Chapter 7

Grid-based Soil Information System*

"It is now quite possible to combine information derived from DEMs and satellite observation with profile data and numerical models of soil processes to produce a rich, predictive models of the soil to meet both the purposes of research in soil formation and landscape development and practical considerations of land suitability assessment, decision making or the review of development scenarios."

[P.A. Burrough, announcing future research in "Continuous classification in soil survey: spatial correlation, confusion and boundaries", Geoderma, vol. 77(2-4): 115-135]

*based on: Hengl, T., 2004? A hybrid grid-based soil information system based on the mixed model of spatial variation. Geoderma, in review.

7.1 Introduction

A Soil Information System (SIS), also referred to as a Soil Geographical Database (SGDB), is a commonly used term for a thematic GIS specifically designed to provide (geo)information on soils (Burrough, 1991). This is a structured digital version of soil maps and soil survey reports associated with data from laboratory analysis. A Conventional SIS consists of:

1. a polygon map, representing the soil bodies;
2. a point map, representing profile observations, and
3. attribute tables representing sampled descriptive and physical or chemical soil properties.

The polygon map is a class-type map, the classes are soil mapping units (further referred to as SMUs) and the profiles are organized into a relational database and linked to the SMUs via their coordinates or soil types (Zinck & Valenzuela, 1990). This system follows the Discrete Model of Spatial Variation (Heuvelink, 1998). The key function of a SIS is to serve the users for data retrieval, spatial queries, statistical analysis and visualisation of results. The profile data is used to make attribute or thematic maps and statistical representations by averaging the values per SMU type or soil type (Burrough, 1993a). Similarly, the SMU's can be directly linked to interpretation tables e.g. soil suitability classes. The above-described system is also referred to as the "conventional approach" to the soil mapping and has been adopted and used in most of the World today, especially at regional and national scales.

For many GIS professionals, working on data integration, a critical layer in a multi-thematic GIS, particularly when utilized in land management decisions, is soil survey information (Maclean *et al.*, 1993). For other SIS external users, such as agronomists, land use planners or civil engineers, the concepts of soil classes and soil mapping units are often harder to grasp and interpret than the land use types or vegetation types. Instead of the map of soil types, the external users are often more interested into the maps of soil interpretations (e.g. suitability for vine production) or limiting land characteristics (e.g. depth to gleying) or technical properties of the soils (e.g. texture fractions, depth to the cemented layers etc.) (Dent & Young, 1981). Moreover, modern users require soil geoinformation at increasingly finer level of detail and increasingly higher accuracy.

There are several likely reasons that conventional soil maps are unpopular among the external users. First, the concept of soil types is probably the fuzziest from all environmental sciences, as the soil bodies are hidden, often irregular or random in

distribution (Burrough *et al.*, 1997). Second, classification systems have been an object of dispute and it was not until the end of the last century that an official international classification system (FAO, 1998) was accepted. Even today, there is still a high chance that two soil surveyors, working independently in the same pit, will identify two different types of soils. Third, analytical procedures are missing in some phases of soil mapping or are not fully documented. For example, the soil boundaries are drawn by following the mental model in surveyor's head rather than by an objective procedure (Cook *et al.*, 1996). Hence, soil survey is still considered by some to be more of an art than a science (Hudson, 1992). The fourth cause of the general low confidence in the soil maps is that their operational quality, i.e. accuracy, lineage and completeness, has often been proved to be lower than expected (Marsman & de Gruijter, 1987; Burrough, 1993a). Finally, the concept of SMUs and related polygon-based organisation of SIS is not immediately suitable for multi-source data integration and quantitative environmental modelling (Ventura *et al.*, 1996). Some more recent conceptual designs of SGDBs, e.g. by Fernandez & Rusinkiewicz (1993), are often unnecessary too complex and therefore user-unfriendly for external users. This is most probably because: (a) the soil surveyors often produce multiple-component mapping units, which are harder (sometimes impossible) to organize and query and (b) SGDB use several entities at the same time (mapping units, pedons, horizons), which can be connected in several ways, thus confusing the external users.

The above-listed problems with the conventional approach have been a major inspiration for researchers in the last decade or two. In early 90's, McSweeney *et al.* (1994) laid the foundation for a new four-stage framework for modelling the distribution of soils. From then, the following two developments have shown to be especially promising: use of auxiliary or secondary data, such as terrain parameters and remote sensing images (Dobos *et al.*, 2000; McKenzie *et al.*, 2000), and use of new concepts and methods, such as continuous classification to model the soils more successfully (McBratney *et al.*, 1997). The use of auxiliary data to improve mapping of soil variables has been especially prominent in Australia (Carlile *et al.*, 2001). Also in the Netherlands, there has been a significant shift towards the quantitative methods for inventarization and utilization of soil data (Buurman & Sevink, 1995). Even in the USA, where the soil mapping is fully dominated by the U.S. Soil Taxonomy and the Soil Survey Manual, there are more and more alternative systems being developed (Zhu *et al.*, 2001). This, however, does not mean that the photo-interpretation or empirical knowledge on soils should be cast out from operational soil survey. On the contrary, case studies have shown that the purely geostatistical methods do not always give prediction maps better than those obtained by subjective photo-interpretation (van Kuilenburg *et al.*, 1982; Boucneau *et al.*, 1998).

In this chapter a grid-based SIS, which integrates the use of photo-interpretation, auxiliary terrain and remote sensing data, hybrid pedometric techniques, continuous classification and advanced visualisation techniques is described. This connects the methods from the previous chapters into a real soil survey application.

7.2 Methods

Three main aspects determine the design of a SIS: (a) concepts and elements used (entities); (b) organizational structure and operations and (c) format and presentation of products. In the following sections, the key concepts and elements used are listed. First, the relation between the grid size and cartographic scale is explained, then a schematic flow of the methodological steps and explanation of algorithms for interpolation, classification, inference, visualisation and (dis)aggregation of data is given. Note that I refer to the proposed SIS as the *hybrid grid-based SIS* in the further text — the adjective 'hybrid' determines both the use of the mixed model of spatial variation and hybrid interpolation technique.

7.2.1 Key concepts

Two key concepts specifically distinguish the SIS proposed in this paper from other similar grid-based SIS applications: use of quantitative methods in all parts of mapping process and combination of different mapping techniques (including photo-interpretation, kriging and correlation with auxiliary maps). The latter ensures a combination of the abrupt and continous transitions in space, which is referred to as the *Mixed Model of Spatial Variation* (Mowrer & Congalton, 2000). This is a combination of the discrete and continuous models of spatial variation, although one might argue that the continuous model already can adopt both continuous and less-continuous (discrete) transitions. The following concepts define the hybrid grid-based SIS more closely:

- The fundamental spatial entity is a grid cell. All GIS layers are brought to same grid resolution in order to make calculations and data integration possible. The grid size (resolution) determines the effective scale.
- The focus is production of maps of key land characteristics. This means that the soil mappers need to interview their users prior to the actual sampling and select the most important land characteristics, level of detail (grid size) and required accuracy. These wishes are then adjusted to the available funds.
- The SIS includes not only maps of soil variables and tables of soil attributes but also auxiliary (non-soil) variables used to assist soil mapping, as well as

derived classifications and interpretations. This means that a SIS user can get a better insight into the original data and extend it with an additional survey or investigate eventual problems with the data.

- Three types of operations are used to produce soil geoinformation from input layers: interpolation, classification and inference. All these are achieved using the GIS operations on grid maps, rather than table calculations.

- Quantitative methods are used to interpolate soil variables (universal kriging), classify (fuzzy k-means) and retrieve them.

- Soil properties, classes, and interpretations are modelled using the mixed model of spatial variation, so that both discrete and continuous transitions are possible.

- The original soil description and measurements are linked to the spatial predictions and interpretations, so that the latter can be updated if the former is augmented or corrected. This linkage is kept in tables built for this purpose. For example, the `interpolation table` records the number of regression coefficients and kriging parameters derived from the regression and geostatistical analysis. Consequently, each prediction or interpretation map can be updated by updating the input maps or adding the new soil samples.

7.2.2 Selection of a suitable grid size

The grid size, i.e. the length of one side of a grid cell, is linearly related to the cartographic scale. However, there are different ideas about the suitable grid size for a given scale. In conventional soil cartography, the scale is commonly assessed by using either the Maximum Location Accuracy (MLA) or Average Size Area (ASA) of the polygons on the ground. For example, MLA on the ground when divided with MLA on the map (e.g. 0.25 mm for maps produced according to common map accuracy standards) gives the scale denominator (Rossiter, 2001). To assess the scale denominator via the ASA, the square root of the nominator should be used. These cartographic definitions can also be used to estimate the suitable grid size for a given mapping scale. As a rule of thumb, Rossiter (2001) suggest that four grid cells should be considered equivalent the Minimum Legible Delineation (MLD). According to the definition of Vink (1975) the MLD is 0.25 cm^2 on the map. The suitable grid size is then:

$$p = \sqrt{\frac{MLD}{4}} = \frac{\sqrt{SN^2 \cdot 0.000025}}{2} = SN \cdot 0.0025 \tag{7.1}$$

where p is the grid (pixel) size, MLD is the Minimum Legible Delineation area on the ground and SN is the scale denominator. This means that for a 1:50 K scale, MLD is 6.25 ha and suitable grid size is 125 m, which seems fairly coarse. Larger grid sizes (0.5 mm to 3 mm on the map) have also been recommended by Valenzuela & Baumgardner (1990). In remote sensing, the relation of the ground resolution and the cartographic scale is somewhat stricter. For example, the Landsat images of 30 m ground resolution are commonly related to the 1:50 K or 1:100 K scale (Lillesand & Kiefer, 2000). Hence, the ground resolution can be defined as two times the MLA on the ground:

$$p = SN \cdot MLA \cdot 2 = SN \cdot 0.0005 \tag{7.2}$$

so for 1:50 K scale, a suitable grid size is 25 m.

The third criterion for the selection of the suitable grid size is empirical knowledge of spatial variation. Ideally, the grid size should equal the minimum size of a pedon (1 m^2), especially if the soils are varying at short distances (e.g. cockpits in the Karst area). If the soils are homogeneous spatially and show smoother transitions, much larger grid sizes will be adequate for spatial modelling (Thompson *et al.*, 2001). This means that the selection of the suitable grid size should be adjusted to the spatial variability of soils to avoid over-sampling. Florinsky & Kuryakova (2000) suggested that, for soil-terrain modelling, adequate grid size is the one that offers the highest predictive power, i.e. correlation coefficient in their case. The spatial variation of soils can be estimated from the terrain data i.e. contour data. Hengl *et al.* (2003b) suggest that the grid spacing should be at least half the average spacing between the contours to represent the most changes in a terrain.

Although these three criteria give a range of possible values, a rule of thumb *the finer the grid size the better* is suggested in the most cases. The importance of the finer grid size has been proven to play an important role especially if the terrain data is used for spatial modelling of soils (Dietrich *et al.*, 1995; Thompson *et al.*, 2001). With increasingly powerful computers and cheap storage, fine grid sizes are feasible for most study areas.

7.2.3 Interpolation, classification and inference methods

Three operations play key roles in the production of geoinformation in the hybrid grid-based SIS: interpolation, classification and inference. Each is explained in more detail down bellow. A flow diagram of the computational procedures is given in Fig. 7.1. The profile data is first combined with a set of predictors to produce continuous field maps of measured soil variables. These are then classified to membership maps using continuous classification and the predefined class centres. Finally,

the interpolated soil variables, auxiliary predictors and derived memberships can be used to derive soil interpretations, i.e. inferred soil geoinformation.

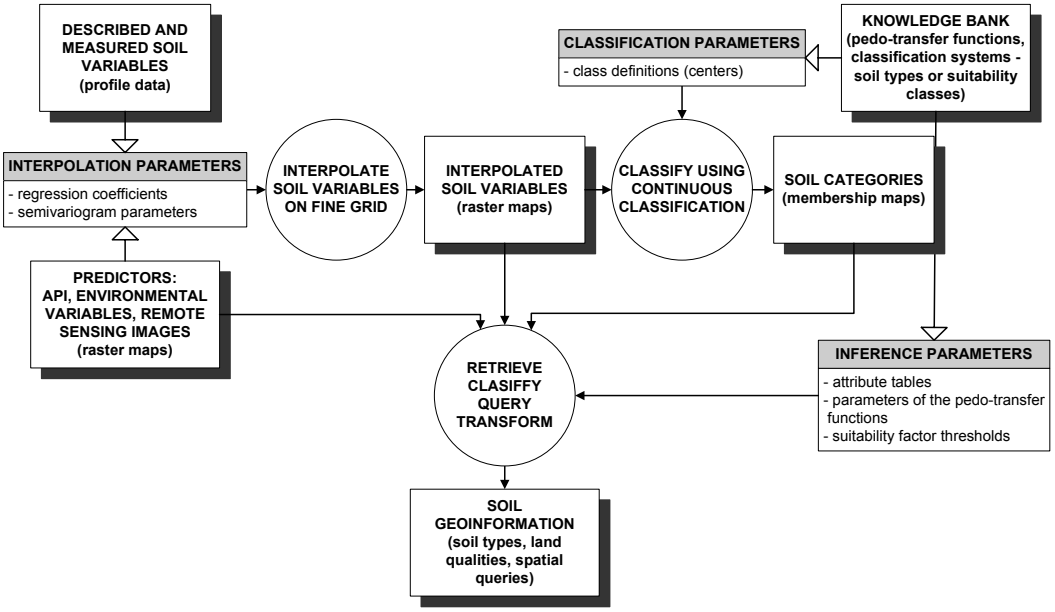

Figure 7.1: Schematic flow of methodological steps.

Interpolation

The generic framework based on the step-wise principal component logistic regression-kriging model, was used to interpolate the soil variables. This algorithm can use information from the photo-interpretation, auxiliary data and spatial auto-correlation at the same time. The algorithm is explained in more detail in chapter 5.

Classification

After all selected soil variables have been interpolated they can be classified using the point observations and the class centres for each category (e.g. soil classes). A flexible classification algorithm is the fuzzy k-means classification, which gives a

membership map for each class. This is the concept of continuous soil mapping, first introduced by **?** and then further on developed by de Gruijter *et al.* (1997). The limitation of their approach, however, is that it employs only geostatistical interpolation while the auxiliary variables are ignored. This approach is somewhat different since first the soil properties are mapped over the whole area and then classified per each grid. This generally means that the produced memberships will follow the pattern of the relief and other predictors, thus giving a more realistic picture. The classification of maps and resulting continuous soil map is explained[2] in more detail in chapter 6.

Inference

The derived memberships, also referred to as similarity values (Zhu *et al.*, 1997), can now be linked to the attribute tables, pedo-transfer functions or suitability ranks (knowledge bank). The key columns can be the soil categories, which is a common way of organizing the SGDB (Zinck & Valenzuela, 1990). The inferred soil attribute is then mapped directly from the membership maps using the linear additive weighting function (Zhu *et al.*, 2001):

$$\hat{S}(i) = \sum_{c=1}^{k} \mu_c(i) \cdot S_c \qquad \sum_{c=1}^{k} \mu_c(i) = 1 \qquad i = 1, .., N \tag{7.3}$$

where $\hat{S}(i)$ is the inferred soil attribute at ith grid position and S_c is the modal value of the inferred soil attribute of the cth category. For example, imagine four membership maps of soil type `A`, `B`, `C` and `D`. The knowledge bank shows that soil type `A` has 10%, `B` 10%, `C` 30% and `D` 40% of clay and the membership values at a grid position are 0.6, 0.2, 0.1 and 0.1, so the Eq. (7.3) will estimate the average clay content of 15%. Note that although the method assumes that a linear weighted average best represents the overall value, the technique can be extended to any aggregation method.

The membership maps can also be used for land suitability assessment. One option is to use the limitation scoring system described by Triantafilis *et al.* (2001). Here, the key issue is to derive limitation scores (or negative points) based on the definition of land qualities and threshold limits. In the case of the hybrid grid-based SIS, the limitation score can be calculated per each pixel by cumulatively using membership maps, interpolated soil variables (e.g. gleying properties) and/or auxiliary variables (e.g. slope):

[2]See also supplementary materials for ILWIS commands.

$$l(i) = \sum_{c=1}^{k} \mu_c(i) \cdot l_c + \sum_{r=1}^{t} S_r(i) \cdot l_r \qquad \sum_{c=1}^{k} \mu_c(i) = 1 \qquad i = 1, .., N \tag{7.4}$$

where l is the accumulated limitation score, l_c is the limitation score of the cth soil type, S_r is the classified auxiliary or soil variable and l_r is the limitation score of the rth class. For example, the same grid position as above (`A`, `B`, `C`, `D`) and the limitations scores 5, 0, 0, 20, give the average limitation score 5. The slope at the same grid position is 10%, which gives 3 more points (9–16%) so that the total accumulative score is 8. The accumulated limitation score, ranging from 0 to ∞ is transformed to continuous land suitability by:

$$L_s = e^{-0.1 \cdot l} \quad L_s \in [0, 1] \tag{7.5}$$

where L_s is the continuous land suitability and l is the accumulated limitation score.

7.2.4 Aggregation and disaggregation

Aggregation or down-scaling is a process of reducing the scale of map and disaggregation is the opposite process. In the grid-based SIS, aggregation means changing towards a coarser resolutions and disagreggation towards finer resolutions, i.e. smaller grid sizes. A schematic example of aggregation and disaggregation in the hybrid grid-based SIS is shown in Fig. 7.2. This models follows the conceptual model of scaling described by McBratney (1998). One advantage of the hybrid grid-based SIS is that the aggregation is easier than with the conventional system where both the soil boundaries and the legend need to be adjusted. In the grid-based SIS, each interpolated continuous soil variable can be resampled to a coarser grid using standard image processing algorithms such as bilinear resampling (Lillesand & Kiefer, 2000). The scaling of the continuous variables is much less problematic than the scaling of categorical variables, such as soil types. The resampling of soil types to a coarser resolution implies that the small local patches will be merged with the dominant types and disappear from the map. Because we deal with maps of soil memberships, we can first resample these to a coarser resolution and then re-standardize them by:

$$\mu_c^{S-}(i) = \frac{\mu_c^{+}(i)}{\sum_{c=1}^{k} \mu_c^{+}(i)} \quad c = 1, 2, ...k \quad i = 1, 2, ...N \tag{7.6}$$

where μ_c^{S-} is the down-scaled membership value and μ_c^{+} is the resampled membership. A longer alternative is to re-calculate soil variables and re-classify soil types from the input maps at finer resolutions.

The hybrid grid-based SIS is also attractive for the purpose of up-scaling, which is in the conventional SIS almost impossible. Because the accuracy of interpolation depends on the quality and detail of auxiliary variables (terrain data, remote sensing images), one can imagine that improving the spatial detail of the predictors will also reflect on the interpolation results. A caution should be made not to 'blow-up' the scale outside the realistic limits defined by the standards. For example, if the inspection density is four observations per km^2 the largest scale that the existing dataset can be disaggregated to is 1:25 K. Additional observations are recommended to achieve larger scales.

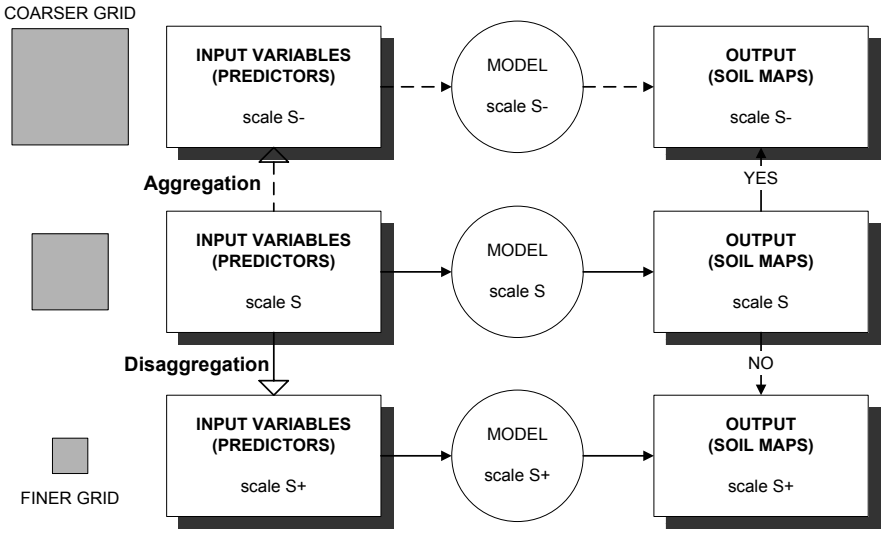

Figure 7.2: Schematic example of aggregation and disaggregation process in the hybrid grid-based SIS. Note that although direct disaggregation of soil maps is possible, it is not recommended. S indicates scale: S^- are smaller scales and S^+ are larger scales.

7.2.5 Case study and data analysis

The methodology was developed and tested using a data set from Baranja hill and a portion of the adjacent Danube terraces in Eastern Croatia. The study area is 3.8×3.8 km square (centred on 45°47'40" N, 18°41'27" E) and corresponds to the

size of a single 1:20 K aerial photo (Fig. 7.3). The main geomorphic facets are hill summits and shoulders, eroded slopes of small vales, vale bottoms and high and low river terraces. The elevations range from 80 to 240 m. I first produced an API map using the geopedological approach of Zinck & Valenzuela (1990). I then made 59 profile observations using a random design (40) and two transect studies (19) (Fig. 7.3c). The boundaries were finally cross-checked on the field to produce a conventional soil map with the legend.

The observed soil types ranged from Calcaric Regosols, Cambisols to Kastanozems (FAO, 1998). The Calcaric Cambisols are the dominant soil type in the hilland, while in the vale bottoms and in the lower floodplain, I observed gleyic properties. At some locations on the hill summits, I observed occurrence of a hypocalcic horizon ($> 15\%$ calcium carbonate equivalent). This layer is neither cemented nor close to the surface so it does not present a limitation for agriculture. I observed the following land use types: vineyards, orchards, natural grasslands, meadows (for animal production), natural forest and woodland (hunting resorts), residential use, fish pond, water control (channels), animal farming and crop fields. The most common crops were maize and wheat, vegetables (manual farming), sugar beet and sunflower.

The most controlling factors for agricultural management in the area are: slope, solum thickness, soil alkalinity and water-saturation conditions. Finally, I selected the following six soil variables as the most important diagnostic land characteristics:

1. Depth to the parent material , i.e. thickness of solum (SOLUM) measured in cm.

2. Occurrence of the gleying properties (GLEY_P) — coded with "0" for not observed, "1" for gleying properties within 50 cm and "0.5" for gleying properties within 50 cm.

3. Occurrence of the Mollic horizon (MOL_H) — coded with "0" for not observed and "1" for observed Mollic horizon.

4. Occurrence of the Calcic horizon (CALC_H) — coded with "0" for not observed and "1" for observed Calcic horizon.

5. Thickness of the topsoil (A_DEPTH) measured in cm.

6. Silt fraction (0.002–0.05 mm) content in topsoil (A_SILT) estimated using the centroids of the textural classes and expressed in percentage. The texture classes ranged from sandy-loam, loam, silt loam to silty clay loam.

Note that the indicator variables GLEY_P, MOL_H and CALC_H have either 0 and 1 value which can not be transformed (see chapter 5, page 94). To avoid

division by zero or ln(0) problems, I introduced a small adjustment of 0.01, so that 0 becomes 0.01 and 1 becomes 0.99. A more optimal approach would be to estimate these threshold iteratively in a statistical package.

The working scale of the project was 1:50 K, hence, a grid size of 25 m, which corresponds to 0.5 mm on the map was selected. For the predictors, I used six terrain parameters (Hengl *et al.*, 2003b): elevation (DEM), slope gradient in % (SLOPE), profile curvature (PROFC), plan curvature (PLANC), wetness index (CTI) and slope insolation (SINS); all derived in ILWIS[3]. As remote sensing-based predictors, I used the intensity (value on the grey scale) of the aerial photo (AP), the standard deviation image filter of the AP map (AP_STD) and NDVI map derived from the Landsat 7 image. The aerial photo was taken in May 1998 and the satellite image in August of 1999. I assumed that these remote sensing-based variables would help explain the occurrence of horizons and depths. The nine maps were first transformed to nine predictive components (SPCs) using factor analysis in ILWIS. This was done to reduce the multicollinearity and optimize the selection of the best subset of predictors[4].

In addition to the SPCs, nine soil mapping units (SMUs) were transformed to nine indicator variables: colluvial footslopes (SMU1), eroded slope (SMU2), floodplain (SMU3), glacis (SMU4), high terrace (SMU5), scarp (SMU6), shoulder (SMU7), summit (SMU8) and vale bottom (SMU9). We also added three land use indicator variables: agricultural land (LU1), natural forest (LU2) and pastures and orchards (LU3). The total number of predictors was 21 (Fig. 7.3a and b). The target soil variables and the predictors were imported to a regression table consisting of 59 observations, 9 target variables and 21 predictors. The 'best' subset of predictors (SPCs, SMUs and LUs) was selected using the stepwise regression in the S-PLUS statistical package (MathSoft Inc., 1999). The regression coefficients and interpolated the residuals were then calculated over the whole study area using the regression-kriging (see chapter 5).

The set of nine interpolated soil maps was further used to classify the whole area. The membership maps were calculated using the supervised fuzzy k-means classification. First the class centres were calculated by averaging the nine soil variables per soil type. For the indicator soil variables the sampled standard deviation was zero, which is unsolvable. The indicator variables follow a binomial distribution, so that the standard deviation can be estimated using:

$$\hat{s}_z = \sqrt{\frac{p \cdot (1-p)}{k}} \tag{7.7}$$

[3]See lecture note "Digital Terrain Analysis in ILWIS", available with supplementary materials.
[4]See chapter 5 for more details.

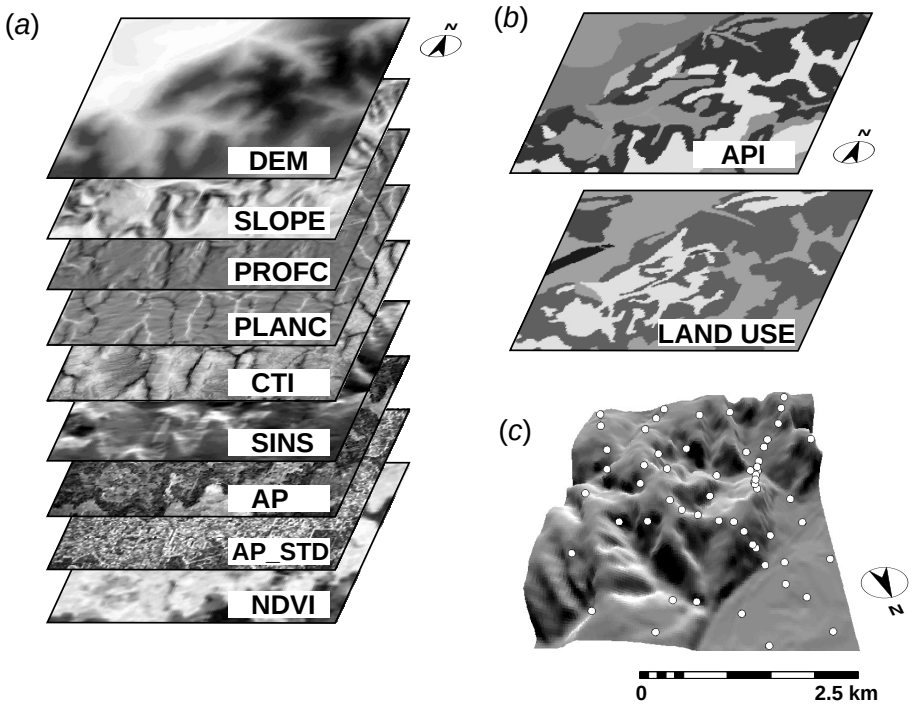

Figure 7.3: Multi-source predictors: (a) auxiliary predictors terrain parameters and remote sensing data; (b) aerial photo-interpretation map (API) and land use map and (c) location of the 59 soil profile observations. DEM – elevation; SLOPE – slope gradient in %; PROFC – profile curvature; PLANC – plan curvature; CTI – wetness index; SINS – slope insolation; AP – intensity of the aerial photo; AP_STD – standard deviation of the AP map and NDVI map derived from the Landsat 7 image.

where p is the threshold probability (e.g. 95%) and k is the number of classes. In the case of MOL_H and CALC_H, the number of classes is two and the standard deviation is 0.15, while in the case of GLEY_P the standard deviation is 0.13.

Membership maps for the six observed soil types were derived: Siltic, Calcisols (CL_s), Calcari-Eutric Cambisols (CM_ce), Gleyi-Calcaric Cambisols (CM_gc), Calcari-Eutric Gleysols (GL_ce), Calci-Siltic Kastanozems (KS_cs) and Calcari-Eutric Regosols (RG_ce). The memberships were then used to derive the limitation score for the land utilisation type (wheat) using the soil types and slope classes as input (Eq. 7.4)). In addition, the membership values were resampled to the 100 m grid using the Eq. (7.6) to demonstrate disaggregation aspects.

7.2.6 Comparison of conventional and hybrid grid-based SIS

The hybrid grid-based SIS was compared with the conventional polygon-based SIS of the same area. I first compared the predictability of SMUs and SPCs, which gives an idea which predictors explain the measured soil variables better. This was done by comparing the correlation coefficient and coefficient of determination between all target variables and predictors. The two systems were also compared for their mapping efficiency using: amount of variation explained and thematic confusion. Amount of variation explained was assessed by calculating the sum of squared residuals, i.e. *RMSE* for each of the six interpolated variables. The lower the *RMSE*, the better is the fitting of the data. The thematic confusion was assessed by calculating the confusion index among each spatial entity:

$$CI = 1 - (\mu_{\max} - \mu_{2nd\,\max}) \tag{7.8}$$

where $\mu_{\max}$ is the highest membership and the $\mu_{2nd\,\max}$ is the second highest membership at the same location (Burrough *et al.*, 1997). The lower the CI, the higher the certainty of the classification system. Note that the CI for SMUs is calculated by first calculating composition of soil types in percentage. The CI value is then attributed to each SMU to derive the overall or average confusion index. In addition to the statistical measures, a summary comparison of the two systems for their cost-effectiveness, flexibility and technical properties was made.

7.3 Results

7.3.1 Mapping soil variables

The factor analysis on the continuous predictors showed that the information overlap is low. The first three SPCs accounted for about 65% of the total variation and it appears that all SPCs need to be taken into account. A first comparison of correlation coefficients between the all combination of SPCs and SMUs with target variables showed that the auxiliary predictors are slightly more correlated with the target soil variables than the SMUs (Fig. 7.4a). However, the amount of variation explained in the multivariate models (adjusted R^2) showed that the SMUs are in general better predictors than the SPCs, except for SOLUM and CALC_H (Fig. 7.4b). In all cases, except for CALC_H, the regression models explained about 40% of variation and were statistically significant (p <0.001). Note that the discrepancy between the univariate correlation coefficients (r) and coefficients of multiple determination (R^2) in Fig. 7.4 is because there is still some thematic overlap in the SPCs. The SMUs (indicator variables) have no overlap by definition so that lower univariate correlations will accumulate more effectively in the multivariate model.

In all cases the step-wise regression selected from 3 to 6 predictors from the 21 possible, or 25% in average (Table 7.1). The best predictors were:

- for SOLUM – SPC1 (CTI, SLOPE), SPC3 (AP_STD) and SMU5 (high terrace);
- for GLEY_P – SMU3 (floodplain area), SMU9 (vale bottom) and SPC9 (CTI)
- for MOL_H – SMU4 (glacis), SMU5 (high terrace) and SPC9 (CTI);
- for CALC_H – LU2 (natural forests) and SMU2 (eroded slope);
- for A_DEPTH – LU1 (agricultural land) and SMU5 (high terrace) and
- for A_SILT – SMU9 (vale bottom), SPC9 (CTI) and SMU3 (floodplain area).

Many predictors, on the other hand, have been ignored by the system, such as SPCs 2,6,7,8, SMUs 1,6,7,8 and LU3. The models in general reflect our empirical idea of the distribution of soils. For example, I observed the gleyic properties in only two mapping units and assumed that these are closely related with the potential of water accumulation, which was also confirmed by the model (SMU3, SMU9 and CTI). In the case of CALC_H, the current predictors are of little help. It seems that this variable is controlled by the parent material and not geomorphology or land use. Note that the adjusted R^2's (Table 7.1) are somewhat higher than the ones in the Fig. 7.4. This is because a lower number of predictors is used for final prediction, which typically means a lower adjusted R^2.

The geostastical analysis of the residuals showed the pure nugget variation for the SOLUM and GLEY_P, fairly long-range spatial dependence for MOL_H and CALC_H and somewhat shorter-range spatial dependence for A_DEPTH and A_SILT (Table 7.1). The pure nugget effect for residuals is reasonable for GLEY_P because most of the variation (70%) has been accounted for by the model. For SOLUM, the pure nugget effect is somewhat more surprising since the residuals are still significant. In this case, only 37% of the total variation has been explained by the regression analysis. This means that SOLUM is much noisier variable and much harder to map, which is probably due to the fuzzy character of the boundary between the solum and parent material. The ordinary kriging of residuals practically 'saved' the prediction of CALC_H, despite the poor regression model. The residuals, however, showed strong spatial dependence, which was sufficient to map it using ordinary kriging.

A visual comparison of the interpolated maps produced using the conventional approach (Fig. 7.5, left) and hybrid interpolation (Fig. 7.5, right) suggests that the hybrid system in general offers more detail and higher contrast. In the case of the

Table 7.1: Soil variables (logit-transforms), selected sub-sample of predictors, adjusted R^2 and estimated variogram parameters.

		Soil variables (logit-transforms)					
		SOLUM^{++}	GLEY_P^{++}	MOL_H^{++}	CALC_H^{++}	A_DEPTH^{++}	A_SILT^{++}
Regression coefficients (predictors)	Intercept	-0.72	-1.953	-2.848	-3.888	-2.014	-0.624
	SPC1	0.0114	-0.002	0.0284	0	-0.002	0.0026
	SPC2	0	0	0	0	0	0
	SPC3	0.0178	0	0	0	-0.008	0
	SPC4	0	-0.004	0	0	0	0.0029
	SPC5	0.013	0	0	0	0	0
	SPC6	0	0	0	0	0	0
	SPC7	0	0	0	0	0	0
	SPC8	0	0	0	0	0	0
	SPC9	-0.013	-0.043	-0.058	-0.018	-0.01	0.0111
	SMU1	0	0	0	0	0	0
	SMU2	-0.18	0	0	1.0332	0	0
	SMU3	0	4.965	0	0	0	-0.622
	SMU4	0	0	8.7559	0	0	0
	SMU5	0.3211	0	8.8444	0	0.7607	0
	SMU6	0	0	0	0	0	0
	SMU7	0	0	0	0	0	0
	SMU8	0	0	0	0	0	0
	SMU9	0	5.9859	0	0	0	-0.777
	LU1	0	0	0	0	0.4484	0
	LU2	0	0	0	1.4065	-0.066	0
	LU3	0	0	0	0	0	0
	R^2_a	0.37	0.70	0.59	0.13	0.41	0.61
Variogram	Variogram model	nugget effect	nugget effect	exponential	exponential	exponential	exponential
	C_0	0.156	3.78	0.27	0	0	0
	C_0+C_1	0.156	3.78	26.2	8.28	0.192	0.122
	R (m)	0	0	10 km	759	194	69

hybrid systems, not only discrete and continuous transitions can be seen, but also the pattern of relief or land use is reflected via the auxiliary maps. This hybrid pattern is especially distinct in the map of A_SILT: the highest values follow the steeper slopes, discrete transitions are visible in the floodplain area but also the kriging pattern with hot spots (Fig. 7.5c, right).

The conventional system is more sensitive to the fairly contrasting inclusions in the mapping unit. For example, the prediction map of the GLEY_P for the conventional system shows a value of 0.1 even at locations where no gleying could have occurred (Fig. 7.5b, left). This is because there was a single profile (inclusion) in this mapping unit, which somehow finished in the neighbouring polygon (probably boundary misplaced during API). This affected then the whole attribute map giving an unrealistic prediction of occurrence of gleying properties.

Comparison of the *RMSE* at observation points for these six variables showed no large difference for SOLUM (17.4 cm vs. 17.8 cm) and GLEY_P (0.18 vs. 0.13), but in all other case was the data better fitted with the hybrid interpolation technique (0.21 vs. 0.02 for MOL_H, 0.26 vs. 0.01 for CALC_H, 8.6 cm vs. 0.7 cm for A_DEPTH and 7.8% vs. 1.1 for A_SILT).

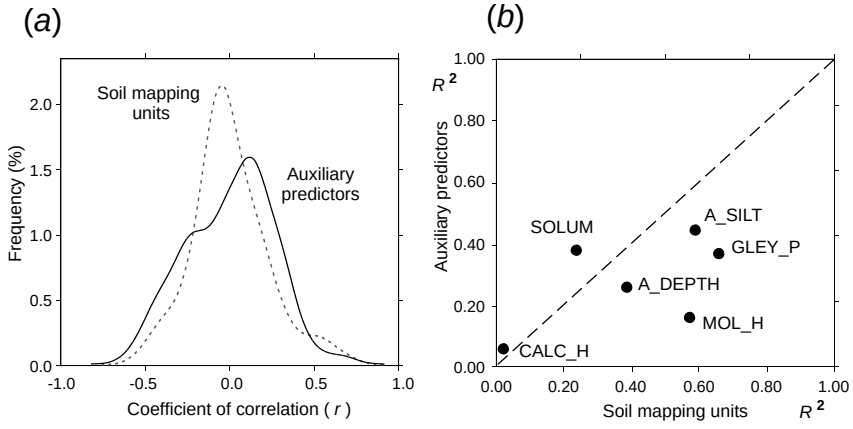

Figure 7.4: Comparison of relationships between the soil variables and soil mapping units and auxiliary predictors: (a) density histograms of the correlation coefficients for univariate models and (b) coefficients of multiple determination for fitted soil variables. SOLUM - depth to the parent material in cm; GLEY_P - occurrence of the gleying properties; MOL_H - occurrence of the Mollic horizon; CALC_H - occurrence of the Calcic horizon; A_DEPTH - thickness of the topsoil in cm; A_SILT - silt fraction content in topsoil.

7.3.2 Classification, down-scaling and inference

The classified map of soil types (Fig. 7.6b) reflects empirical ideas, following the fieldwork experience, about the distribution of the soils. The CM_ce is the dominant soil type covering 61% of the study area, CM_gc and GL_ce occur as expected at lowest convex positions, while the RG_ce occurs more locally (slopes). The CL_s was depicted as the highest membership in only 0.6% of the study area and as the mapping of Calcic horizon was difficult.

From the sampled class centres for the six soil types (Table 7.2), it can be seen that some classes can be distinguished in the attribute space more easily then others. For example, KS_cs is clearly a distinct soil type: deep soil, with occurrence of Mollic horizon and no gleying properties. The factor analysis of class centres also showed that especially CM_ce and RG_ce; and CM_gc and GL_c are similar soil types. This information about the similarity of soils was then used to produce a fuzz-metric legend and then visualise soil taxa and problematic areas as a continuous soil map (Fig. 7.6c). This mixed-colour map indeed shows highest classification uncertainty between the CL_s and KS_cs (note the white patches in Fig. 7.6c). This information can now be used to collect additional samples or cross-check accuracy of our classification system. Also note that the continuous soil map shows three major groups of soil types indicated as bluish (CM_ce, RG_ce and CL_s), greenish (GL_ce, CM_gc) and reddish (KS_cs).

The average confusion index for the conventional SIS, calculated using Eq. (7.8), was 51% (±28%) for the whole map. The confusion index for the hybrid grid-based SIS was 17% (±14%) in average (see the legend in Fig. 7.6a). This means that the spatial confusion between the membership maps is significantly lower ($p < 0.05$) than the confusion within the SMUs for the conventional SIS. After the down-scaling (100 m grid), the less frequent classes did not disappear from the map as we would have expected. For example CL_s occupies about 9 ha in the 100 m scale map, while it occupied 7.9 ha in the 25 m scale map (Fig. 7.6d). This means that the proposed aggregation algorithm retains smaller-size features if their membership is more distinct.

From the membership maps and classified slope map the accumulated limitation score and the resulting continuous land suitability for wheat were derived. The schematic example of the calculation is shown in Fig. 7.7. I used the following limitation scores: 3 (CL_s), 1 (CM_ce), 1 (CM_gc), 9 (GL_ce), 0 (KS_cs) and 9 (RG_ce) for soil types and 0 (0-2%), 1 (2-8%), 3 (9-16%), 9 (17-25%) and 27 (> 25%) for the slope classes.

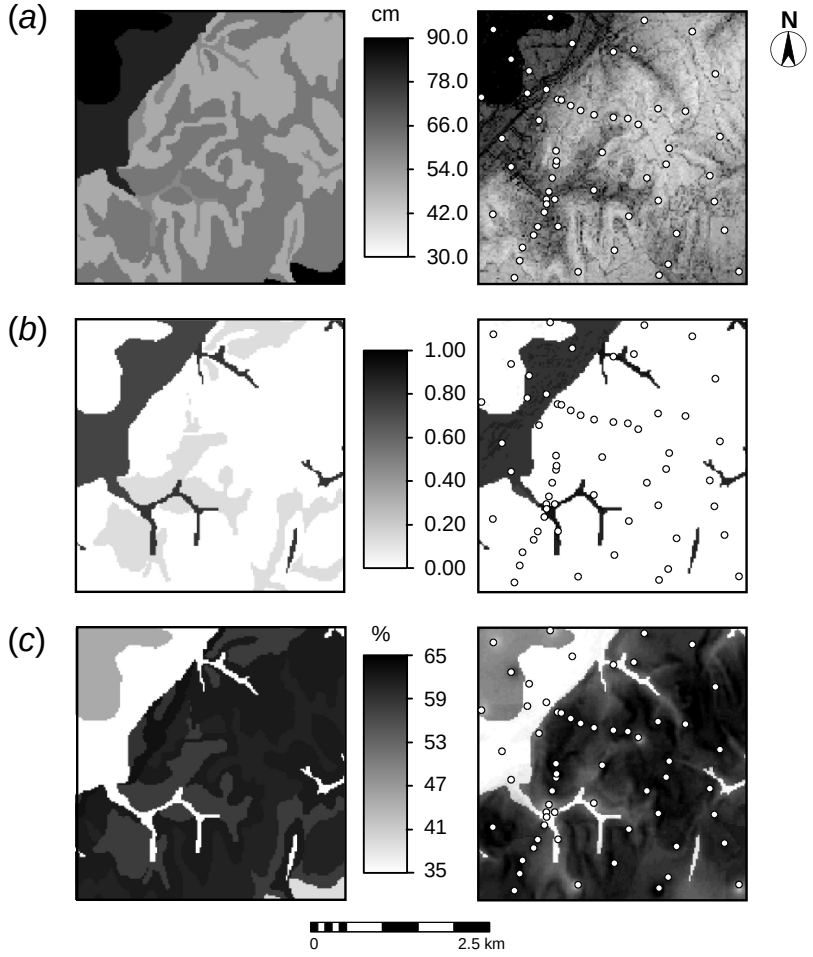

Figure 7.5: Comparison of (a) depth to the parent material (SOLUM); (b) occurrence of the gleying properties (GLEY_P) and (c) silt fraction content in topsoil (A_SILT), interpolated using the mapping units only (left) and the hybrid interpolation algorithm (right).

7.4 Conclusions and discussion

In this chapter I have presented some key concepts, operations and organizational issues of a grid-based SIS as an alternative to the conventional polygon-based SIS and plain geostatistical techniques. The proposed hybrid grid-based SIS was not developed for purpose of replacing conventional techniques and concepts, replac-

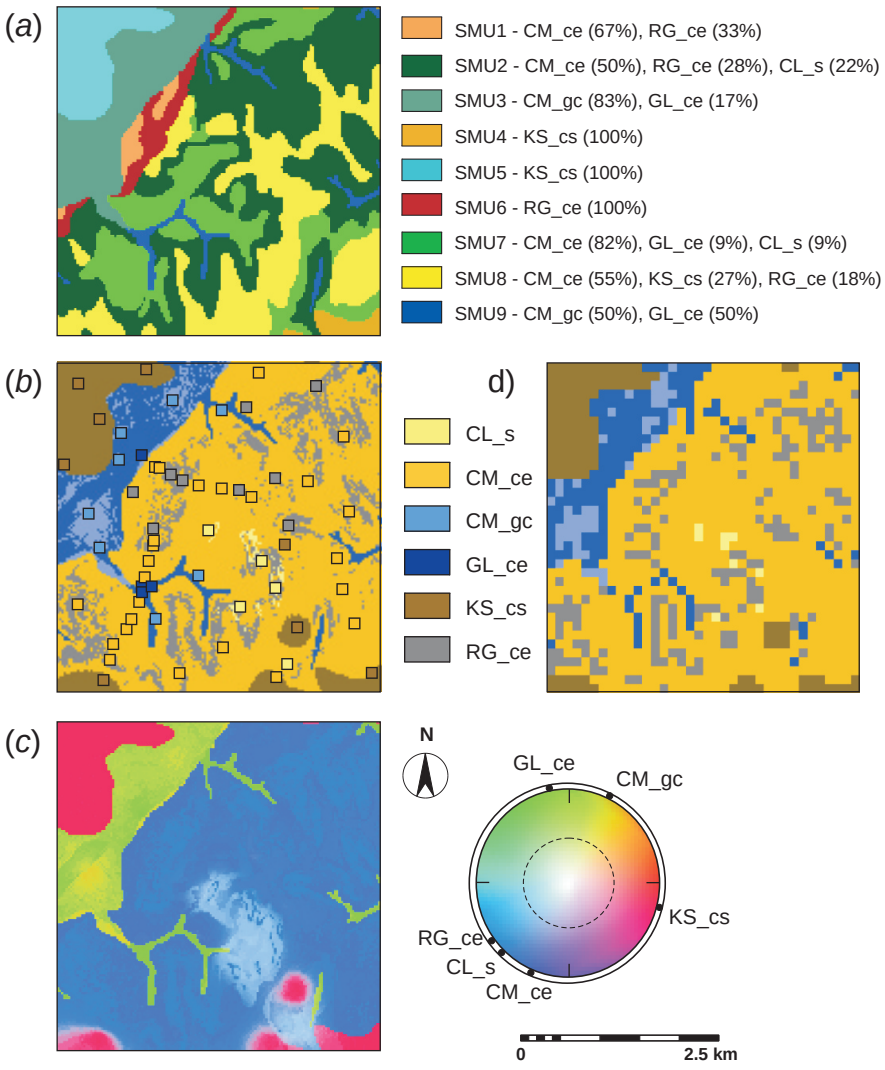

Figure 7.6: Comparison of (a) the conventional soil map with compound composition of mapping units, (b) defuzzified (highest) membership map from the supervised fuzzy k-means classification with freely selected colours; (c) the continuous soil map with a circular legend and (d) down-scaled map to 100 m grid. CL_s - Siltic, Calcisols; CM_ce - Calcari-Eutric Cambisols; CM_gc - Gleyi-Calcaric Cambisols; GL_ce - Calcari-Eutric Gleysols; KS_cs - Calci-Siltic Kastanozems and RG_ce - Calcari-Eutric Regosols.

Table 7.2: Class centres used to classify the six soil types from six attributes.

	Sampled class centres and variation around the central values					
	SOLUM	GLEY_P	MOL_H	CALC_H	A_DEPTH	A_SILT
	(σ)	(σ)	(σ)	(σ)	(σ)	(σ)
	cm	-	-	-	cm	%
CL_s	37.4	0	0	1	17	63
	(11.4)	(0.13)	(0.15)	(0.15)	(12)	(5.1)
CM_ce	60.16	0	0	0	22.48	61
	(16.3)	(0.13)	(0.15)	(0.15)	(6.9)	(8.6)
CM_gc	77.75	0.5	0	0	32.5	37.3
	(14.5)	(0.13)	(0.15)	(0.15)	(14.5)	(3.2)
GL_ce	63.75	1	0	0	23.25	29.5
	(25)	(0.13)	(0.15)	(0.15)	(4.6)	(7.1)
KS_cs	92.88	0	1	0	47.13	51
	(14.2)	(0.13)	(0.15)	(0.15)	(5.5)	(12.8)
RG_ce	36.67	0	0	0	17.22	61.6
	(15.2)	(0.13)	(0.15)	(0.15)	(6.5)	(4.3)

ing existing soil databases or devaluating the importance of photo-interpretation or existing classification systems, but to employ these in a more objective manner. Moreover, the proposed hybrid grid-based SIS is a generalization of the conventional approach. One can imagine that if the within-unit variability is infinitively small and if there is no overlap between class definitions, than the hybrid SIS will show the same, so-called, "double-crisp" form (crips objects and crisp classes) as a conventional map. In fact, in our case study the API units played an important role and the transition of soils was, consequently, more discrete in many parts of the area.

The summary comparison of the two systems can be seen in Table 7.3. The important advantages of the hybrid grid-based SIS that need to be emphasized are:

- It directly offers a map of soil types rather than a map of the soil-mapping units.

- All variables, including the soil types and land suitability are mapped in a continuous manner and on fine grain of detail. In this case study, the average

Table 7.3: Summary comparison between the conventional polygon-based and grid-based SIS. The technical details apply to the study area.

Aspect	Polygon-based	Grid-based
Entity	Polygon	Grid Average
Detail (average size area)	33.8 ha (581 m)	0.0625 ha (25 m)
Content	Polygon class-type map linked with attribute tables (profile observations)	Set of grid maps linked with attribute tables (regression coefficients, variogram parameters, central values, limitation scores)
Interpolation method	Averaging per SMU or soil type	Regression-kriging
Products	Distribution of soil mapping units with composition; soil profile database; crisp land suitability	Distribution of soil variables (land characteristics), soil types and land suitability with estimated uncertainty
Purity of entities (confusion index)	low (51% in average)	high (17% in average)
Level of detail and reliability of predictions	Only average or modal values; contrasting inclusions may be listed separately	Higher level of detail; the predictions follow the pattern in relief, vegetation or land use, according to factors included in the model
Data input and analysis	API by surveyors conceptual knowledge; lines are digitized; topology is created in GIS ; soil profile observations are organized in a relational database	Auxiliary maps are obtained from secondary sources; computations can be demanding and the end product depends on the quality of the input data and algorithms used for interpolation
Memory use	Single vector map and set of tables (very low); 10 KB per km^2 at 1:50 K	About 21 map of predictors, 9 maps of transformed predictors (SPCs), 6 maps of soil variables, 6 maps of soil types etc. (very high); 400 KB per km^2 at 25 m resolution

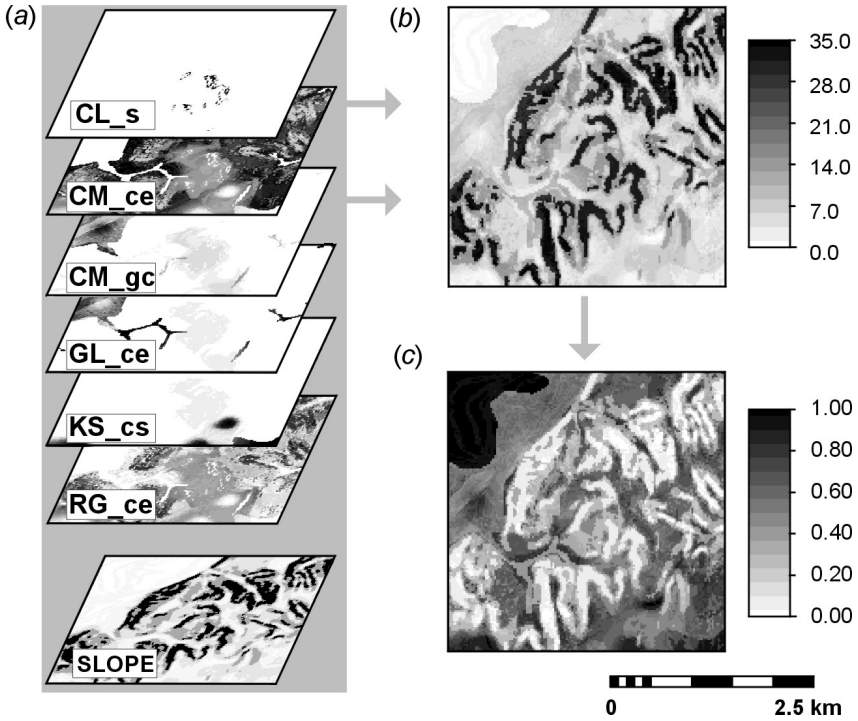

Figure 7.7: Mapping continuous land suitability for wheat: (a) memberships for soil types and slope classes; (b) accumulated limitation score and (c) continuous land suitability. CL_s - Siltic, Calcisols; CM_ce - Calcari-Eutric Cambisols; CM_gc - Gleyi-Calcaric Cambisols; GL_ce - Calcari-Eutric Gleysols; KS_cs - Calci-Siltic Kastanozems; RG_ce - Calcari-Eutric Regosols and SLOPE - slope gradient in %.

size of detail was about 25 times smaller for the grid-based SIS.

- The products of mapping are not only maps of soil variables but also the respective prediction uncertainty (i.e. prediction error or confusion index).
- Maps are more suitable for integration with other geo-data.
- It in general provides more reliable soil geoinformation with lower thematic confusion and higher level of detail than the conventional survey.
- The original soil observations and interpolation/classification parameters are linked to the GIS calculations via the special tables and can be updated.

On the other hand, the disadvantages of the hybrid grid-based SIS are:

- It is computationally demanding as it requires number of GIS, statistical operations with each variable. It also consumes a lot of memory: I estimated that for this case study the memory consumption per km^2 is about 40 times bigger for grid-based SIS.
- It requires number of auxiliary variables, which also means somewhat higher investments.
- Because it is data-driven, it fully depends on the quality of the input data. The prediction maps, however, can always be saved with a good API map and manual correction of problematic features.

The number of observations also plays an important role. In this case study I have dealt with a small case study and relatively small number of profile observations. This caused some problems for the fitting of the data, variogram modelling and factor analysis of the thematic similarity. A much larger number of observations, predictors and soil variables will probably be more satisfactory to the real users. I also experienced problems with interpolation of some variables. In this case study this was occurrence of the calcic horizon, which seems to be difficult with this set of predictors. This feature could have been probably explained better with the use of parent material as auxiliary map.

Also note that some of the applied algorithms, such as the continuous land suitability, are not completely satisfactory. Although this method objectively combines limitations, it depends entirely on the subjective assignment of limitation scores to classes, and also on the concept that a linear combination best expresses suitability.

A more flexible system will be to keep all original data in original cell size (or as sample points) and up or downscale as necessary depending on the algorithm. The input data often comes at different resolutions (multi-source data), for example, terrain data may be available at finer resolution (10 m), satellite data at coarser resolutions (30 m) or very coarse resolutions (1 km). Calculations with raster maps of different resolutions without resampling, however, are still not possible in many GIS packages. Another improvement would be to use the kriging by moving window and not the global estimation of the regression residuals. This would, however, require even more input points and computational power.