PREDICTIVE QUALITY OF THE BELGIAN SOIL SURVEY INFORMATION[1]

Van Meirvenne M.

University Gent, Dept. Soil Management, Coupure 653, 9000 Gent
Tel. : 09-2646056, Fax : 09-2646247, e-mail : Marc.Vanmeirvenne@rug.ac.be

ABSTRACT

Due to the continuous and sometimes highly variable nature of soil properties and the relatively limited number of point observations compared to the investigated area, soil survey information is predictive rather than descriptive. To quantify the predictive quality of this information, several parameters are available like : map purity, map bias and intraclass correlation coefficient. The Belgian soil survey information was characterized using these parameters and twelve different prediction algorithms were compared. Both soil properties taken into account into the mapping legend or not were used to illustrate the different aspects of the predictive quality of the Belgian soil survey information.

Keywords : soil survey, map quality, predictive quality

1. INTRODUCTION

General-purpose soil surveys, like the Belgian soil survey, aim to describe the spatial distribution of a selected number of soil properties, be it directly (e.g. by using a parametric legend) or indirectly (through the use of a soil classification system). The outcome of such a survey is a soil map and its complementary documentation. These are then used as a source of information on the soils inside the mapped area. However, we have to be aware that there is a difference in viewpoint towards soil survey information between the soil surveyor and the user of this information.

For the soil surveyor, the soil survey information is an end-product of his interpretation of the soil-landscape-vegetation-geology-climate-landuse interaction, based on research, field work and laboratory analyses. All procedures which are followed (field observations, map generalization etc.) are oriented towards this goal, i.e. to produce a soil map. Therefore, for the soil surveyor, the soil survey information is *descriptive* in nature.

For the user of soil survey information, this information represents a starting-point of his analysis (in or outside a Geographical Information System). He wants to obtain information on soil properties at every location within his study-area. So for the user soil survey information is *predictive* rather then descriptive. This difference in approach may lead to differences in expectations concerning the usefulness of soil survey information.

Since the soil surveyor didn't visit all locations the soil map contains uncertain information. Hence the user needs to be informed about the *predictive quality* of the information he is using.

[1] Published in : Pedologie-Themata **5**: 71-74 (1998)

It is the aim of this paper to bring an overview of research results obtained over the last years at the Dept. Soil Management of the University Gent, concerning the predictive quality of the Belgian soil survey information. But first an brief presentation of the available soil survey information is given.

2. BELGIAN SOIL SURVEY INFORMATION = SOIL MAP + SOIL DATABASE

2.1 Soil Map (1/20 000)

The soil map of Belgium contains a parametric legend of three classified soil properties : soil texture class, soil drainage class and type of soil profile development. Occasionally other properties were indicated (like the presence of an undeep subsurface substrate), but we will not consider these additions here.

It must be realized that the direct predictive quality of the soil map is limited, not only because this information is qualitative, but also due to wide class interval definitions. If we compare the Belgian soil texture triangle with the internationally accepted USDA soil texture triangle (Fig. 1) then we see that the Belgian soil texture triangle contains 7 textural classes, while the international triangle has 12 classes. Consequently we can expect wider class definitions, and hence a lower predictive quality, of the Belgian textural classes. E.g. texture class E ("clay") is defined to contain a sand content between 0 and 82.5 %, for a variable which is limited between 0 and 100 % ! Clearly a user who wants to know a prediction of the sand content at a location mapped as texture class E receives little informative value.

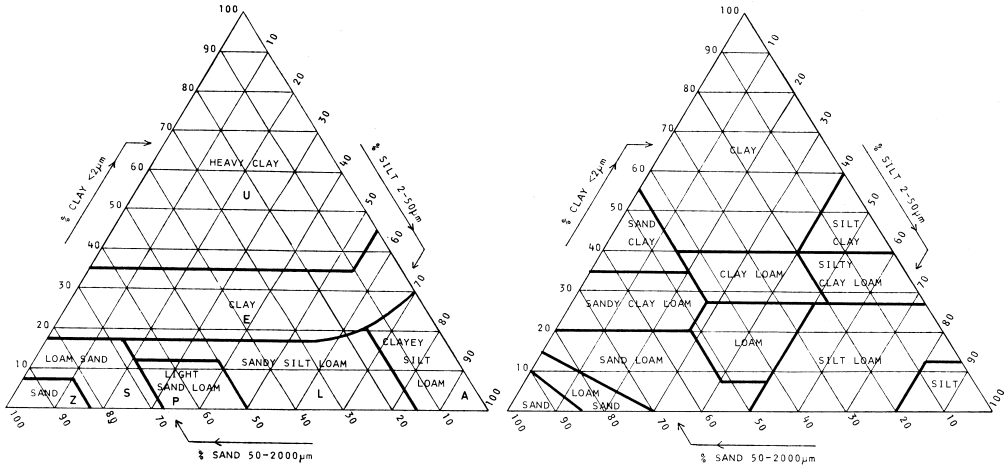

Figure 1 : Belgian (left) and internationally used USDA (right) soil texture triangle (from Verheye and Ameryckx, 1984).

2.2 Soil Database

Besides the soil map, a detailed analytical soil database with a spatial resolution of about 1 observation per 1.5 km² on a national scale is available, which was created between 1947 and the early seventies. These results were published as addenda to the soil map sheets. Through the common classification these soil profiles can be linked with the soil map to provide quantitative information.

Some limitations concerning the usefulness of this database for the present-day user are related to the outdating of information (like organic matter, pH, A-horizon thickness,…) and modifications in the procedure to determine soil properties. E.g. before 1961 organic matter (OM) was calculated as the organic carbon determined according to the Walkley & Black method (OC_{WB}) multiplied with 1.724 (so organic carbon (OC) = (OM/1.724) x (4/3)). After 1961 this changed to : OM = OC x 2 = OC_{WB} x 2.667 (see addendum of map sheet 80/E). This leads to confusions still today.

However, a major drawback of the database is the unclear sampling strategy which was followed. This sampling strategy differed over the years and different soil surveyors followed different approaches. For some map sheets typical ("representative") profiles were sampled, for others the a-typical situation (like Tertiary outcrops) were targeted. Mostly a more or less even spatial distribution was aimed, but some map sheets have been sampled according to a spatially clustered scheme (Fig. 2). As a result, the user must realize that the soil database might contain biased information with different degrees of spatial autocorrelation. By no means, this database can be regarded as a probabilistic sample, thus further statistical processing should be done with care (like the use of distribution parameters, like the mean or variance, for predictive purposes).

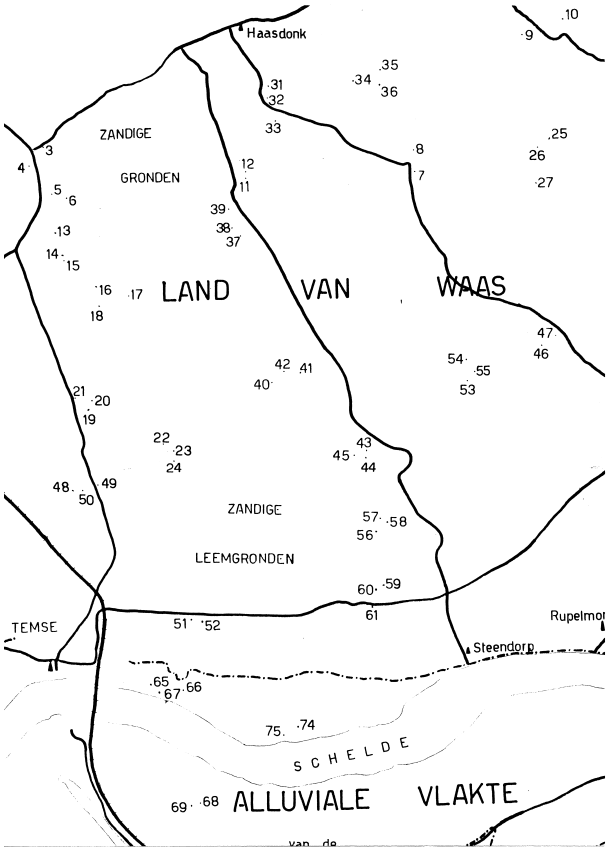

Fig. 2 : Localization of soil profiles of a part of map sheet Temse 42/E according to a clear spatially clustered sampling configuration, mostly of 3 observations per cluster (numbers represent sample no.).

3. MEASURES OF PREDICTIVE QUALITY

3.1 Map purity and bias

The *map purity* is defined as the number of observations of a given soil property located on a soil map which match the definition of the unit in which they are located (Dent & Young, 1981). The result is conveniently expressed as a percentage. This parameter can only be verified for those soil properties which are included in the mapping legend, but it has the advantage to work for both quantitative and qualitative information. It is evident that we can use only observations obtained independently from the soil survey to check its map purity. Formerly it was stated (Soil Survey Staff, 1951) as a quality goal for soil maps that they should have a purity of 85 % (so 15 % deviation from the mapping legend was considered to be acceptable).

A *bias* is a statistical term indicating a systematic deviation in a selection. In the context of a soil survey this means that there was a tendency to systematically over- or underestimate a soil property.

As an example we will use two datasets on soil texture which were largely obtained independently from the soil survey. The first consists of 118 texture analyses in the Polder area of north E.-Flanders (Watervlietse polders) (Van Meirvenne & Hofman, 1992) and the second dataset contains 326 texture analyses taken in the Sandy Loam region in W.-Flanders by J. De Smet and K. Scheldeman in the frame of research projects concerning the delineation of phosphate saturated soils. The latter data were processed by S. Depuydt (1994). These textural analyses were classified according to the Belgian texture triangle and located by a point-in-polygon search on the soil map. To allow a comparison, the Belgian texture classes (Fig. 1) were ranked as follows : Z>S>P>L>A>E>U. Figure 3 shows the results.

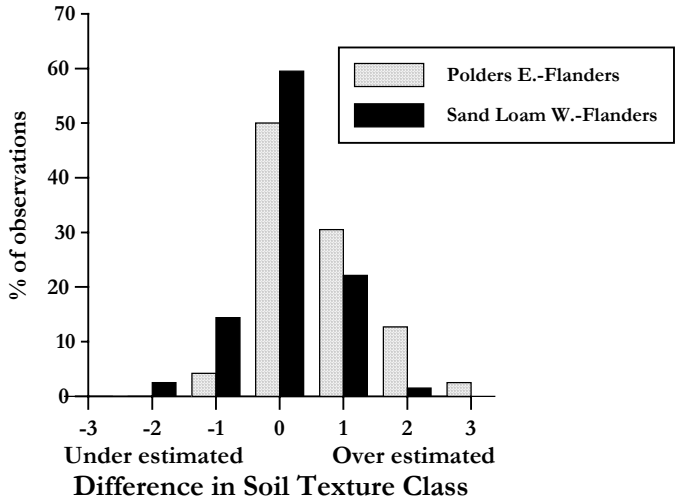

Figure 3 : Map purity and bias of soil texture in two study areas.

The map purity in the Sandy Loam region (no difference in soil texture class) reached 59.5 %, with almost equal parts being underestimated (16.9 %) or overestimated (23.6 %) by 1 or 2 textural classes. So only a slight bias towards overestimation could be found. However, in the polder area, the results were different. First, the map purity dropped to 50 % (probably due to the larger heterogeneity of this region compared to the Sandy Loam region). Second, a strong bias was observed : only 4.2 % of the observations were underestimated by the surveyor, the

remaining 45.8 % were overestimated (30.5 % by 1 class). So the soil surveyors systematically mapped the soils of this region to be more clayey than in reality.

Clearly, the 85 % map purity set forward by the USDA is unrealistic, even for a parametric legend with wide class definitions, like in the case of the Belgian soil map.

3.2 Intraclass correlation coefficient

If one wishes to evaluate the usefulness of a soil map to predict a variable which was not taken into account by the soil map legend, the map purity cannot be used. An alternative is given by the intraclass correlation coefficient (Webster & Beckett, 1968). It is beyond the scope of this paper to go into the mathematical details concerning this parameter, these can be found in Webster & Oliver (1990).

In short, the intraclass correlation coefficient r_i is a measure of the part of the total variability of a variable which is taken up by the soil map classification. In this sense its interpretation is similar to R^2 in regression analysis. If all soil classes would be absolutely uniform then $r_i \approx 1$ and we would have a perfect map. The other extreme situation is when the differences between soil map classes are not significant. Then $r_i \approx 0$ and we would have a useless map for the variable considered. A given map could be found to be more informative for one variable than for another, or different maps could be compared to select the most suitable one.

We investigated the intraclass correlation coefficients for a number of soil properties of the soil map units of the Sand Loam region of W.-Flanders. The same 326 independent samples of texture and organic matter (OM) were used and in addition 612 samples of pH-H_2O, ammoniumoxalate extractable Al and Fe and the Phosphate Sorption Capacity (PSC - being defined as (Al + Fe)/2) were used. More details are given by Depuydt (1994). The results are given in Table 1.

Table 1 : Intraclass correlation coefficient (r_i) of a number of variables for the Sandy Loam region of W.-Flanders

Variable	r_i
Clay	0.38
Silt	0.53
Sand	0.61
OM	0.06
pH	0.09
Fe (oxalate)	0.36
Al (oxalate)	0.28
PSC (oxalate)	0.09

For the soil variables taken into consideration by the map legend, like silt and sand fractions, the r_i's are relatively large, i.e. 0.5-0.6. The clay fraction is mostly more variable due to pedogenetic processes (migration, dissolution by acids, neogenesis, …) which act more actively on clay particles than on silt or sand. As a result the r_i drops to 0.38 for clay. For OM and pH the soil map is nearly useless ($r_i < 0.1$). Fe and Al are strongly related with the parent material and thus with the textural composition, hence its intermediate r_i's of 0.36 and 0.28 resp. However, when the latter two variables are combined into the PSC, their uncertainties are also combined and the r_i drops to 0.09. So also for the PSC the soil map is unsuitable as predictor. According to published values (Webster & Oliver, 1990) we can conclude that the soil map is a reasonably good information source for textural fractions and strongly related variables, however for chemical properties its predictive quality is very poor.

3.3 Prediction algorithms

Once the soil survey information is found to be useful, a choice has to be made about which prediction algorithm is going to be used. To investigate this question we need an independent test dataset to evaluate the predictions based on the soil survey information. Such a dataset became available during the second survey of the Sea Polders (W.-Flanders) by De Leenheer & Van Ruymbeke (1960). As a result we had two datasets (Figure 4) :

- A "prediction" dataset obtained from the National soil survey containing analytical results of 697 soil profiles.
- A "validation" dataset of 123 soil profile observations taken from De Leenheer & Van Ruymbeke (1960).

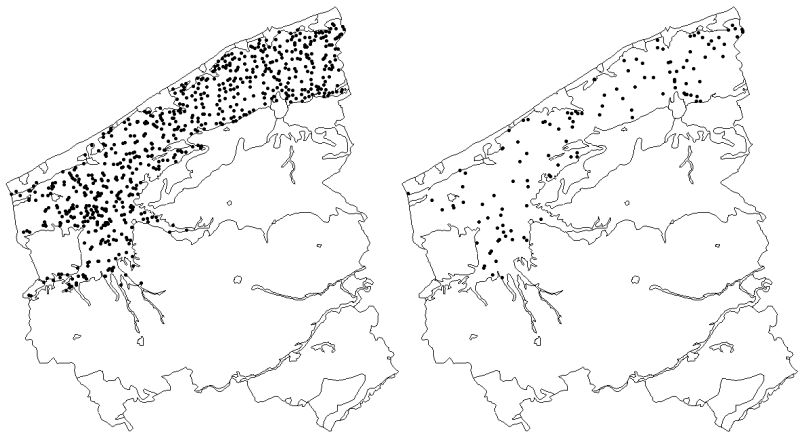

Figure 4 : Province of West-Flanders with the Sea Polders study area with localization of the 697 observations of the "prediction" dataset (left) and the 123 observations of the "validation" dataset (right)

The variables which were considered were topsoil clay and the background value of lead (BPb in mg kg^{-1} Dry Matter). The latter was defined by the Soil Sanitation Decree of the Flemish Government (VLAREBO, 1996) as depending on the clay and OM content of the soil (both in %) :

$$BPb = 33 + 0.3*clay + 2.3*OM.$$

Clearly the first variable represents a variable which was taken directly into consideration during the soil survey, where the second variable is only indirectly related to the criteria used during the survey. However it is an example of a present-day environmental application of the use of soil survey information.

Twelve prediction algorithms were used to predict these two variables at the locations of the "validation" dataset using the "prediction" dataset (more details can be found in Rienckens, 1997). The applied prediction algorithms can be combined into three groups :

1. Algorithms independent from the soil map :
    - Global Mean of entire region (GM).
    - Nearest Profile (NP).
    - Mean of 5 nearest profiles (M5).
    - Mean of all profiles within a radius of 3 km (M3km).

- Inverse squared Distance interpolation (ID).
- Point Kriging (Journel & Huijbregts, 1978) interpolation using the nearest 8 observations and a variogram of clay or the background value of Pb determined on the entire "prediction" dataset (K).

2. Algorithms depending on the generalized soil map with 3 strata :
   - Mean of stratum (Ms).
   - Inverse squared Distance interpolation within same stratum (IDs).
   - Point Kriging interpolation using the 8 nearest observations within same stratum and a stratum specific variogram (Ks).
3. Algorithms depending on the mapping units of the 1/20.000 soil map :
   - Mean of mapping unit (Mm).
   - Nearest Profile within the same mapping unit (NPm).
   - Inverse squared Distance interpolation within same mapping unit (IDm).

Point kriging could not be used in combination with the soil map due to the limited number of observations within the map units. Solutions for this limitation exist (e.g. Heuvelink, 1996) but were not taken into account here.

Since we predicted the values of both variables at the locations of the "validation" dataset, we compared predicted and measured values. This was done by calculating the Pearson correlation coefficient (r) and two indices :

1. the Mean Square Error (MSE) :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Z(\mathbf{x}_i) - Z^*(\mathbf{x}_i)\right)^2$$

with $Z(\mathbf{x}_i)$ the measured value of Z at location $\mathbf{x}_i$ representing its coordinates in space, $Z^*(\mathbf{x}_i)$ the predicted value of the same variable at the same location and n = 123. The MSE evaluates the magnitude of the average error, and so it should be as small as possible.

2. the Standard Deviation of the Square Error (SDSE) :

$$SDSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\left(Z(\mathbf{x}_i) - Z^*(\mathbf{x}_i)\right)^2 - MSE\right|^2}$$

This index evaluates the spread of the errors, and since a homogeneous distribution of errors is preferred over a highly variable distribution, small values are aimed.

A useful way to compare the outcome of the different prediction algorithms is to plot SDSE versus MSE. The best performing algorithms are those close to the origin of the graph.

The results are shown in Figure 5 for the clay content and in Figure 6 for BPb.

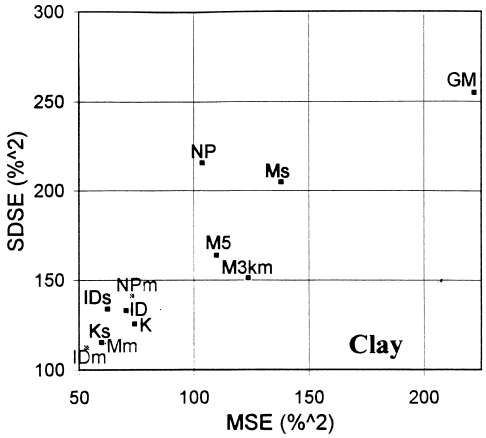

Figure 5 : SDSE versus MSE of the twelve prediction algorithms for the clay content

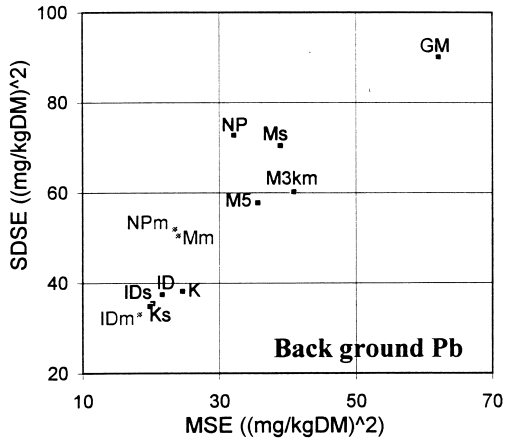

Figure 6 : SDSE versus MSE of the twelve prediction algorithms for the background value of Pb

The following points can be concluded from this validation :

- For any of the three groups of methods, the interpolation algorithms (ID and K) always performed best. Even when a detailed soil map is available, a combination with an interpolation method was found to be advantageous.
- As the detail of the map increases, so does the quality of the prediction for any given method.
- Predictions by non-interpolation algorithms based the soil map (Mm, NPm) are only acceptable when it concerns a variable which was directly related to the mapping legend.

CONCLUSIONS

To use the Belgian soil survey information, one should pay attention to the predictive quality of this information.

First, a bias might be present both in the mapped information and in the selection of the analyzed soil profiles. This limits the possibility to make probabilistic statements based on this information.

Generally, the map purity can be situated around 50 to 60 % , but this parameter is limited to mapped variables.

The intraclass correlation coefficient of mapped soil variables was found to be sufficiently large (0.4-0.6). However, for other variables, like chemical variables, low values were found indicating that for these variables the usefulness of the map is very limited.

To predict quantitative information a combination of soil survey information with an interpolation algorithm is strongly recommended. If the sampling size is small, an inverse distance algorithm could be used, otherwise geostatistical procedures, like kriging, should be used.

ACKNOWLEDGMENTS

The author wishes to thank all persons involved in the research topics reported in this overview paper : G. Boucneau, S. Depuydt, J. De Smet, G. Hofman, K. Scheldeman and K. Rienkens.

REFERENCES

De Leenheer L. & Van Ruymbeke M., 1960. *Monografie der Zeepolders.* Rijkslandbouwhogeschool Gent, 416 p.

Dent D. & Young A., 1981. *Soil Survey and Land Evaluation.* George Allen & Unwin, London, 278 p.

Depuydt S., 1994. *Bruikbaarheid van bodemkarteringsinformatie voor het voorspellen van fosfaatsorptiekarakteristieken in West-Vlaanderen.* Afstudeerwerk Universiteit Gent, 69p.

Heuvelink G., 1996. Identification of field attribute error under different models of spatial variation. *Int. J. GIS* **10**:921-935.

Rienckens K., 1997. *Bruikbaarheid van bodemkarteringsinformatie voor het inschatten van saneringsdrempelwaarden.* Afstudeerwerk Universiteit Gent, 126p

Soil Survey Staff, 1951. *Soil survey manual.* Agric. Handbook no. 18. Washington DC : Dept. of Agriculture (USDA).

Van Meirvenne M. & Hofman G., 1992. Combining GIS and geostatistics for quantitative soil texture mapping. *EGIS '92 Conference proceedings* (Eds. J. Harts, H. Ottens & H. Scholten), EGIS Foundation Utrecht, p. 1426-1433.

Verheye W. & Ameryckx J., 1984. Mineral fractions and classification of soil texture. *Pedologie,* pp. 215- 225.

VLAREBO, 1996. Vlaams Reglement betreffende de Bodemsanering. *Belgisch Staatsblad,* 27 maart 1996.

Webster R. & Beckett P.H.T., 1968. Quality and usefulness of soil maps. *Nature* **219**:680-682.

Webster R. & Oliver M., 1990. *Statistical Methods in Soil and Land Resource Survey.* Oxford University Press, 316 p.