DSM comparison: setting up a framework for an objective evaluation of DSM predictors (Indiana State case study)

T. Hengl¹, G.B.M. Heuvelink¹, J. Hempel², Z. Libohova², N. Odgers³, & J. Thompson³
¹ISRIC - World Soil Information, Wageningen, the Netherlands
²USDA-NRCS-National Soil Survey Center, Lincoln (NE), USA
³West Virginia University, Morgantown (WV), USA

The *GlobalSoilMap.net* project requires a number of case studies with complete solutions that meet the *GlobalSoilMap.net* specifications and that can be used to demonstrate various mapping steps and product functionality. This paper outlines a framework for an objective evaluation of DSM predictors, illustrated with the Indiana State case study. This case study has been packaged into the the GSIF package for R and can now be used to demonstrate various DSM processing steps. There is a need to set up a virtual and/or physical workshop where anyone in the world could submit their DSM predictor and compete for a sort of award or professional acknowledgement. The DSM comparison should be based on at least 5 study areas in the different parts of the world. The evaluation should be based on robust criteria — accuracy, bias, model robustness, model reliability, computational burden and production costs.

1 INTRODUCTION

Soil property maps are a critical inputs for ecosystem services. There is now a need for global consistent soil maps that are geographically continuous, scalable and which includes uncertainty estimates. Digital Soil Mapping provides a number of frameworks to achieve these goals. Many countries in the world have already accepted the specifications set by the *GlobalSoilMap.net* project, so that the only remaining issues are: (a) how to generate soil property maps for large areas — which methods to use, which soil covariates? (b) how to assess, visualize and communicate uncertainty of produced maps? and (c) how to continuously update and improve them?

The *GlobalSoilMap.net* project requires a number of showcases — case studies with complete solutions that meet the *GlobalSoilMap.net* specifications — that can be used to demonstrate various mapping steps and product functionality. There is also still quite some discussion going on whether is it better to let each country produce and deliver maps using freely selected methods, or let the continental nodes (8) produce maps for the whole continents and then stitch them together, or generate maps for the whole world using standardized global methods. Does a method applicable to map the whole world exist at all? Does a DSM method exist that is maybe superior? There are still many open issues for the global DSM community to tackle.

An objective way to evaluate if one method outperforms the alternatives is to organize a DSM comparison, such as the “Spatial Interpolation Comparison” (Dubois and Galmarini 2004). This would help us review the complete field of DSM, but also allow an operational planning of the large-scale DSM projects. DSM comparison would be an useful contribution, both to science and practice of DSM.

In this paper we show some initial results of testing various soil mapping algorithms for production of soil property maps according to the *GlobalSoilMap.net* specifications. Our objective is set up a framework for an objective evaluation of DSM predictors, and then compare some common DSM techniques using ideal data sets. All results, methods and the input and output data sets presented in this paper have been documented via the GSIF (Global Soil Information Facilities) package for R.

2 METHODS AND MATERIALS

2.1 DSM spatial predictors

There are many potential candidates to generate predictions from soil samples (McBratney et al. 2011). The literature however tends to contain more names than there are actual methods (Li and Heap 2010). For
example the technique called regression-kriging (RK) has several synonyms — Kriging with External Drift (KED), Universal kriging (UK) etc — which might be based on slightly different processing steps, but then are mathematically equivalent (Hengl 2009). Sometimes the statistical theory is the same, but is implemented through a different computational framework and which hence results in different predictions.

Therefore, the first step toward organizing a DSM comparison is to inventory and group the potential candidates. In general, the DSM predictors can be classified following the scheme in Fig. 1. To evaluate differences between various predictors, one can for example run the Principle Component Analysis. Fig. 2 shows, for example, results of PCA for the meuse data sets used in the gstat package to demonstrate different processing steps.

![Figure 2: Visual comparison of spatial predictors for Zinc concentration using the Principal Component Analysis: zinc.glm (GLM regression), zinc.lm (multiple regression), zinc.rk (regression-kriging), zinc.tr (trend surface), zinc.ok (ordinary kriging) and zinc.id (inverse distance interpolation). In this case, the inverse distance interpolation and GLM predictions seem to be the least correlated.](image)

Figure 2: Visual comparison of spatial predictors for Zinc concentration using the Principal Component Analysis: zinc.glm (GLM regression), zinc.lm (multiple regression), zinc.rk (regression-kriging), zinc.tr (trend surface), zinc.ok (ordinary kriging) and zinc.id (inverse distance interpolation). In this case, the inverse distance interpolation and GLM predictions seem to be the least correlated.

2.2 The DSM comparison criteria
For any comparison to be fair, it needs to be transparent and open, relevant for the practice of DSM, and based on a diversity of pre-specified criteria. From the application point of view, the following six performance criteria seem to be especially interesting for a DSM comparison (Hengl 2009):

1. **The mapping accuracy** — The standardized RMSE at control points, which also shows the amount of variation explained by the predictor expressed in % (0–100). The more variation explained by the model, the better the predictor.

2. **The bias** — The mean error i.e. the accuracy of estimating the population parameters (mean and standard deviation). The lower the bias the better the predictor.

3. **The model robustness** — Also known as the model sensitivity. This can be measured as the number of cases in which the algorithm completely fails (see Fig. 3). The lower the number of cases where the algorithm fails the better the predictor.

4. **The model reliability** — RMSE between the actual and predicted model error. This shows how good is the model in estimating the prediction error i.e. how accurate is the prediction variance considering the true mapping accuracy. The lower the RSME the better the predictor.

5. **The computational burden** — The time (in seconds) required to complete predictions. The less time required the better the predictor.

6. **The production costs** — The costs required to run DSM predictor. The lower the costs, the better the predictors.

From the six criteria listed above, one could derive a single composite measure or the mean ranking that would then allow us to select ‘the optimal’ predictor for any given data set. This is highly controversial as not all criteria are equally important. In addition, costs of running DSM are not equal in different parts of the world. It is also often hard to measure criteria such as the model robustness and production costs.

To avoid some controversial factors such as the production costs, computational burden and the robustness, we can exclude them from the experiment design. Here are some ideas on how to design the DSM comparison experiment so it provides more robust results:

- **Output products must be specified in detail** — target variables, prediction locations, support size, delivery format etc.

- **The production costs and the total survey costs need to be kept constant**, i.e. there should be no distinction between the basic costs and added costs. The total costs need to be standardized to global measures and expressed in e.g. US Dollars.
The comparison of different methods should be run on the same computer with the same operating system and system variables, i.e. using the same hardware.

The cross-validation needs to be done in a standardized way i.e. using exactly the same technique.

All potential candidates should ideally be analytical techniques that can be automated and re-run as many times are required. Because cross-validation includes random subsetting of the point data, it would be advisable to repeat cross-validation even >100 times to reach stable outputs.

The RMSE i.e. the amount of information explained by the model, which is the possibly the most relevant performance criteria, can be further used to derive the composite measures of the mapping success such as for example the mapping efficiency (Hengl et al. 2012):

$$\theta = \frac{X \cdot \text{RMSE}}{A} \quad [\text{EUR} \cdot \text{km}^{-2} \cdot \%^{-1}]$$ (1)
where \( X \) is the total costs of a survey, \( A \) is the size of area in \( \text{km}^{-2} \), and \( \text{RMSE}_r \) is the amount of variation explained by the spatial prediction model.

2.3 The Indiana data set

The NRCS (USDA) maintains and distributes one of the most detailed and most complete soil survey databases in the world. Almost whole country has been mapped at scales 1:12,000–1:24,000 and there are several detailed harmonized national soil profile and polygon databases. For example, Soil Survey Geographic database (SSURGO) maps consists of about 3000 individual soil surveys covering most of USA. These are possibly the most detailed soil survey reports for a continent-size country in the world.

USDA’s/USGS soil data is also suitable for testing mapping algorithms also because of the US Government’s open data sharing policy: the complete soil survey data is available publicly and can be downloaded from the Soil Data Mart. Within USA, Indiana state has been commonly used to produce soil property maps using the state-of-the-art DSM techniques. This can be consider one of the soil-data richest areas in the world.

We have prepared this data for a public use and testing of various DSM methods. The data set comprises:

NCSS — A subset of the National Cooperative Soil Survey National Cooperative Soil Characterization Database\(^1\) containing 2386 soil profile observations for Indiana.

NASIS — A subset of the National Soil Information System\(^2\) (NASIS) containing 489 representative profiles.

hoosier — A number of covariate layers for the southern Indiana State, including MODIS EVI and LST products (year 2011), Land Cover map\(^3\), surface and bedrock geology maps, SRTM DEM, and soil mapping unit keys from the generalized (STATSGO) national map. This data set also contains 28 taxonomic membership maps (0-1) derived as percentage of cover by soil great groups (e.g. "Epiaqualfs") from the Soil Survey Geographic database\(^4\) (SSURGO; at nominal scale 1:24k). The original maps are available at various resolutions from 100 m up to 1 km resolution.

These dataset are now built-in into the GSIF package, and all examples shown in this paper can be reproduced by loading the GSIF package (R). To get a detailed description of each layer and method used please refer to the GSIF package documentation\(^5\).

The following section shows some preliminary results for mapping some key soil properties using data.

3 RESULTS

3.1 Mapping accuracy

We have focused on mapping some basic soil properties, following the GlobalSoilMap.net specifications: depth to bedrock, organic carbon, pH, clay, silt, sand, and coarse fragments. We generated predictions of these properties to 2 m depth (where possible) with data reported for 6 depth intervals of 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm and 100–200 cm. The results shown in Fig. 4 refer to 1 km grid resolution; higher resolution imagery can also be obtained by loading the GSIF package.

![Figure 4: Predicted pH in H2O for topsoil (5 cm depth) using regression-kriging and the STATSGO soil polygon map in the background.](http://gsif.r-forge.r-project.org/)
land surface temperatures seem to have a high potential for DSM and GSM applications as they were correlated with soil textures (R-square in the range 0.241–0.455) and pH (R-square 0.244). MODIS products, publicly available high-temporal resolution remote sensing images (Savtchenko et al. 2004), are of major importance for the global soil mapping initiatives and this case study shows that they might well be one of the most important soil covariate layers for global soil mapping applications.

The results of cross-validation (5–fold), however, warn us that the soils are highly complex and often difficult to map, even in such data rich areas. Our models are rarely able to explain more than 50% of variability in the target variable. This makes it also difficult to claim that any of the predictors is ready for operational mapping (if such is possible in this study area at this level of detail at all).

3.2 Lessons learned and next steps
The results from the Indiana case study indicate that there can be quite some differences between maps produced by polygon averaging and e.g. hardcore geostatistics (Fig. 4). Note also that the results of comparison in this paper are based on cross-validation only. We are aware that the points in the National Soil Characterization Dataset have been collected using the concepts of free survey, hence it is possible that the sampling bias also propagates to the cross-validation.

Table 1 lists some common predictors used by the DSM teams in the world. Here two things need to be emphasized. First, not every DSM method one can think off can or should be considered for a DSM comparison. It should at least satisfy some basic criteria. Here are some important points:

• Each spatial predictor should be described in detail so that anyone can repeat all steps and come to the same result. Ideally, a computer data processing script or a workflow describing all steps should be submitted together with the produced data outputs.

• A spatial predictor should be based on statistical or mathematical theory; i.e. it should be an adaptation of some technique that is described in the literature. If it can not be formalized in mathematical syntax, then it probably can not be implemented as an automated process.

• Each spatial predictor should produce the same predictions, or at least results that are statistically not significantly different from each other, when repeatedly applied to the same inputs.

Secondly, not all techniques are not truly independent from each other. Although many mappers may not realize it, many non-spatial techniques can be viewed as a special case of regression-kriging and/or its variants. Consider for example a technique commonly used to generate predictions of soil properties from polygon maps — weighted averaging. It can be shown, in fact, that the weighted averaging per unit is just a special version of regression-kriging where spatial autocorrelation is ignored (assumed non-existent), and where all covariates are in principle binary variables (mapping units). This illustrates that there are not so many original statistical techniques at the first place.

4 CONCLUSIONS
This paper introduces some ideas and suggest some major discussion points to design a platform for a Digital Soil Mapping comparison. Our next objective is to define 4–5 international case studies, prepare and document the data and plug it into a package. We hope to get several local DSM teams involved and convince them to collect completely independent validation data (e.g. 100–150 points sampled using some design based sampling), and then use only the such truly independent, standardized soil profiles for validation purposes. This would make the whole Olympics ever more interesting, and certainly more convincing.

There is a need to set-up a virtual and/or physical workshop where anyone in the world could submit their DSM predictor and compete for a sort of award or professional acknowledgement — a sort of “The DSM olympics”. The Indiana show case and the GSIF package are now available publicly and anyone can consider using it to test new DSM algorithms. We hope to add several more data sets to this package in the coming months, and then open the door for a more representative comparison of the DSM predictors.

An ideal environment to implement DSM comparison is probably some scripting language such as Matlab, Python and/or R. In this paper we show the results of using R as a programming platform because R, in comparison with other listed scripting environments, has the most extensive support for processing spatial data and the largest number of geostatistical functions (Bivand et al. 2008). Although software is irrelevant for DSM quality, open source software such as R is fully transparent, so that anyone can analyze the method and check all steps.

The main motive to run a DSM comparison is to select or recommend the optimal procedure for global soil mapping. If the winner (i.e. the ‘killer’) method can be automated, it could potentially be used to produce consistent soil property maps for various continents while soil mappers sleep (automated DSM via server network). Unfortunately, access to the primary soil data (soil field observations and laboratory mea-
Table 1: Overview of some spatial predictors of interest for the DSM comparison.

<table>
<thead>
<tr>
<th>Spatial predictor</th>
<th>Code</th>
<th>2D data</th>
<th>3D data</th>
<th>Variogram modeling</th>
<th>Soil mapping units</th>
<th>Remote sensing data</th>
<th>Multiscale methods</th>
<th>Global model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polygon averaging</td>
<td>DSMPAVG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Correlation with soil taxa</td>
<td>DSMCTAX</td>
<td>✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stratified kriging</td>
<td>DSMSOK</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regression trees</td>
<td>DSMREGT</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regression-kriging 2D</td>
<td>DSM2DRK</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regression-kriging 3D</td>
<td>DSM3DRK</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Global Multiscale RK 3D</td>
<td>DSMMNRK</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Measurements) in many countries is still limited, and often clouded by the administration and unclear data licenses. This will remain the main problem of applying the DSM methods globally.

REFERENCES


