Available online at www.sciencedirect.com
ScienceDirect
Geoderma 141 (2007) 1–14
GEODERMA
www.elsevier.com/locate/geoderma

Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping

F. Carré [a,*], Alex B. McBratney [b], B. Minasny [b]

[a] European Commission, DG Joint Research Centre, Institute of Environment and Sustainability, Land Management Unit, TP 280, 21020 Ispra (Va), Italy
[b] Australian Centre for Precision Agriculture, Faculty of Agriculture, Food & Natural Resources, The University of Sydney, NSW 2006, Australia

Received 19 January 2006; received in revised form 24 August 2006; accepted 24 January 2007
Available online 29 May 2007

Abstract

Legacy soil data form an important resource for digital soil mapping and are essential for calibration of models for predicting soil properties from environmental variables. Such data arise from traditional soil survey. Methods of soil survey are generally empirical and based on the mental development of the surveyor, correlating soil with underlying geology, landforms, vegetation and air-photo interpretation. There are no statistical criteria for traditional soil sampling, and this may lead to biases in the areas being sampled. The challenge is to test the use of legacy data for large-area mapping (e.g. national or continental extents) in order to limit the funds of field survey for large-area mapping. The problem is then to assess the reliability and quality of the legacy soil databases that have been mainly populated by traditional soil survey, and if there is a possibility of additional funding for sampling, to determine where new sampling units should be located. This additional sampling can be used to improve and validate the prediction model.

Latin hypercube sampling (LHS) has been proposed as a sampling design for digital soil mapping when there is no prior sample. We use the principle of hypercube sampling to assess the quality of existing soil data and guide us to locations that need to be sampled.

First an area is defined and the empirical environmental data layers or covariates are identified on a regular grid. The existing soil data are matched with the environmental variables. The HELS algorithm is used to check the occupancy of the legacy sampling units in the hypercube of the quantiles of the covarying environmental data. This is to determine whether legacy soil survey data occupy the hypercube uniformly or if there is over- or under-observation in the partitions of the hypercube. It also allows posterior estimation of the apparent probability of sample units being surveyed. From this information we can design further sampling. The methods are illustrated using legacy soil samples from Edgeroi, New South Wales, Australia, and from a large part of the Danube Basin. One third of the total number of sampling units are added to the original dataset. These new sampling units improve the representation of the feature space of the covariate. The standard deviation of the overall density is consequently smaller.

Keywords: Legacy soil data; Soil sampling; Hypercube sampling; Pedometrics; Soil survey; Digital soil mapping

1. Introduction

Legacy soil data arise from traditional soil survey (Bui and Moran, 2001). Methods of soil survey are generally empirical and based on the mental development of the surveyor, correlating soil with underlying geology, landforms, vegetation and air-photo interpretation. There are no statistical criteria for traditional soil sampling, this may lead to bias in the areas being sampled.

de Gruijter et al. (2006) offer some very thoughtful definitions in relation to sampling which we paraphrase here and use subsequently. *Sampling sensu lato* comprises selecting parts from a universe with the purpose of taking observations on them. The selected parts may be observed *in situ*, or material may be taken out from them for subsequent measurement in a laboratory. It is the collection of selected parts that is referred to as the *sample*. A single part that is, or could be, selected, is referred to as a *sampling unit*. The total number of sampling

* Corresponding author. Tel.: +39 0332 78 65 46; fax: +39 0332 78 63 94.
E-mail address: Florence.Carre@jrc.it (F. Carré).

0016-7061/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2007.01.018

units in a sample is referred to as the *sample size*. The material possibly taken from a sampling unit is referred to as an *aliquot*.

Using these definitions a legacy soil sample is a collection of sampling units that have been selected (probably) with unequal probability.

Usually, for national scale survey, the possibility of using legacy soil observation data is really expected since it can avoid the expense of new soil survey. But these data can also have limitations depending on:

- their description: the data have to answer the target mapping issue;
- their location/number: if the data must answer specific issues like "where to find soils with high agronomic potential", their location might not be representative of the pedogenesis of the whole area. The main issue is that usually, these existing data have been sampled at different times in order to answer various and manifold questions. In this case, soil samples can have different soil variable descriptions and some areas of interest can be relatively over- or undersampled.

In digital soil mapping, soil samples may be used to elaborate quantitative relationships or models, between soil attributes and soil covariates, the *scorpan* variables (McBratney et al., 2003). Because the relationships are based on the soil observations, the quality of the resulting soil map depends also on the soil observation quality. Usually, a digital soil mapper tries to optimise the accuracy of the models and minimize the errors (Heuvelink et al., 2006), without taking into account the quality of the legacy data. (Legacy) sample quality evaluation has been done in forest ecology (Vanclay et al., 1995) and in computer science (Bisbal et al., 2005) but not in soil science. In this paper, we focus only on the location quality of the soil samples in the feature and geographic spaces.

An appropriate sampling design for digital soil mapping depends on how much data are available and where the data are located. Since the '90s, some statistical methods have been developed for optimizing data sampling for soil surveys. Some deal with the use of ancillary information. Heuvelink et al. (in press) designed the sampling minimizing the spatially averaged universal kriging variance. Simbahan and Dobermann (2006) compared three different optimization criteria: the minimization of the mean of the shortest distances, a uniform distribution of point pairs for variogram estimation (Warrick and Myers, 1987) and a combination of both. All of these methods involved simulated annealing algorithms (Van Groeningen and Stein, 1998; Fereyra et al., 2002). Hengl et al. (2003) proposed sampling along the principal components of the ancillary variables. The number of samples taken from each component is proportional to the total variance described by each of the component. Other methods do not involve ancillary information. Lark (2000) introduced fuzzy sets of grid spacings when there are uncertainties

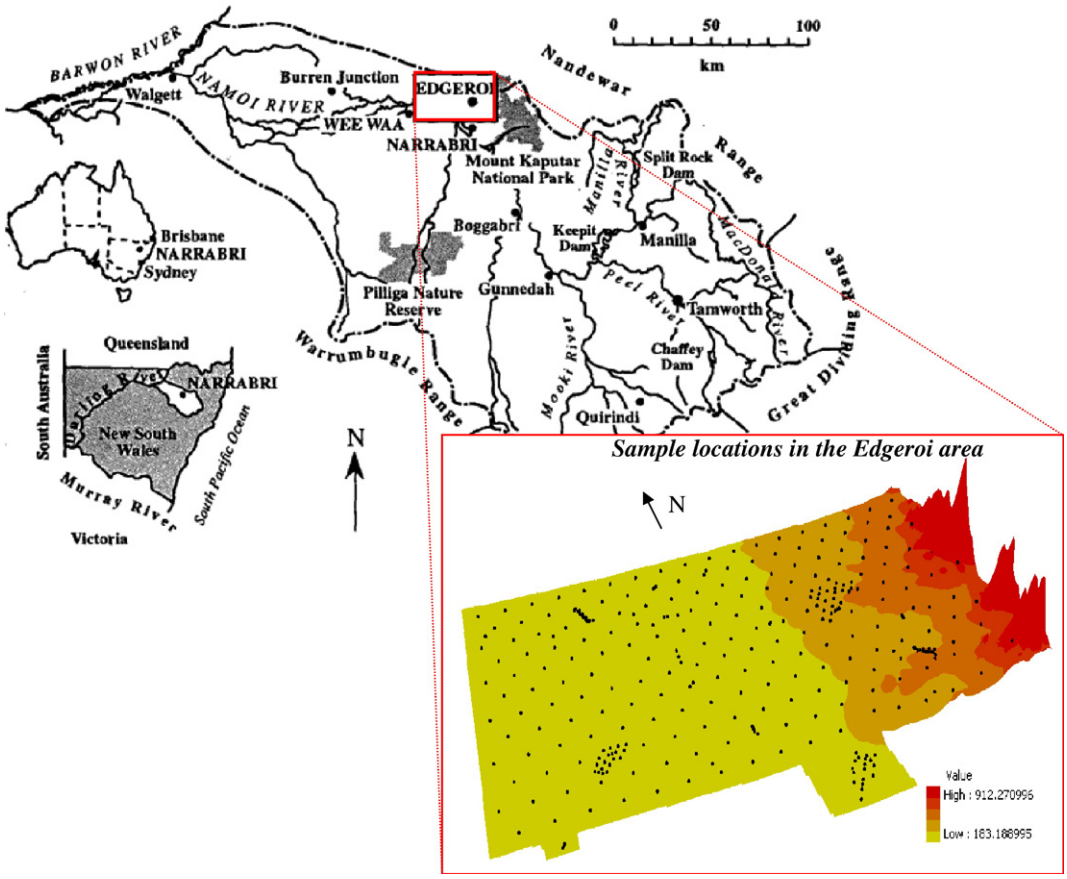

Fig. 1. Study location and the legacy sample location on the altitude of the Edgeroi area.

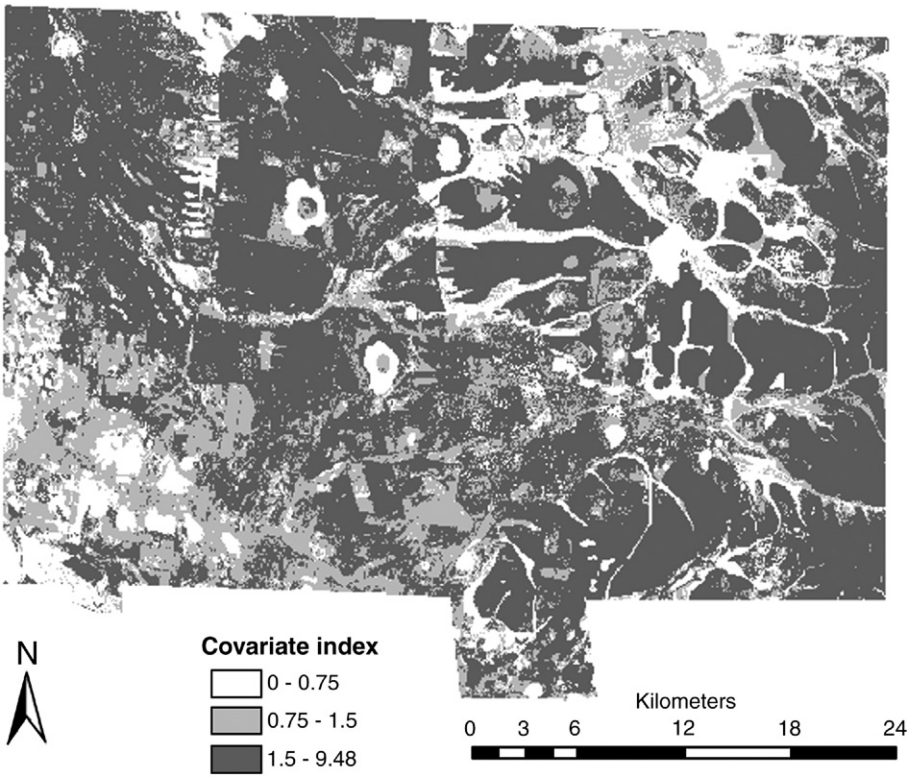

Fig. 2. Covariate index of the hypercube strata in the Edgeroi area.

about the variograms and minimized also the associated kriged variances. Brus et al. (in press) used *k*-means algorithms for minimizing the mean of the shortest distances.

Latin hypercube sampling (LHS) (Minasny and McBratney, 2006) has been proposed as a sampling design for digital soil mapping when there is no prior soil sample (only information

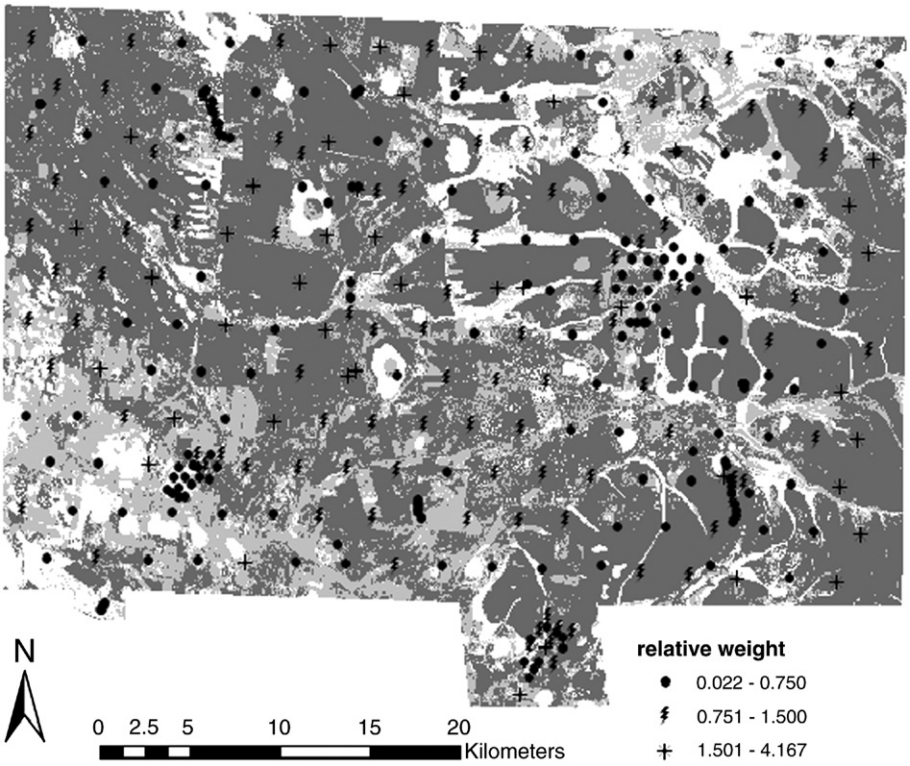

Fig. 3. The relative weights of the legacy sampling units in the Edgeroi area.

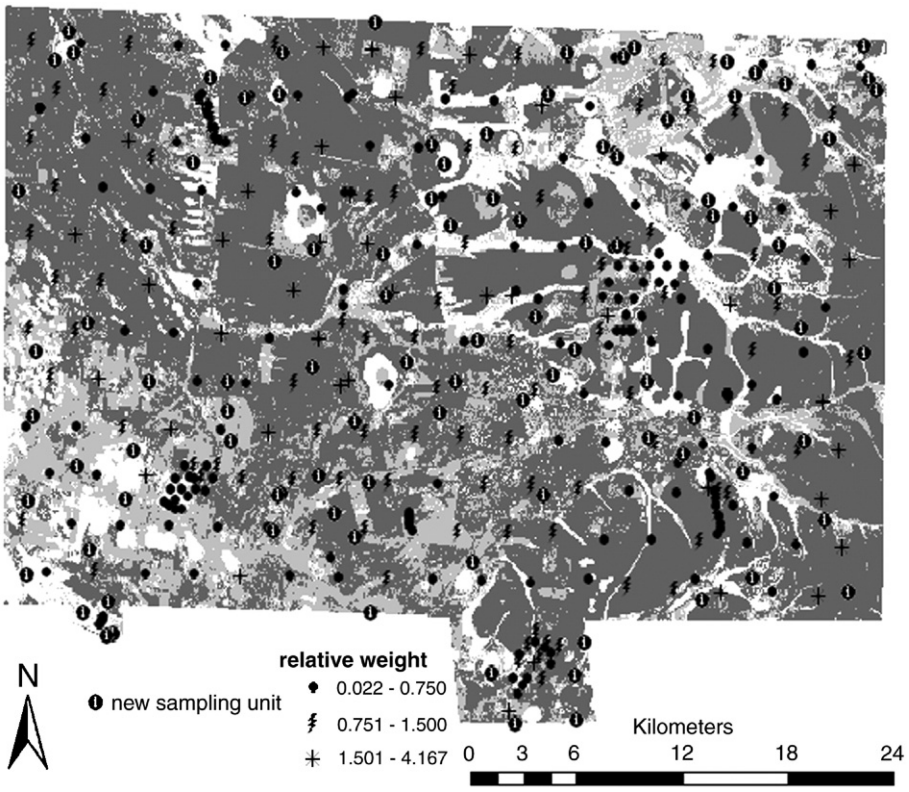

Fig. 4. The locations of the 100 new (additional) sampling units in the Edgeroi area.

on ancillary data). LHS is a constrained Monte Carlo sampling scheme (McKay et al., 1979; Xu et al., 2005). It is a stratified-random procedure that provides an efficient way of sampling variables from their multivariate distributions. A square grid containing sample positions is a Latin square if (and only if) there is only one sampling unit in each row and each column. A

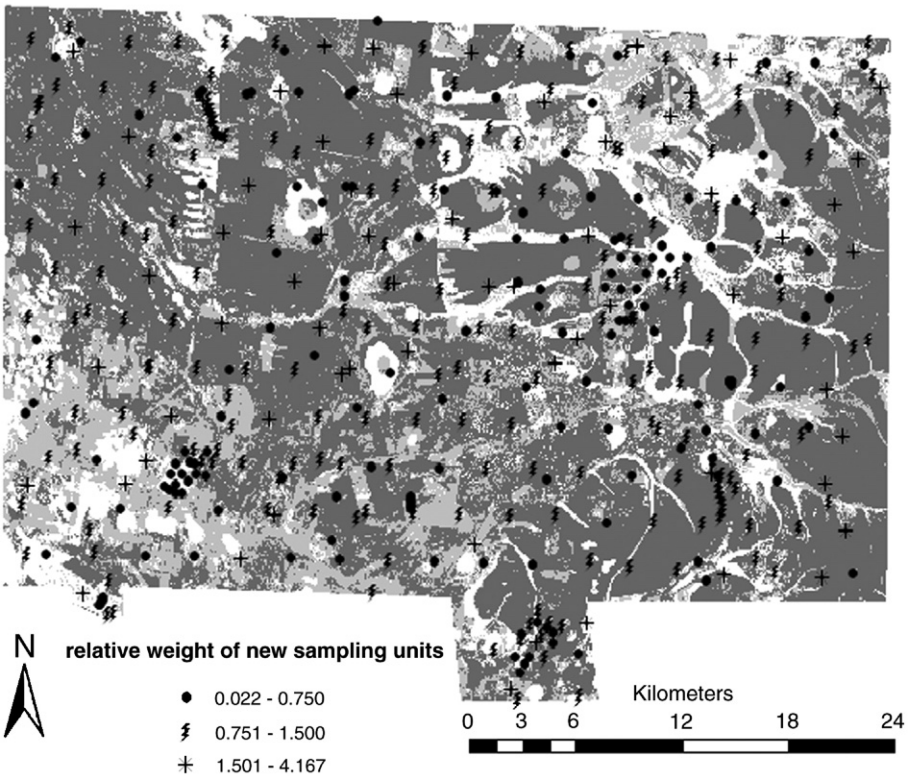

Fig. 5. The relative weights of the enhanced sample (341 legacy plus 100 additional sampling units) in the Edgeroi area.

Table 1
Comparison of the distribution of relative weights between the enhanced sample and the legacy sample for the Edgeroi area

	Legacy sample	Enhanced sample
Number of sampling units	341	441
Mean of relative weights	0.42	0.59
Standard deviation of the relative weights	3.48	2.39
Percentage of sampling units with a relative weight <0.75 (representing oversampling)	45.5%	32.4%
Percentage of sampling units with a relative weight >1.5 (representing undersampling)	16.3%	18.9%

Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions, whereby each sampling unit is the only one in each axis-aligned hyperplane containing it. LHS involves sampling n values from the prescribed distribution of each of the variables. The cumulative distribution for each variable is divided into n equiprobable intervals, and a value is selected randomly from each interval. The n values obtained for each variable are then paired with the other variables. This method ensures a full coverage of the range of each variable by maximally stratifying the marginal distribution. We use the principle of hypercube sampling to assess the quality of existing soil data and guide us to the area that needs to be sampled.

The purpose of the study is then two-fold: (1) evaluating the quality of sampling of existing data and (2) locating new soil observations in non-sampled areas (in the covariate space) using ancillary soil information. The use of an existing methodology (Minasny and McBratney, in press) is proposed for this and illustrated with two examples.

2. Methodology

The methodology is divided into two steps: the first one is the evaluation of the legacy sampling with hypercube sampling and the second one is the improvement of the sample quality.

2.1. Evaluation of a legacy sample

2.1.1. Rationale

The basic rationale is to set up a hypercube, the axes of which are the quantiles of completely enumerated (rasters of)

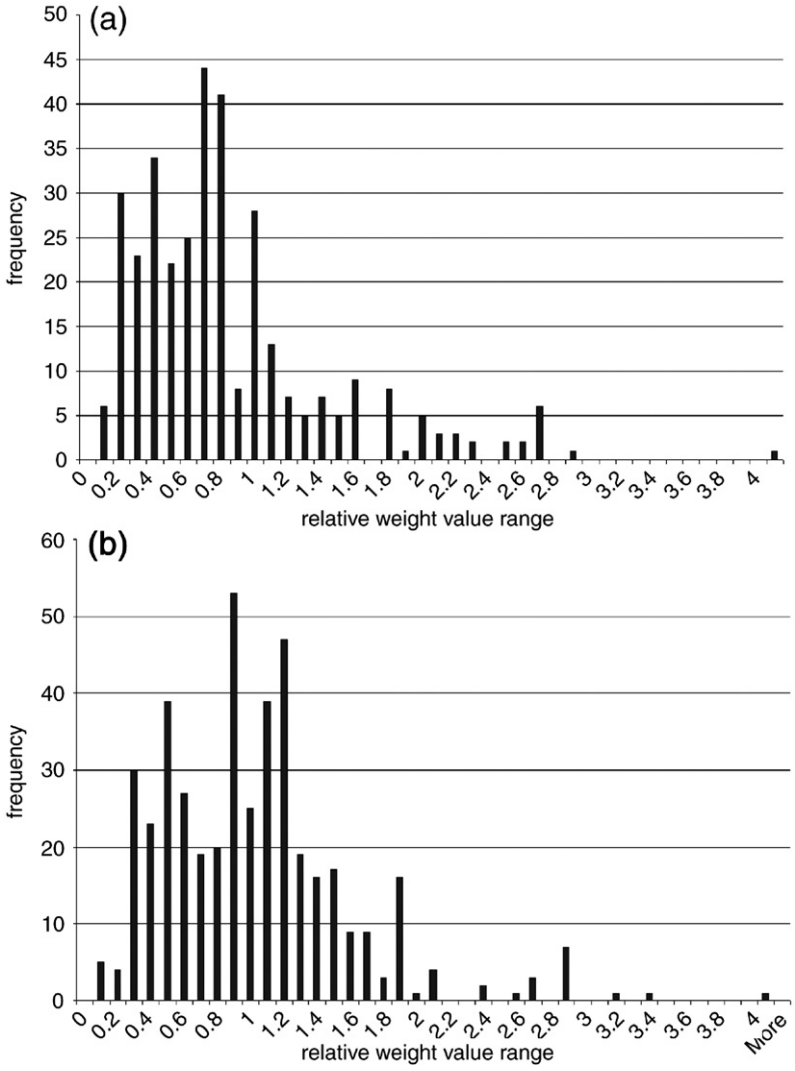

Fig. 6. Comparison between the distributions of the relative weights of the legacy sample (a) and the enhanced sample (b) in the Edgeroi area.

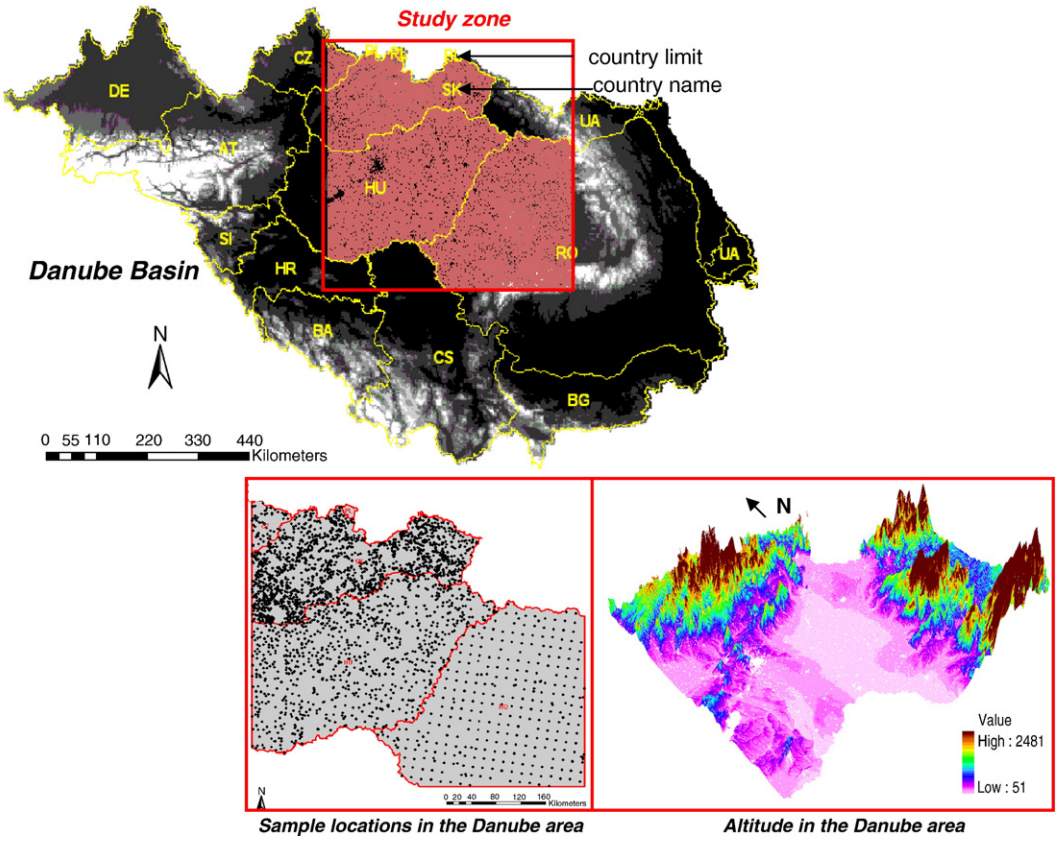

Fig. 7. The Danube Basin study site and the legacy soil sampling unit locations on the SRTM altitude.

environmental covariates, e.g., DEM, and to observe how the sampling units of the legacy sample are located within this hypercube. A good quality sample will have the same density of sampling units within each of the strata of the hypercube. The relative densities of the legacy sampling units and the environmental covariate rasters can be measured and a relative density ratio calculated. This relative density ratio can be used to suggest posterior selection probabilities and relative weights for the legacy sampling units which because of the non-statistical nature of the sampling process are usually assumed to be unknown or unknowable. The success of this of course relies on some dependency of the targeted soil attribute on the environmental covariates. The evaluation is achieved using the Hypercube Evaluation of a Legacy Sample (HELS) algorithm described below. Before running the HELS algorithm, covariates are chosen that are completely enumerated (observed on a complete raster for the area of interest) and show some predictability for the target attribute(s). For digital soil mapping these covariates will represent some of the *scorpan* factors (McBratney et al., 2003).

2.1.2. HELS algorithm

For p covariate rasters of r grid cells, Q quantile ranges for each covariate, and n sampling units in the legacy sample (usually $r>>n$). We call a "stratum" a division of the feature space composed of the combination of quantile ranges of the covariates. There are Q^p different strata.

(1) Choose Q such that Q is maximized given $Q^p<n$ (this will ensure that the hypercube below will have on average at least one sampling unit in every stratum).
(2) Find quantiles Q^p of each of the r grid cells of the covariates.
(3) Form a p dimensional hypercube with axes the quantiles Q of each covariate.
Locate the Q^p in the hypercube.
(4) Form Q^p strata s in the hypercube.
(5) For each stratum s, estimate the number of grid cells M_s.
(6) Calculate the density of grid cells M_s/r in each stratum.
(7) Give the coordinates of each of the n legacy sampling units in the hypercube.
(8) For each stratum, calculate the number of legacy sampling units n_s.
(9) Calculate the density of each observation n_s/n in each stratum.
(10) Calculate the relative weight, $r_w=1/w$, where $w=(n_s/n)/(M_s/r)$ and allocate it to each legacy sampling unit.

2.2. Improving the sample quality

2.2.1. Resources to get more samples

The HELS algorithm will tell us where there is over- or under-observation in the legacy soil sample relative to the

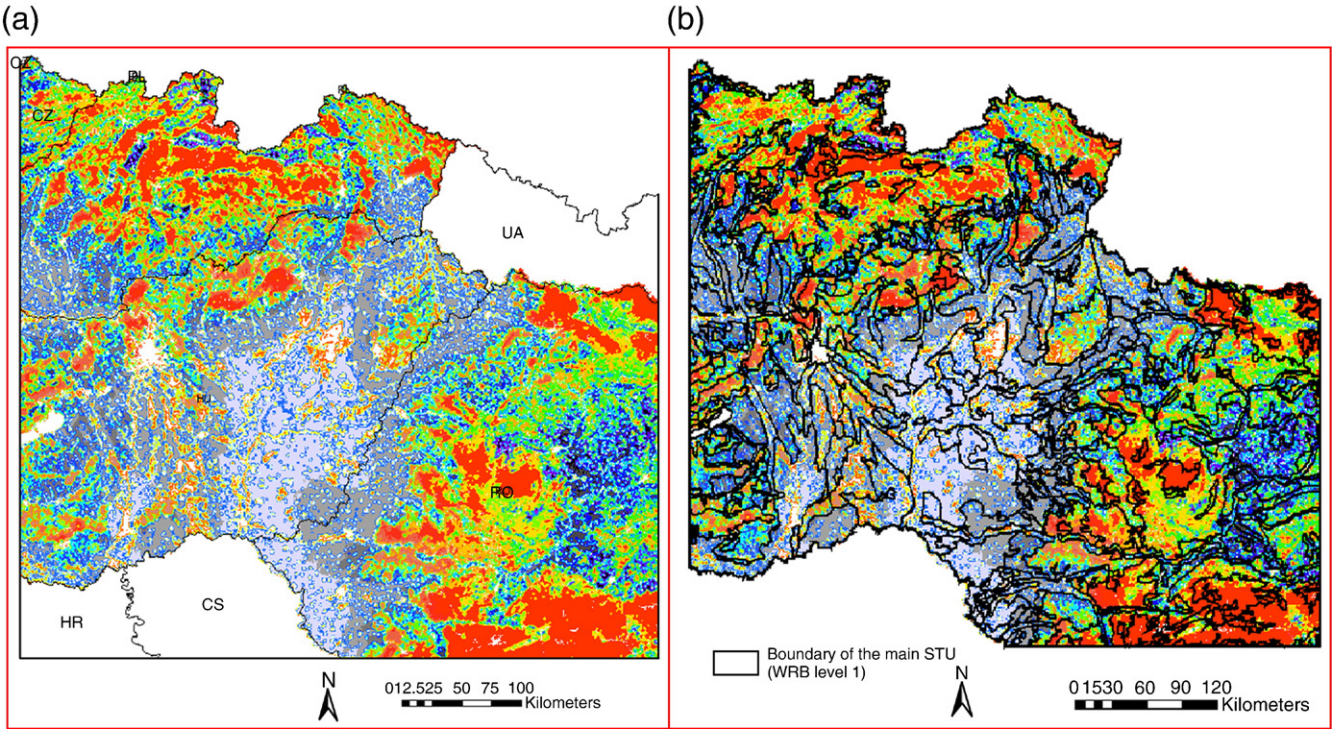

Fig. 8. Comparison between the stratification of the covariates map (a) and the soil map of the area at the scale of 1:1M (b).

covariates. If additional resources are available we would like to include further sampling units to improve our sample of the soil cover. These can be found first by removing all locations or grid cells in the environmental raster which are included in the legacy sample. The remaining grid cells are ordered in terms of their degree of under-observation. Sampling units are first placed at random in strata with no sampling units but in order of decreased density of covariates, i.e., ordered by decreasing density ratio. This will try to ensure that the hypercube is as maximally occupied as possible.

2.2.2. The HISQ algorithm

For p covariates with a raster of size r, Q quantile ranges, and n sampling units in the legacy sample.

The HELS algorithm provides s strata equal to Q^p.

The legacy sampling units n_s contained in strata s, are weighted according to their density (n_s/n) and the covariate density (M_s/r).

In order to add k sampling units in the area, we have to:

(1) Check if some quantile ranges are missing soil samples (some strata s are empty of legacy data). If:
(2.a) There are s_e empty strata.
(2.a.1) Calculate the covariate density (Ms_e/r).
(2.a.2) Check the total number of empty strata Ns_e. If:
(3.a) $Ns_e<k$, each stratum s_e have to be filled by at least one sampling unit. To choose the raster to be filled by one sample unit:
(4.1) For the Ns_e sampling units, obtain the coordinates of each grid cell rs_e contained in the stratum s_e with the highest density.
(4.2) Choose a grid cell randomly within the highest density strata.
(4.3) For the $k-Ns_e$ other samples to choose, sort the covariate density (Ms/r) of each stratum s from the highest to the lowest.
(4.4) Repeat (4.1) and (4.2) until having k sampling units.
(3.b) $Ns_e>k$, sort out the covariate density (Ms_e/r) of each strata s_e from the highest density to the lowest one.
(4.1) For the Ns_e sampling units, obtain the coordinates of each grid cell rs_e contained in the strata s_e with the highest density.
(4.2) Choose a grid cell randomly within the highest density strata.
(4.3) Repeat (4.1) and (4.2) until we having k additional sampling units.
(2.b) There are no empty strata.
(2.b.1) Calculate the covariate density (Ms/r) of each stratum s.
(2.b.2) Sort the density (Ms/r) from the highest to the lowest.
(2.b.3) Obtain the coordinates of each grid cell r_s contained in the stratum s with the highest density.
(2.b.4) Choose a grid cell randomly within the highest density strata.
(2.b.5) Repeat the steps (2.b.3) and (2.b.4) until having k additional sampling units.

3. Examples and applications

Two examples were used to illustrate and discuss the methodology. The first one is an Australian study area, the Edgeroi dataset, and the second one, which is much larger, deals with a small part of the Danube Basin.

3.1. The Edgeroi area, NSW, Australia

The study site is located at Edgeroi area, near Narrabri, NSW, Australia. It is a typical part of the North-western slopes and plains of New South Wales (Ward, 1999).

3.1.1. The Edgeroi dataset

The soil dataset consists of 341 soil profiles: 210 are sampled on systematic, equilateral triangular grid with a spacing of 2.8 km and 131 are distributed randomly or according to transects (Fig. 1). Morphological, physical and chemical attributes are described in the soil dataset.

The covariate dataset is represented by:

- Altitude from a 25 m resolution DEM.
- Compound topographical index (Gessler et al., 1995).
- Potassium content derived from airborne gamma-radiometric survey.
- Clay Index derived from Landsat ETM (band 5/band 7).

All data layers were interpolated to a common grid of 25 m resolution. The size of the raster layer is 2006 by 1425 grid cells.

3.1.2. Results of the hypercube sampling

For the Edgeroi area, each of the four covariates is divided into four quantiles. The hypercube character space has 4^4, i.e. 256 strata. The average number of grid cells within each stratum is therefore 11 166 grid cells.

We now present the indices we can derive from the HELS algorithm results. They are:

- the covariate index (CI) of the grid cells in each stratum corresponding to $CI=(M_s/\overline{Ms})$, where M_s is the number of grid cells in the stratum s and $\overline{Ms}$ is the average number of grid cells per stratum. This index allows us to explain if high or low relative weight of the legacy sampling units (see below) is due to soil sampling or covariate density ($CI<0.75$ means low covariate density, $CI>1.5$ means high covariate density);
- the relative weight of the legacy sampling units in each stratum (*cf. (10) of the HELS algorithm*), $r_w=1/w$, where $w=(n_s/n)/(M_s/r)$ i.e. the ratio between the density of real observations (n_s/n) and covariate hypercube cells (M_s/r) in each stratum (if the density of real observations per covariate hypercube cell is large then the relative weight is small which means that no further sampling is necessary).

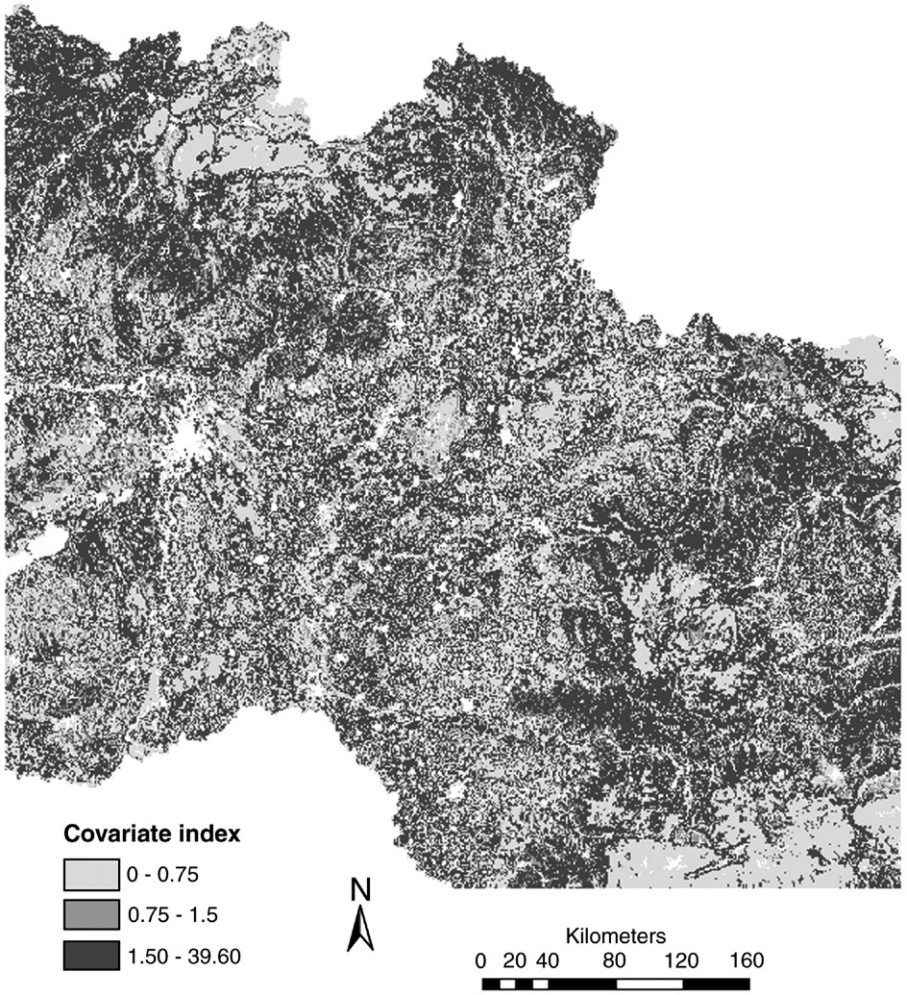

Fig. 9. Covariate index of the hypercube strata in part of the Danube Basin.

The map of the covariate index (Fig. 2) shows some patterns representative of the covariate variability. The values of the covariate index range between 0.001 and 9.475. This means that some strata are very dense compared to others. The strata with the lowest density are in the talwegs of medium altitude, in the areas of high clay content in low altitudes and in dry areas in medium altitude.

This figure (Fig. 2) can explain the distribution of the weights of the legacy samples (Fig. 3).

Of the 256 strata, 16% represent undersampling, 38.4% adequate sampling and 45.7% oversampling. The oversampling appears in the research experiment areas where all soil sample units are clustered. Usually, undersampling appears above all in the plain area but it is most of the time surrounded by well-sampled areas. Sampling according to transects seems a good way to obtain adequate sampling (see the transects in Fig. 3). In regards to the covariate index, most of undersampling appears in high covariate index, where soil covariates are then highly variable. Actually, it is difficult to collect field samples in these highly variable areas. But the reciprocity is not true: adequate and oversampling can also appear in the high covariate index areas. This is due to regular grid sampling which appears not to be the best strategy for field sampling (pedodiversity is not taken into account for field sampling).

One hundred seventeen strata out of 256 were empty of legacy sample units. The number of quantiles p was chosen according to the number of original sampling units in order to avoid too much field survey which is very costly. Then, only 100 new sampling units were added to the legacy sample (Fig. 4). They are selected in order of decreasing covariate raster density. Usually, the grid cells belonging to strata with highest covariate raster density are spatially clustered. This means that feature space also reflects geographic space (the covariate are spatially and feature continuous variables). The new sampling units were chosen at random from within each hypercube stratum, i.e., one of the covariate raster points within the particular stratum is chosen (with uniform probability). This has an associated location in the geographic space which is the new sampling unit location. 85.4% (100/117) of the empty strata were filled by new sampling units chosen in this manner. This gave an 'enhanced' sample with 441 sampling units.

We recalculated the weights of all the sampling units in the enhanced sample (Fig. 5), and compared the quantiles and the statistics of the weights distribution of the legacy sample and the enhanced one (Table 1).

The weights of the sampling units in the enhanced sample are more distributed between 0.75 and 1.5 (Fig. 6). Since these values represent adequate sampling, the area becomes better sampled with the enhanced sample.

Also, adding more sampling units changes the mean of the weights. The proportion of weight values less than 0.75 (representing undersampling) increases. The number of units

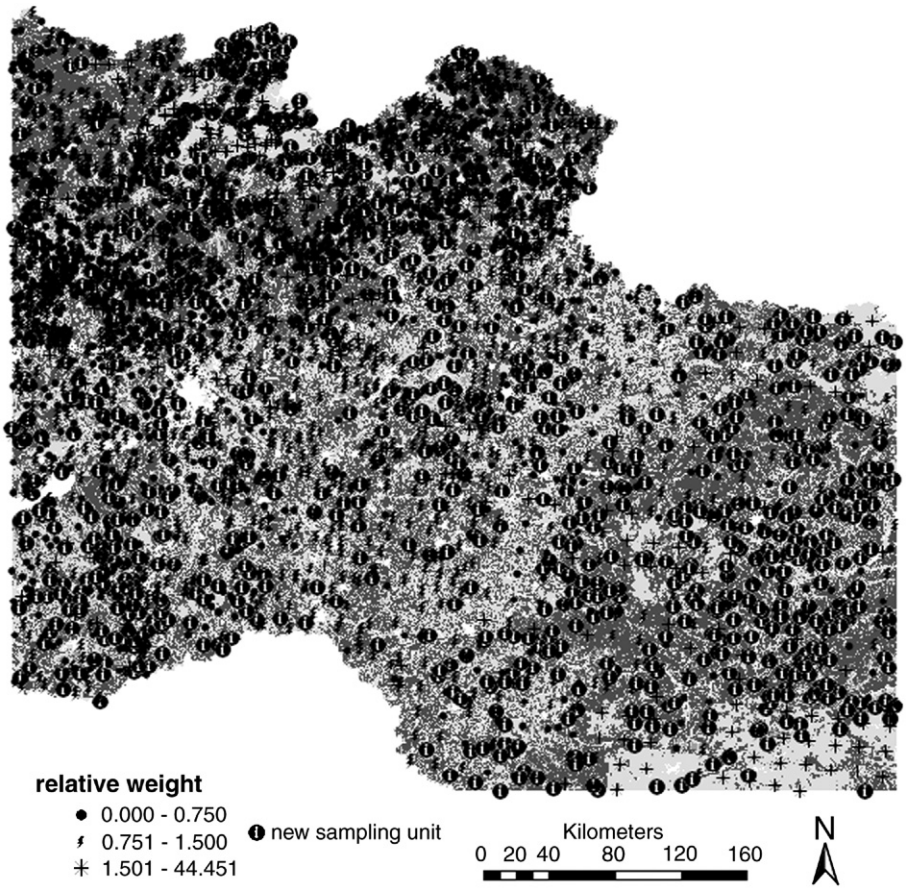

Fig. 10. The locations of the 1000 new (additional) sampling units in part of the Danube Basin.

representing oversampling decreases by 13% when adding the new sampling units. This can be explained by the fact that adding new sampling units in the originally unsampled areas homogenises the distribution of sampling units in the area. Oversampling globally decreases, whereas undersampling areas tend to increase.

3.2. The Danube Basin area

The Danube Basin is located in Eastern Europe and covers 11 countries which are: Germany (DE), Austria (AU), Czech Republic (CZ), Slovak Republic (SL), Hungary (HU), Croatia (HR), Slovenia (SI), Bulgaria (BU), Romania (RO), Moldova (MO) and Ukraine (UA).

3.2.1. The Danube Basin dataset

In this area, we studied an area of 536 km by 536 km (Fig. 7) on a 180-m square raster. It contains 2945 soil sampling units described by their physical, chemical and morphological attributes.

This legacy sample covers parts of Slovakia, Hungary and Romania. In the Romanian part, the data have been regularly sampled on a 16-km square grid (Fig. 7) because they are part of the UNECE ICP (United Nations Economic Commission for Europe International Collaborative Program on Effects of Air Pollution on Natural Vegetation and Crops (Lorenz et al., 2005). For Hungary and Slovakia, the soil data are irregularly sampled.

The covariates that are used for this study are:

- altitude from the SRTM (Shuttle Radar Topographic Mission) at about 90-m resolution;
- slope derived from the altitude;
- compound topographical index derived also from the DEM;
- landuse from Corine Land Cover (Büttner et al., 2002). Only the agricultural landuses were taken into account for this study. We allocate to each pixel the proportion of grid cells in a 10 by 10 neighborhood which is agricultural. This gives a quantitative variable.

All data layers were resampled to a common grid of 180-m resolution. The whole area is composed of 2979 by 2979 raster grid cells.

3.2.2. Results of hypercube sampling

We choose 6 quantiles to divide each feature space of the 4 covariates. The hypercube feature space therefore comprises 6^4 or 1296 strata. The average quantity of grid cell within each stratum is therefore 6847 grid cells.

Because in this area a soil map is existing, we can test the reliability of the soil covariates by comparing the limits of the

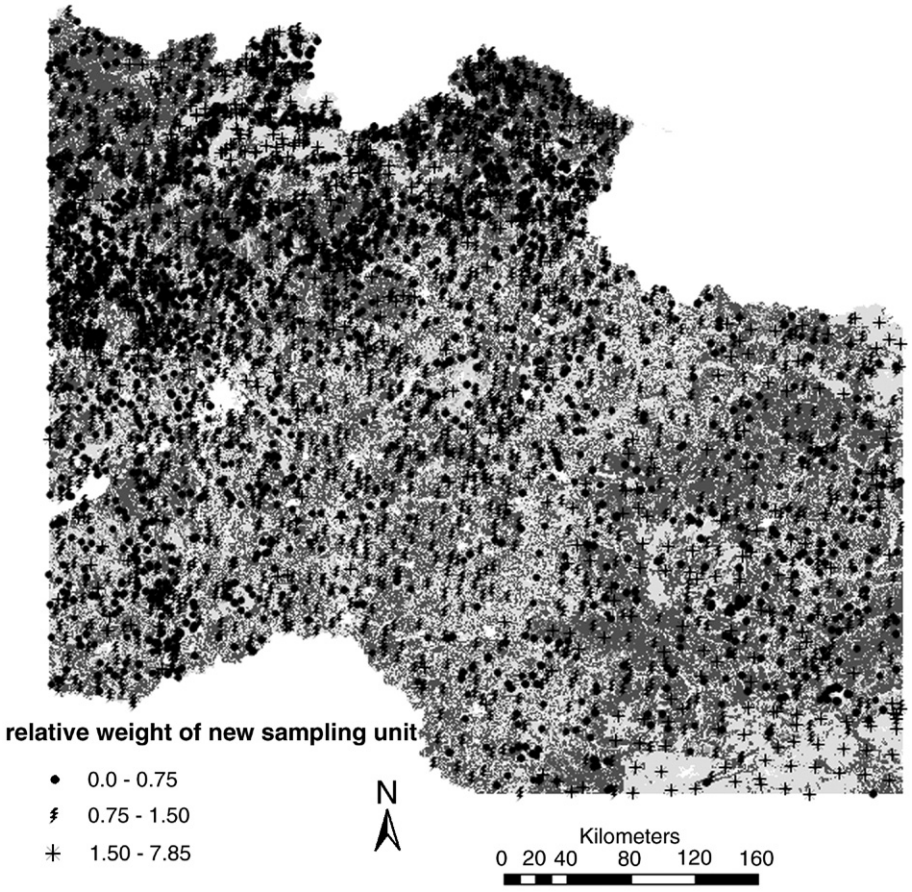

Fig. 11. The relative weights of the enhanced sample (2945 legacy plus 1000 additional sampling units) in part of the Danube Basin.

strata (Fig. 8a) with the ones of the major soil types in Europe (Lambert et al., 2003) described according to the first level of the WRB (IUSS Working Group WRB, 2006) at the 1:1M scale (Fig. 8b) by traditional soil survey. In Fig. 8b, we overlaid the strata (of Fig. 8a) with the limits of the major Soil Typological Units (STU) of the 1:1M scale soil map (in black).

Most of the time, the limits show the same patterns above all the mountainous areas of Slovakia and Romania. However, some differences appear in the Hungarian plain: the small depression which appears with the landform and with the landcover attributes is not considered as the main source of soil differentiation. We can imagine that other soil forming processes have to be taken into account in this area like land management practices. But since we don't have any data on the accuracy of the soil map, we cannot conclude that the covariates we used are insufficient for the study. The accuracy of the soil map has to be verified first but it is not the purpose of the study.

The covariate index map of the area (Fig. 9) shows very diverse distribution of covariate index. The low covariate indices are located in areas with extreme values: high compound topographical index with high altitude and with high percentage of agricultural areas, and low compound topographical index with low altitude and with low percentage of agricultural areas. The areas with high slope percentages present high covariate index. This means that for the slope, the covariate density is high and in areas with extreme values of altitude, compound topographical index and agricultural percentage, the covariate density is low. This fits with general pedodiversity distribution model: slopes are very pedodiverse and plateau and plains present usually less pedodiversity.

Of the 2925 soil sampling units, 27% represent undersampling (relative weight upper than 1.5) and 47% oversampling (relative weight lower than 0.75). Oversampling is very present in the mountainous areas, in Slovakia where sampling is not regular and undersampling is more represented in Romania where the sampling is regular. When covariate density is low, sampling units are representing oversampling. Like for Edgeroi, we can conclude that regular grid sampling is not appropriate in areas where soil covariates are highly variable. In Hungary, the sampling units tend to be more oversampled because this plain area is quite homogeneous.

The map of the relative weights of the legacy sampling units (Fig. 10) is concordant with the previous explanation since all the samples representing undersampling are in the mountainous areas and the depression.

The interesting result is that sampling with a regular grid of 16 km in the mountainous area often leads to an undersampling of the area. In a more homogeneous area, like the plains, this grid seems sufficient.

New sampling can be designed by prioritising the empty strata. One thousand additional sampling units are added to the legacy sample to give an enhanced sample. First, the empty

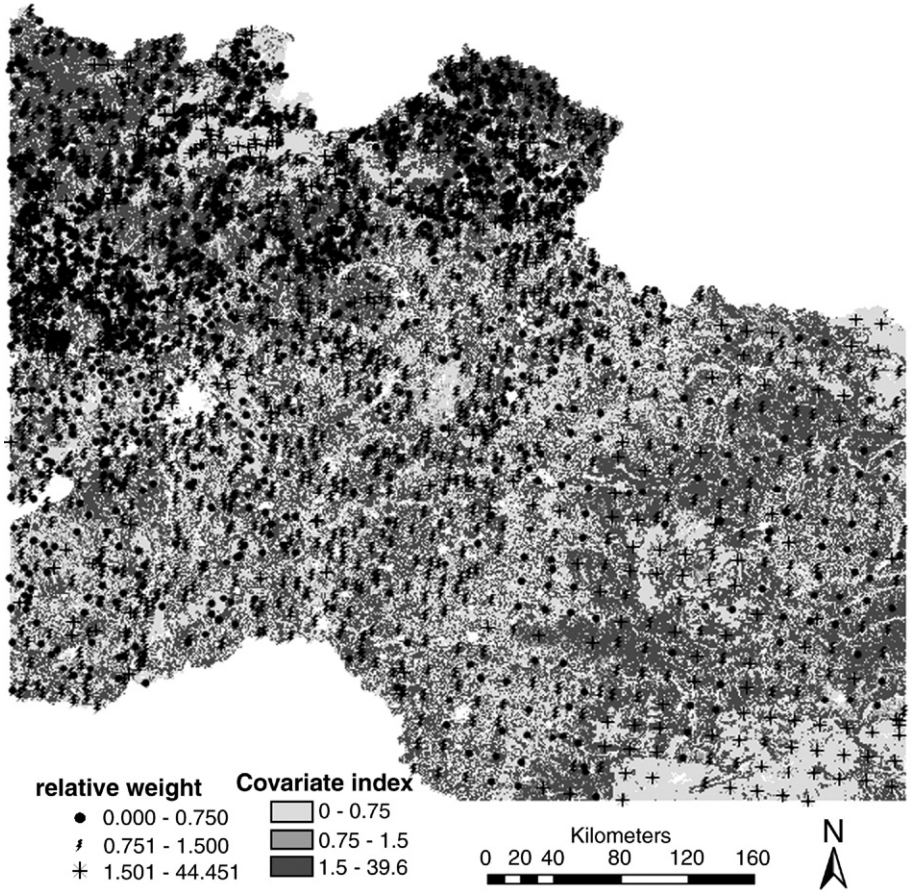

Fig. 12. The relative weight of the legacy sampling units in part of the Danube Basin.

Table 2
Comparison of the distribution of relative weights between the enhanced sample and the legacy sample for part of the Danube Basin

	Legacy sample	Enhanced sample
Values		
Number of sampling units	2 925	3 925
Mean of relative weights	0.58	1.02
Standard deviation of the relative weights	1.32	1.63
Percentage of sampling units with a relative weight <0.75 (representing oversampling)	46.6%	10%
Percentage of sampling units with a relative weight >1.5 (representing undersampling)	27%	52%

strata are filled and then, the 143 (1000 − 857) highest covariate raster densities are filled randomly. This means that all the feature space strata are represented by at least one sample unit.

As for the Edgeroi area, it is possible to map the relative weights of the enhanced sample (Fig. 11).

Compared with the legacy sample (Fig. 12), the sampling units are more evenly distributed with less oversampling. The sampling units are now predominantly considered to be well-sampled (relative weight 0.75–1.5). This result can be evidenced in Table 2 and Fig. 13.

Table 2 shows that, for the enhanced sample, the mean value of the relative weights is now closer to 1, which means that there is globally a good sampling of the area. The percentage of oversampled units is largely reduced going from 46.6% (legacy sample) down to 11% (enhanced sample).

In the legacy sample, the relative weights were largely heterogeneous and quite far from 1.0 (Fig. 13) whereas in the enhanced sample, the values were very closed to 1.0 and the distribution seemed quite normal.

The conclusions for the enhanced sampling in the two study areas are similar: attempting to cover the whole feature space (hypercube sampling strategy) leads to:

- a more homogeneous distribution of relative density between sample units and covariates;
- better sampling locations even if the sampling was done according to the feature space and not to the geographical space.

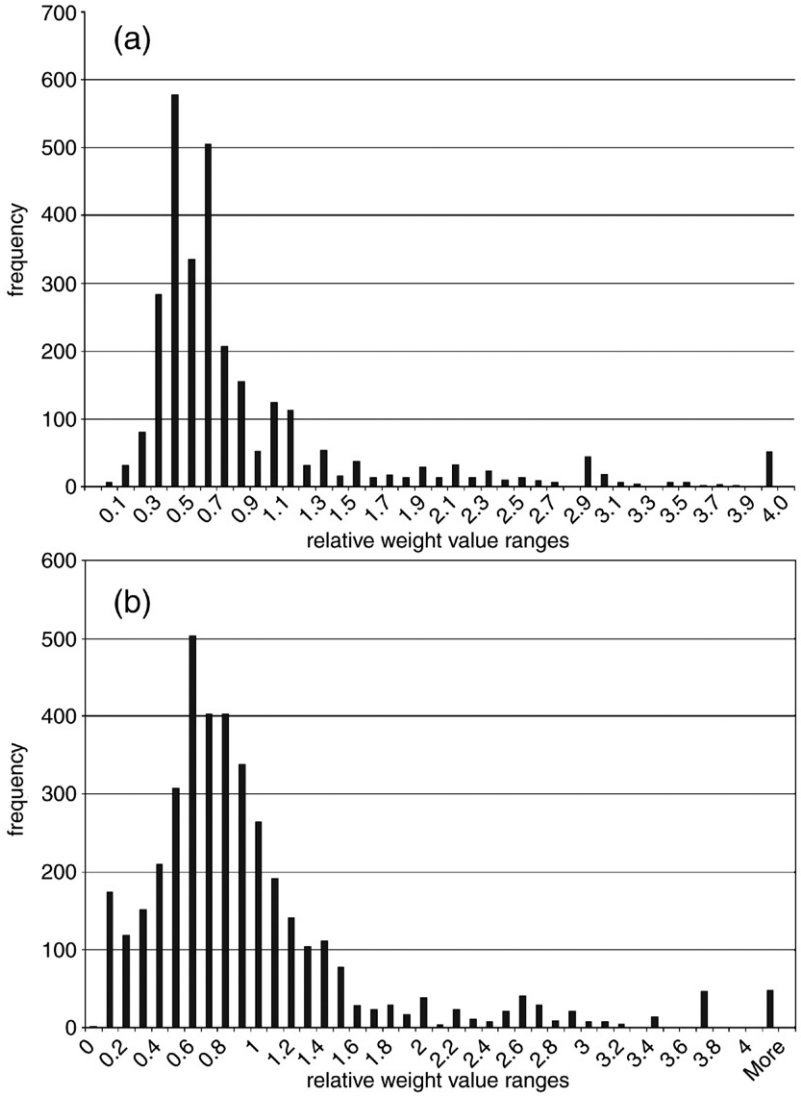

Fig. 13. Comparison of the distributions of the relative weights of the legacy sample (a) and the enhanced sample (b) in part of the Danube Basin.

When the legacy soil sampling units present low mean relative weight with high standard deviation (for the Edgeroi area, see Table 1), we advice to add more than one third of new sampling units (if the budget is sufficient) in order to increase the mean and to decrease the standard deviation.

4. Discussion

Since, we assume that soil attributes to be mapped can be predicted by the environmental covariates, our estimation of the legacy sample units is based on the covariates. Then, the results are very dependent on the covariates (number and spatial resolution of the covariates and quality of their measurement or description). If the covariate resolution is more precise, the distribution will be smoothed. Then, only a small number of quantiles must be chosen. In the future, we should test the sampling using different sets of covariates. What would the results be if we add one covariate or if we change the spatial resolution of the covariates?

Also, only the feature space has been used for the evaluation of the sampling. Is the addition of geographic coordinates useful or not? This can also be tested in the future by firstly introducing the covariate coordinates in the hypercube and secondly, weighting them in order not to overemphasize the geographical space.

For the enhanced sample, most of the new relative weights are close to 1, which means that the new sample units well represent the covariate variability.

The problem that appears when new samples are added is that the percentage of sample units that represent undersampling increases. The user has to bear in mind that the relative weight depends on the number of sampling units added. The best indicator is the mean value of the relative weights. In both cases, the mean increases, so the relative importance of undersampled units increases too.

5. Conclusions

Hypercube sampling provides a mean to evaluate adequacy of a legacy sample according to the soil covariates. The main advantage of such a method is that all the legacy sample units can be estimated according to their density in the feature space that represents soil variability. From the results, it is possible to add new sampling units in order to:

- cover the whole feature space if some parts are missing
- enhanced some parts of the feature space that appear to be undersampled.

The algorithms described here allow the digital soil mapper to test and quantify the distribution of soil variables according to the covariates. The HELS and HISQ algorithms have several advantages, among them:

- they take into account the detailed information of covariates;
- they are not restricted to a specific number of covariates;
- they can be applied in any area (for which information is available).

The methodology described here only deals with the covariate feature space, in the future, we should test the advantage or disadvantage of additionally taking geographic information into account for evaluating the legacy sample and enhancing it if required. This methodology can be also compared to other sampling methods (use of kriging variance, minimization of the mean of the shortest distances…).

Acknowledgment

This work was supported by an Australian Research Council Discovery Grant entitled “Digital Soil Mapping”. The Danube Basin database used here has been collated by the Land Management Unit of the Joint Research Centre of the European Commission from contributions from agencies in the Czech Republic, Hungary, Romania and Slovakia.

References

Bisbal, J., Grimson, J., Bell, D., 2005. A formal framework for database sampling. Information and Software Technology 47, 819–826.

Brus, D.J., de Gruitjer, J., van Groeningen, J.W., in press. Chapter 14. Designing purposive and random spatial coverage samples by the K-means clustering algorithm. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Initial Perspective. Developments in Soil Science 31. Elsevier Amsterdam, 185–192.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79–94.

Büttner, G., Feranec, J., Jaffrain, G., 2002. CORINE land cover update 2000. Technical Guidelines. EEA Technical Report, vol. 89. European Environment Agency, Copenhagen, Denmark. 56 pp.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.

Fereyra, R.A., Apezteguia, H.P., Sereno, R., Jones, J.W., 2002. Reduction of soil water spatial sampling density using scaled semi-variograms and simulated annealing. Geoderma 110, 265–289.

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modeling and spatial prediction of soil attributes. International Journal of Geographical Information Science 9, 421–432.

Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Geoderma 120, 75–93.

Heuvelink, G., Brus, D., de Gruijter, J., in press. Chapter 11. Optimisation of sample configurations for digital soil mapping with universal kriging. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Initial Perspective. Developments in Soil Science 31, Elsevier Amsterdam.

IUSS Working Group WRB, 2006. World reference base for soil resources 2006, 2nd ed. World Soil Resources Reports, vol. 13. FAO, Rome.

Lambert, J.J., Daroussin, J., Eimberck, M., Le Bas, C., Jamagne, J., King, D., Montanarella, L., 2003. Soil geographical database for Eurasia and the Mediterranean: instructions guide for elaboration at scale 1:1,000,000. Version 4.0. EUR 20422 EN. Office for Official Publications of the European Communities, Luxembourg. 64 pp.

Lark, R.M., 2000. Designing sampling grids from imprecise information on soil variability, an approach based on fuzzy kriging variance. Geoderma 98, 35–59.

McBratney, A., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Minasny, B., McBratney, A.B., in press. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computer & Geosciences.

Lorenz, M., Becher, G., Mues, V., Fischer, R., Becker, R., Calatayud, V., Dise, N., Krause, G.H.M., Sanz, M., Ulrich, E., 2005. Forest condition in Europe. Technical Report 2005. ISSN 1020-3729UNECE, Geneva. 96 pp.

Simbahan, G.G., Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma 133, 345–362.

Vanclay, J.K., Skovsgaard, J.P., Hansen, C.P., 1995. Assessing the quality of permanent sample plot databases for growth modeling in forest plantations. Forest Ecology and Management 71, 177–186.

Van Groeningen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal for Environmental Quality 27, 1078–1086.

Ward, W.T., 1999. Soils and landscapes near Narrabri and Edgeroi, New South Wales, with data analysis using fuzzy k-means. CSIRO Division of Soils Divisional Report.

Warrick, A.W., Myers, D.E., 1987. Optimisation of sampling locations for variogram calculations. Water Resources Research 23, 496–500.

Xu, C., He, H.S., Hu, Y., Chang, Y., Li, X., Bu, R., 2005. Latin hypercube sampling and geostatistical modeling of spatial uncertainty in a spatially explicit forest landscape model simulation. Ecological Modelling 185, 255–269.