IMPLICATIONS OF DIGITAL SOIL MAPPING FOR SOIL INFORMATION SYSTEMS

G.B.M. Heuvelink[1], D.J. Brus[1], F. De Vries[1], R. Vašát[2], D.J.J. Walvoort[1], B. Kempen[1] and M. Knotters[1]
[1]Alterra, Wageningen University and Research Centre, Wageningen, The Netherlands, gerard.heuvelink@wur.nl
[2]Department of Soil Science and Soil Protection, Czech University of Life Sciences in Prague, Czech Republic

Abstract
Soil information systems (SIS) as we know them store information about the soil as point observations and maps. This may seem obvious and sensible, but in fact in the present time it is a suboptimal way of storing soil information. Nowadays soil maps are often derived using digital soil mapping models and this offers the possibility to store the models used to derive the maps, instead of the maps themselves. This short paper lists the advantages of storing models instead of maps and illustrates the approach with examples from SIS+, a prototype developed for the Netherlands.

Keywords: Soil Information System, geostatistics, database, accuracy, R

1. Introduction

Soil information systems (SIS) were introduced in the 1970s and have proven very instrumental for the disclosure of soil data and soil information. These systems stimulated the transformation of analogue soil data to digital form and facilitated their storage. The database functionality of a professional SIS provides flexible query capabilities such that users can easily search and extract point data and maps from the database. Nowadays, many of the national and institutional SI systems around the globe can also be accessed from the internet. Current SI systems serve a clear goal and have a large group of satisfied users. However, they also have limitations. Some of these limitations can be overcome by making use of the fact that nowadays, many soil maps are produced using digital soil mapping techniques.

2. SIS+: storing models instead of maps

The rapid development and application of digital soil mapping techniques during the past decade calls for a next generation of SI systems. Such a system, named SIS+ for short, stores (pedometric) models instead of maps. For instance, rather than storing the result of a kriging interpolation, it is more sensible to store the source data and kriging parameters, such that a map can be delivered on demand. Thus, SIS+ still needs the source data (i.e. observations) and explanatory variables, such as a legacy soil type map, a DEM or remote sensing imagery, all of which can be read from a conventional SIS or geodatabase, but it no longer contains resultant maps of soil type and soil properties.

Storing models instead of maps has several important advantages:

1. It gives much more flexibility in terms of the spatial and temporal extent, resolution and support of the requested map. Users can log on to SIS+, submit their specific request and

a customized map will be produced on the fly, provided that a model and sufficient data are available. Note that flexibility with respect to spatial extent, resolution and support applies both to the horizontal as well as the vertical dimension.

2. It saves storage capacity, because the number of different maps that can be produced (i.e. different soil property, resolution, extent, support, depth or time period) is extremely large and it is practically impossible to store all of these in a conventional SIS.
3. Many digital soil mapping approaches make use of (geo)statistical models that not only produce a map but also quantify the associated uncertainty, which is indispensable information in today's environmental policy. Hence SIS+ also includes methods for stochastic simulation of hundreds or thousands 'possible realities', such as required for Monte Carlo uncertainty propagation analyses (e.g. Goovaerts 2001). Again, it would be very difficult to store all of the simulated realities in a conventional SIS. There is also no need for that, because storing the geostatistical model and the seed of the pseudo-random number generator is sufficient to be able to reproduce the realizations when needed.
4. It enables easy updating with new data. When new data are added to the database, all that needs to be done to update an existing map is to rerun the model, which draws on all relevant data (old and new) stored in the SIS.
5. It automatically archives the way in which a map is made. The model code tells precisely how a map was obtained, which point data and covariates were used and in what way. Note that this also allows using a model that was built for a particular soil property as starting point for a model of another soil property. Thus, model building need not commence from scratch each time a new model is built.
6. It can help solve data sharing problems because it need only store models that can be applied to data that are stored elsewhere. For instance, a data owner may not be willing or allowed to transfer datasets to others. However, perhaps it is allowed that others make use of the data to create a map. In such a case, using a web-based approach, the SIS+ model can be submitted to and run on the computer of the data owner, returning only the resulting map. In other words, there is no need to physically export data. Note that web-based implementation of SIS+ also facilitates its use from all around the world. Anyone with permission rights can connect to the SIS+ host and load and run the SIS+ models on their own data, stored in their local SIS.

The principle of storing models instead of maps is not new. For instance, the INTAMAP project (Pebesma et al. 2010) developed a web-based interpolation service for real-time automatic interpolation of environmental variables. The service can be approached from any place around the world, data can be submitted together with an interpolation request, after which an interpolated map and associated accuracy map are produced in real time in a fully automated fashion (see http://www.intamap.org/tryIntamap.php).

The use of SIS+ can vary with the expertise of the user. Roughly, three categories of users are envisaged:

1. Plain users that only work with ready-made models (recipes) for which only a few parameters need to be set (e.g. which soil property, extent, resolution and support).
2. More advanced users that in addition can work with the system in an interactive way and can slightly modify existing models. For instance, based on the results of an exploratory data analysis tool contained in SIS+ these users might decide to transform the data or

covariates and run the model on the transformed data. These users are expected to be able to adjust the model code accordingly.
3. Fellow model developers that help extend the library of recipes stored in SIS+ with entirely new functions. For instance, these users might wish to implement new statistical models published in the scientific literature.

Many more issues and opportunities will turn up once the development of SIS+ progresses and matures. Indeed, perhaps the best way to discover these opportunities is to simply build an operational SIS+ and learn from that. In the Netherlands, we have explored the possibilities of building a prototype SIS+ since 2007 (Brus and Heuvelink 2007). More recently, we have implemented a first version of SIS+ that as yet can only create maps of a limited number of continuous soil properties (Brus et al. 2010). The project is ongoing and aims to have a comprehensive prototype ready by the end of 2014. The next section briefly describes the current version of the Dutch SIS+ and illustrates its functionality with a few examples.

3. Examples from the Dutch prototype SIS+

The Dutch prototype SIS+ is implemented in the R language for statistical computing (http://www.r-project.org/). It automatically loads data from the existing Dutch SIS (www.bodemdata.nl) and stores geostatistical models that map soil properties from data and covariates as R functions. An important feature of the prototype SIS+ is that it also quantifies the accuracy of the resulting maps. It is composed of six stages:

1. *Importing data from the Dutch Soil Information System*. A function was developed that reads data from the SIS without requiring that users know the SQL language. Information such as soil property, time frame, extent and depth are provided as function parameter values.
2. *Data preprocessing*. Among others, this stage converts observations from soil horizons at locations to values of the soil property for an arbitrary soil layer (with arbitrary top and bottom values chosen by the user). For soil properties such as organic matter and clay content, information on horizon bulk density is employed to calculate the weights with which soil horizon values contribute to the average of the soil layer, because this type of soil properties are not related to soil volume but to soil mass.
3. *Exploratory data analysis*. Basic statistics such as the mean, variance, minimum, maximum and median can be computed to obtain insight in the distribution of the soil property. Histograms and Q-Q plots can be constructed to evaluate whether the data depart from the normal distribution. These plots can also be used to evaluate the effect of data transformation. Furthermore, a set of boxplots describing how the soil property distribution varies with soil type may be useful when taking a decision about the structure of the geostatistical model. If appropriate, outliers can also be removed from the data set during this stage.
4. *Building models of spatial variation*. The current version of the Dutch SIS+ is restricted to geostatistical models, which include simple, ordinary, regression and cokriging models. Change of support can be defined using block kriging. The most important covariate currently used is a generalized soil map of the Netherlands, which distinguishes 21 soil types (Wösten et al. 1988).
5. *Geostatistical (co)prediction and (co)simulation*. Standard R-libraries such as *gstat* (Pebesma 2004) for geostatistics and *maptools* for GIS operations are used.

6. *Exporting resulting maps*. Resulting raster maps can be exported in a variety of formats, such as graphical formats, GIS layers, ASCII files and database tables.

Figures 1 to 3 show results of the six stages for mapping the soil pH at point support at depth 0–25 cm, using a regression kriging model. Figure 4 shows final results of mapping the clay content at 0–25 cm with regression kriging for the entire Netherlands and for two subareas. In all three cases exactly the same model and data were used, only the extent differed. Figure 5 shows maps of the organic matter content of the topsoil, using either a kriging or cokriging model. These examples are merely shown for illustration, details are given in Brus et al. (2010). However, the important message is that SIS+ stores the geostatistical models with which these maps are made. These examples show the ease with which a given model can be extended to other depths or soil properties.

```
# 02 Data extraction - PFB


# Connect to database
channel <- odbcConnect(dsn = "deltaBIS", uid = "BISUSER", pwd = "[redacted]")

# Fetch all records from V_pfb_deltabis view
pfb <- sqlFetch(channel = channel, sqtable = "V_PFB_DELTABIS")

# Disconnect from database
odbcClose(channel = channel)
rm(channel)
```

Figure 1. Snapshot of R code used to extract pH observations from the Dutch SIS.

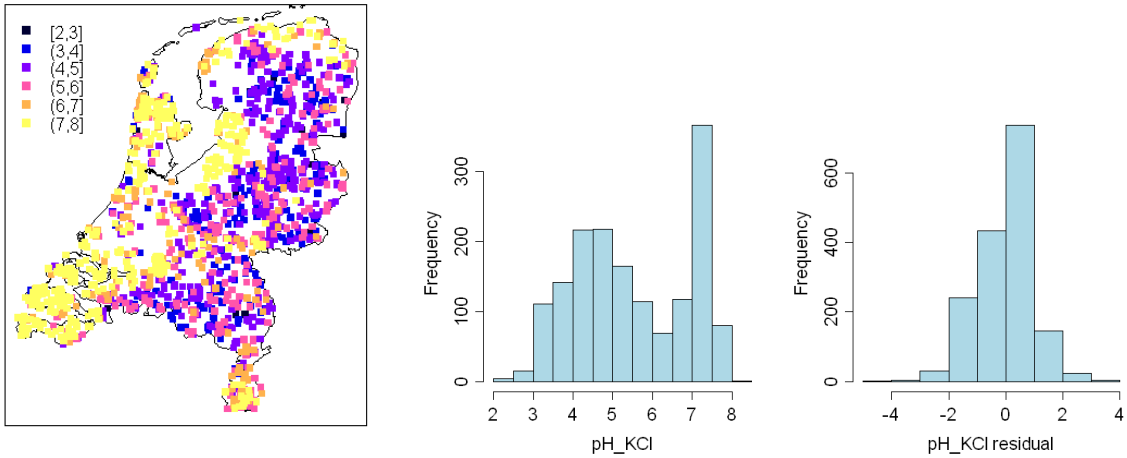

Figure 2. Spatial distribution and histograms of soil pH observations (n=1621) for 0–25 cm depth. Bimodality is caused by the large difference between soil pH in the poor sandy soils (low values) and the loess, clay and calcaric sandy soils (high values). Histogram of pH residuals shows that the generalized Dutch soil map explains much of the spatial variation.

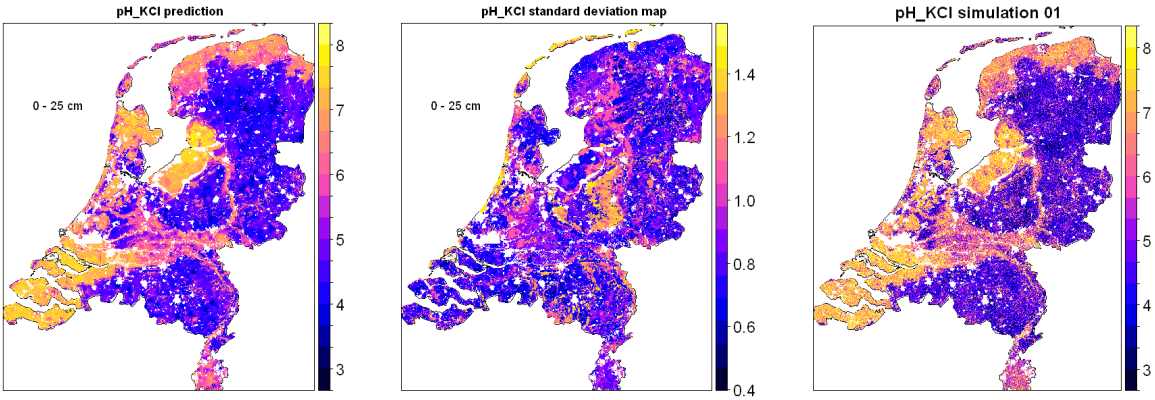

Figure 3. Maps of the predicted (left), prediction error standard deviation (centre) and example realization (right) of the soil pH for the Netherlands at 0–25 cm depth.

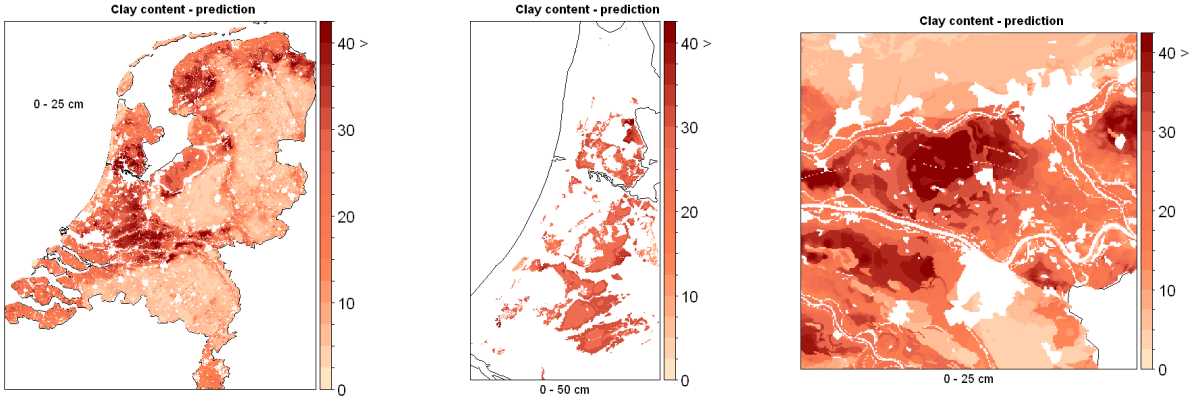

Figure 4. Kriging maps of the clay content (mass %) at 0–25 cm depth for the Netherlands (left), at 0–50 cm depth for a coastal subarea (centre) and at 0–25 cm depth for a subarea in the east of the Netherlands (right).

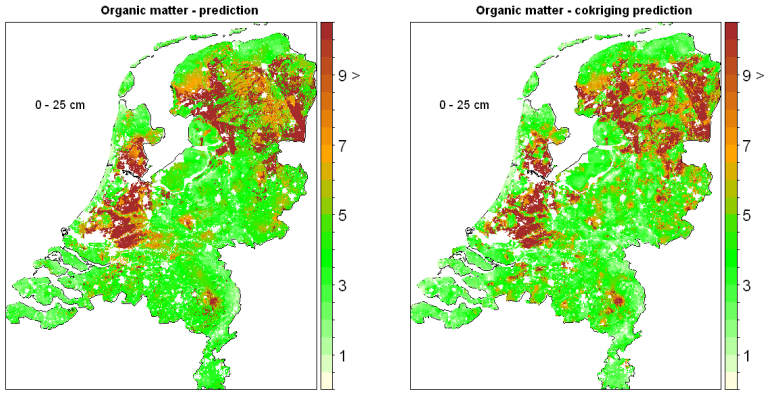

Figure 5. Maps of the organic matter content (mass %) for the Netherlands at 0–25 cm depth obtained with regression kriging (left) and regression cokriging (right), taking correlation with organic matter content at 25–50 cm and 50–100 cm into account.

4. Conclusions

- The development of a SIS+ that stores pedometric models instead of maps has many important advantages that can help overcome the limitations of conventional SIS.
- SIS+ does not replace conventional SIS but needs it for delivery of point soil data and basic soil maps. It also needs other geodatabases for delivery of covariates. In turn, maps produced by SIS+ that are frequently used can be transferred to and stored in the conventional SIS.
- In the long term, SIS and SIS+ may be integrated into one system, but during the development stage it is better that these are separate systems that communicate through data exchange.
- The development of SIS+ can make use of current developments in automated mapping, particularly when it concerns web-based implementations.
- Experiences so far with the development of a prototype Dutch SIS+ are very positive. The current implementation can already automatically download data from the Dutch SIS and map multiple continuous soil properties for arbitrary depths, extents, resolutions and supports.

5. References

Brus, D.J. and G.B.M. Heuvelink, 2007. Towards a Soil Information System with quantified accuracy. Three approaches for stochastic simulation of soil maps. WOt report 58, Wageningen, The Netherlands.

Brus, D.J., R. Vašát, G.B.M. Heuvelink, M. Knotters, F. de Vries and D.J.J. Walvoort, 2010. Towards a Soil Information System with quantified accuracy. A prototype for mapping continuous soil properties. WOt report, Wageningen, The Netherlands (in press).

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3–26.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers and Geosciences 30: 683–691.

Pebesma, E.J., D. Cornford, G. Dubois, G.B.M. Heuvelink, D. Hristopoulos, J. Pilz, U. Stöhlker and J.O. Skøien, 2010. INTAMAP: the design and implementation of an interoperable automated interpolation web service. Computers & Geosciences (in press).

Wösten, J.H.M., F. de Vries and J. Denneboom, 1988. Generalisatie en bodemfysische vertaling van de bodemkaart van Nederland, 1 : 250.000, ten behoeve van de PAWN-studie. Technical Report 2055, Stichting voor Bodemkartering, Wageningen, the Netherlands (in Dutch).