda.gov/technical/fieldbook/, 2012.

655 Science Committee: Specifications: Tiered GlobalSoilMap.Net Products; Release 2.3, Tech. rep., GlobalSoilMap.net, http://www.ozdsm.com.au/resources/GlobalSoilMap%20specs%20version%202point3.pdf, 2012.

Scull, P., Franklin, J., Chadwick, O., and McArthur, D.: Predictive Soil Mapping: A Review, Progress in Physical Geography, 27, 171–197, https://doi.org/10.1191/0309133303pp366ra, 2003.

Soil Survey Division Staff: Keys to Soil Taxonomy, US Government Printing Office, Washington, DC, twelfth edn., www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/class/, 2014.

660 Soil Survey Division Staff: Soil Survey Manual, no. 18 in USDA Handbook, Government Printing Office, Washington, DC, http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/planners/?cid=nrcs142p2_054262, 2017.

Thompson, J. A., Kienast-Brown, S., D'Avello, T., Philippe, J., and Brungard, C.: Soils2026 and Digital Soil Mapping – A Foundation for the Future of Soils Information in the United States, Geoderma Regional, 22, e00 294, https://doi.org/10.1016/j.geodrs.2020.e00294, 2020.

665 United States Department of Agriculture, Natural Resources Conservation Service: National Soil Survey Handbook, United States Department of Agriculture, Natural Resources Conservation Service, Washington, D.C., https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/home/?cid=nrcs142p2_054242, n.d.

Uuemaa, E., Mander, U., and Marja, R.: Trends in the Use of Landscape Spatial Metrics as Landscape Indicators: A Review, Ecological Indicators, 28, 100–106, https://doi.org/10.1016/j.ecolind.2012.07.018, 2013.

670 Vink, A.: Land Use in Advancing Agriculture, no. 1 in Advanced Series in Agricultural Sciences., Springer-Verlag, New York, 1975.