European Journal of Soil Science

European Journal of Soil Science, June 2011, **62**, 394–407 doi: 10.1111/j.1365-2389.2011.01364.x

Sampling for validation of digital soil maps

D.J. Brus, B. Kempen & G.B.M. Heuvelink

Soil Science Centre, Wageningen University and Research Centre, PO Box 47, 6700 AA Wageningen, The Netherlands

Summary

The increase in digital soil mapping around the world means that appropriate and efficient sampling strategies are needed for validation. Data used for calibrating a digital soil mapping model typically are non-random samples. In such a case we recommend collection of additional independent data and validation of the soil map by a design-based sampling strategy involving probability sampling and design-based estimation of quality measures. An important advantage over validation by data-splitting or cross-validation is that model-free estimates of the quality measures and their standard errors can be obtained, and thus no assumptions on the spatial auto-correlation of prediction errors need to be made. The quality of quantitative soil maps can be quantified by the spatial cumulative distribution function (SCDF) of the prediction errors, whereas for categorical soil maps the overall purity and the map unit purities (user's accuracies) and soil class representation (producer's accuracies) are suitable quality measures. The suitability of five basic types of random sampling design for soil map validation was evaluated: simple, stratified simple, systematic, cluster and two-stage random sampling. Stratified simple random sampling is generally a good choice: it is simple to implement, estimation of the quality measures and their precision is straightforward, it gives relatively precise estimates, and no assumptions are needed in quantifying the standard error of the estimated quality measures. Validation by probability sampling is illustrated with two case studies. A categorical soil map on point support depicting soil classes in the province of Drenthe of the Netherlands (268 000 ha) was validated by stratified simple random sampling. Sub-areas with different expected purities were used as strata. The estimated overall purity was 58% with a standard error of 4%. This was 9% smaller than the theoretical purity computed with the model. Map unit purities and class representations were estimated by the ratio estimator. A quantitative soil map, depicting the average soil organic carbon (SOC) contents of pixels in an area of 81 600 ha in Senegal, was validated by random transect sampling. SOC predictions were seriously biased, and the random error was considerable. Both case studies underpin the importance of independent validation of soil maps by probability sampling, to avoid unfounded trust in visually attractive maps produced by advanced pedometric techniques.

Introduction

In recent years, digital soil mapping activities have strongly increased all over the world and, as a result, there are many soil maps produced by quantitative (statistical) methods. Some recent examples are Bui & Moran (2003), Thompson & Kolka (2005), Mora-Vallejo *et al*. (2008) and Taylor *et al*. (2009). Although the educated user knows that digital soil maps, similarly to conventional soil maps, are not perfect and contain errors, they are also frequently not validated (Grunwald, 2009). However, the impurity of soil map units is often large, meaning that in a large part of the area the soil type as depicted on the map does not correspond with the true soil type. Thematic soil maps depicting a single, quantitative (discrete or continuous) soil property such as pH, clay content or stoniness, are imperfect as well, so that the values depicted on the map will deviate to a greater or lesser extent from the true value.

The determination and quantification of the quality of conventional soil maps has been a research topic for many years. Early papers on this include Webster & Beckett (1968), Bie & Beckett (1971), Burrough *et al*. (1971) and Beckett & Bie (1978). These were followed by work that aimed more explicitly at the quantification of the accuracy of conventional soil maps (Steers & Hajek, 1979; van Kuilenburg *et al*., 1982; de Gruijter & Marsman, 1985). The approaches outlined in these publications are still valid and useful today for the validation of digital soil maps, but more recent developments in map validation outside soil science are important too. It is worthwhile to incorporate these developments and to present a comprehensive strategy for validation of digital soil maps.

The quality of a soil map can be determined by comparing the predictions at the calibration sites with the observed values.

Correspondence: D.J. Brus. E-mail: dick.brus@wur.nl

Received 14 December 2009; revised version accepted 25 February 2011

However, the accuracy thus obtained, referred to as the internal accuracy, often over-estimates the actual accuracy. Therefore, predictions are preferably compared with independent data not used in the modelling (Chatfield, 1995). This accuracy is referred to as the external or test accuracy.

If the soil map is produced by a statistical method, then a direct estimate of the quality of the soil map is also obtained, well-known examples are the kriging variance and theoretical purity (Webster & Oliver, 2007; Brus *et al*., 2008). However, these direct statistical estimates rely on the model-assumptions used in mapping, and may therefore be biased. This paper focusses on determination of the external accuracy of soil maps, which can be estimated by several methods such as data-splitting, cross-validation and additional probability sampling.

The aims of our paper are, first, to review map quality measures; second, to show how, if the calibration data are a non-probability sample, validation by additional probability sampling is superior to validation by data-splitting or cross-validation with the calibration data; third, to evaluate basic types of probability sampling design for validating soil maps, to explain how we can decide on the sample size and how the quality measures can be estimated; and fourth, to illustrate validation by additional probability sampling with two case studies.

Measures of map quality

Point support versus block support

Digital soil maps typically are raster maps. It is important to be aware of what the values of the soil variable assigned to pixels represent. There are two possibilities. The first possibility is that the value of a pixel represents the value at any point location within that pixel. Such a soil map is referred to as a soil map on point support. The second possibility is that it represents the average value of a quantitative soil attribute, or the dominant soil class with the largest area within the pixel. Such soil maps are referred to as maps on block support.

This paper is restricted to raster maps with pixels having exactly one value. We do not consider validation of, for instance, fuzzy soil class maps, where each pixel has partial membership in several classes simultaneously, nor soil class maps depicting the predicted areal proportions of soil classes within pixels. Different methods have been proposed for comparison and validation of soft-classified maps, see for instance Woodcock & Gopal (2000), Lewis & Brown (2001) and Pontius & Cheuk (2006).

Quality measures for quantitative soil maps

The difference between the predicted value at a location and the true value at that location (the value that would be observed or measured by an errorless measurement device) is considered as the prediction error:

$$e(\mathbf{s}) = \hat{z}(\mathbf{s}) - z(\mathbf{s}), \tag{1}$$

where $\hat{z}(\mathbf{s})$ denotes the predicted soil property at location $\mathbf{s}$, and $z(\mathbf{s})$ denotes the true value of the soil property at that location. The quality of soil maps depicting quantitative soil properties can be expressed by the spatial cumulative distribution functions (SCDF) of the error, absolute error and squared error. The SCDF of the error is defined as:

$$F(t) = \frac{A_{e \leq t}}{A}, \tag{2}$$

where $A_{e \leq t}$ denotes the surface area of the region where the value of the error is smaller than or equal to the threshold error t, and A denotes the total surface area. If the soil map is on block support, the errors in Equation (1) are defined as the difference between the predicted grid cell average of the soil property of interest and the measured (true) grid cell average. In this case, the SCDF is defined as the proportion of pixels with an (absolute or squared) error smaller than or equal to t. Note that the total number of pixels on the soil map is finite (be it in general very large), whereas the number of point locations is infinite. By replacing the errors in Equation (2) by their absolute or squared values, we obtain the SCDFs of the absolute and squared errors.

Commonly, only parameters of the SCDFs are reported, such as the mean error (bias), the variance of the error, the mean absolute error and the (root) mean squared error. Interest may also be in some quantile of the SCDF of the (squared or absolute) error, such as the 50th percentile (median error) or the 90th percentile of the error.

In some cases, we may be interested in the SCDF or its parameters for sub-areas, for instance, geomorphic units or soil landscapes. If a categorical soil map is available, it is natural to choose the map units of this soil map as sub-areas.

Quality measures for categorical soil maps

For assessing the quality of categorical soil maps, which depict a qualitative variable measured on the nominal or ordinal scale, several quality measures have been proposed, all based on the error or confusion matrix. The most important quality measures are (Congalton, 1991)

- **(i)** overall accuracy (overall purity, map purity),
- **(ii)** user's accuracy,
- **(iii)** producer's accuracy,
- **(iv)** kappa coefficient of agreement and
- **(v)** weighted overall, user's and producer's accuracies.

Stehman (1997a) describes the properties and relationships between these quality measures. He argues that quality measures should be directly interpretable in terms of the probability of occurrence of a misclassification, and for that reason recommended the overall accuracy, user's accuracy and producer's accuracy as quality measures. Hereafter, we will define and use these quality measures only.

Lark (1995) discusses map quality measures (i)–(iii) extensively, and elaborates on which quality measures are helpful for answering different types of questions. He concludes that the map quality measure that determines the utility of the map for a specific use depends on how the questions are framed and qualified. Lark (1995) also questions the appropriateness of the terms user's and producer's accuracies, as both quality measures can be important for users as well as producers. He proposes using the terms 'map unit purity' for the user's accuracy, and 'class representation' for the producer's accuracy, which we adopt here, as well as the term overall purity for overall accuracy.

Table 1 shows an error matrix. Note that the entries of the matrix are true surface areas. In such a case the error matrix is also referred to as the population error matrix. Note also that the row marginals (the areas covered by the map units) of the population error matrix are known, whereas the column marginals (the areas covered by the true classes) are unknown, and must be estimated from the sample. It is important to distinguish the population error matrix from a sample error matrix. The latter has the number of locations (pixels) in the sample as entries. These numbers can be used to estimate the entries in the population error matrix, that is the surface areas. The estimation method depends on the sampling design (see later).

For categorical soil maps on point support, the overall purity is defined as the proportion of the mapped area in which the predicted soil class, which is the soil class as depicted on the map, equals the true soil class. In other words, it is the areal proportion correctly classified:

$$p = \sum_{u=1}^{U} A_{uu}/A \ , \tag{3}$$

where U denotes the number of classes, A_{uu} denotes the surface area of the correctly classified part of map unit u and A denotes the total surface area (Table 1).

For categorical maps on block support the overall purity is defined as:

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i , \tag{4}$$

Table 1 Population error matrix

		Field				
		1	2	⋯	U	$\sum$
Map	1	A_{11}	A_{12}	⋯	A_{1U}	A_{1+}
	2	A_{21}	A_{22}	⋯	A_{2U}	A_{2+}
	.	.	.	⋯	.	.
	.	.	.	⋯	.	.
	.	.	.	⋯	.	.
	U	A_{U1}	A_{U2}	⋯	A_{UU}	A_{U+}
	$\sum$	A_{+1}	A_{+2}	⋯	A_{+U}	A

A_{ij} = surface area mapped as class c_i with observed soil class c_j.

where N denotes the total number of blocks (pixels), and y_i is defined as an indicator variable equal to 1 if the predicted soil class $\hat{c}_i$ in block i equals the true dominant soil class in this block, $c_i^{(\mathrm{dom})}$, and otherwise zero.

The purity can also be defined at the level of the map units, leading to the proportion of the area of a map unit correctly classified, referred to as the map unit purity for mapped class u:

$$p_u = \frac{A_{uu}}{A_{u+}}, \tag{5}$$

where A_{u+} denotes the surface area mapped as soil class u (Table 1). The complement of p_u, $1 - p_u$, is referred to as the error of commission for mapped class u. It is the proportion of the area incorrectly mapped as class u.

The class representation for soil class u is the proportion of the area where in reality soil class u occurs that is also mapped as class u:

$$r_u = \frac{A_{uu}}{A_{+u}}, \tag{6}$$

where A_{+u} denotes the surface area with true soil class u, see Table 1; r_u is also referred to as the sensitivity, and its complement, $1 - r_u$, is referred to as the error of omission, i.e. the proportion of the area with true class u not mapped as class u.

The definitions of the map unit purity and class representation for maps on block support can be obtained by replacing A_{uu}, A_{u+} and A_{+u} in Equations (5) and (6) by N_{uu} (total number of blocks with predicted and true dominant soil class u), N_{u+} (total number of blocks with predicted dominant soil class u), and N_{+u} (total number of blocks with true dominant soil class u), respectively.

Clearly, a trade-off exists between the information content and the purity of a categorical soil map. Soil classifications typically are hierarchical, and the purity depends on the level in the hierarchy: the purity decreases with the level. For this reason, Marsman & de Gruijter (1986) proposed that the quality of categorical soil maps be quantified by both purity and homogeneity. Homogeneity can be quantified, for instance, by the standard deviation of quantitative soil properties within map units:

$$S_u(z) = \sqrt{\frac{1}{A_{u+}}\int_{\mathbf{s}\in\mathcal{A}_u} (z(\mathbf{s}) - \overline{z}_u)^2 \, \mathrm{d}\mathbf{s}}, \tag{7}$$

where $z(\mathbf{s})$ denotes the value of the target soil property at location $\mathbf{s}$, $\overline{z}_u$ denotes the mean of soil property z in map unit u, $\mathcal{A}_u$ denotes the area covered by map unit u and A_{u+} is defined as the size of this area.

Validation methods

Data-splitting

Many papers have been published on validation by data-splitting of land-cover and other maps derived from remote sensing images (Friedl *et al*., 2000; Muchoney & Strahler, 2002; Foody, 2002).

A recent example on validation of a soil map by data-splitting is given by Grinand *et al*. (2008). A problem of validation by data-splitting is that it is far from evident how to split a data set such that unbiased and valid estimates of map accuracy are obtained. The problem is caused by the spatial auto-correlation of prediction (classification) errors. Soil maps are frequently calibrated on existing data, collected during earlier soil surveys. In some cases, the soil map is calibrated on data collected with a sampling design tailored at this specific aim by purposive (targeted) sampling on the basis of the ancillary variables that may be derived from a digital elevation model and airborne gamma-ray imagery for example. Locations are typically selected such that a good spread in property space is achieved, so that the whole range of values of the predictors is present in the calibration sample (Brus & Heuvelink, 2007). With legacy and sample data collected with the aim of calibrating a digital soil mapping model, the calibration data do not form a probability sample from the mapped area to be validated.

For a non-probability sample, regardless of how the validation sub-sample is selected, the sub-sample is not a probability sample from the study area. So, even when the validation sub-sample is selected randomly from the data set, it is a non-probability (purposive or haphazard) sample from the study area. This implies that in predicting the spatial mean of the prediction error, the auto-correlation of the errors must be taken into account. However, even when this is done, this still is no guarantee of unbiased estimates. The sample itself can be biased towards, for instance, areas that are easily classified, so that correctly classified units are over-represented in the sample.

Cross-validation

A second option is n-fold cross-validation, the most common form of which is leave-one-out cross-validation (Efron & Tibshirani, 1993). The difference between this and data-splitting is that in cross-validation the splitting is repeated, which makes it more efficient than data-splitting. In leave-one-out cross-validation the data set is split n times into a set of $n - 1$ locations for calibration and one for validation. For each sampling location, the model is refitted leaving that location out of the calibration data set. The target variable is then predicted for that location and the prediction error is computed. This is done for all sampling locations, and the SCDF of the (squared, absolute) error or its parameters are computed.

When the sampling locations are not selected by probability sampling but, for instance, by purposive sampling, then cross-validation suffers from the same problems as data-splitting. The prediction errors will generally be spatially correlated and the sample itself can be biased, so that a valid prediction of the spatial mean of the prediction error is difficult to obtain. If the calibration data are a simple random sample, then the average of the cross-validation error is a nearly unbiased estimate of the spatial mean error or classification error rate (Krzanowski, 2001; Steele *et al*., 2003).

Additional probability sampling

Unbiased and valid estimates of the quality measures can best be obtained by selecting additional test units (measurements at point locations or pixels not used for calibration) by probability sampling (Stehman, 1999). In probability sampling, all units in the study area have a positive probability of being selected, and the selection probability of any combination of sampling units is known (Särndal *et al*., 1992). These probabilities are determined by the sampling design. Early examples of validation of conventional soil maps by probability sampling are White (1966), Steers & Hajek (1979), van Kuilenburg *et al*. (1982), de Gruijter & Marsman (1985) and Marsman & de Gruijter (1986).

In the case of probability sampling the map quality measures and their standard errors can be estimated unbiasedly by 'design-based' inference, by making use of the selection probabilities of samples as determined by the applied sampling design. The attractiveness of these estimators is that they are model-free, so no model-assumptions on the spatial variation of the errors associated with the sampling units need to be made (de Gruijter & ter Braak, 1990; Brus & de Gruijter, 1997). The independence of model-assumptions is referred to as validity. Validity is especially important for validation purposes, as debates on the quality of the estimated map quality measures must be avoided. Stehman (2000) describes the practical implications of probability sampling and design-based inference, and contrasts design-based inference with model-based inference for map accuracy assessment.

Probability samples are often confused with haphazard samples. A haphazard sample is the one that is selected arbitrarily. Sometimes these are referred to as 'more or less random samples'. We must stress, however, that haphazard samples cannot be analysed as if they were probability samples. With haphazard samples the probability of selection of any combination of sampling units is unknown. This makes design-based statistical inference impossible.

Mueller *et al*. (2004) found poor correlations between map accuracy estimates obtained by cross-validation and by additional probability sampling. We conclude here that validation by additional probability sampling is to be preferred over validation by data-splitting and cross-validation. We now describe how to design such a probability sample.

Designing a sampling scheme for validation

Basic sampling units and sample support

In sampling, we must specify the population units that serve as the basic sampling units. These units can either be defined as all point locations in the mapped area, or as all the pixels of a raster map.

The basic sampling units for validation should have the support (size, shape and orientation) of the model output, the soil map. So for a soil map on point support, the basic sampling units are point locations. In this case, sampling by selecting grid cells is

incomplete, as there is an infinite number of point locations within each grid cell. Taking the centres of the selected grid cells is incorrect, because then all other point locations have a zero probability of being selected, and consequently the estimated quality measures pertain to a finite set of point locations and not to the entire study area. In this case, one or more point locations should be selected randomly in each selected grid cell.

To validate a soil map on block support, the pixels serve as basic sampling units. In practice, it is often difficult or not feasible to observe the selected validation pixels exhaustively. Ground truth data are then collected at a limited number of point locations within the selected validation pixels, and the average of the soil variable at these locations is taken as representative of the entire pixel. The design for this sub-sampling is part of what is referred to as the response design by Stehman & Czaplewski (1998). The sub-sampling introduces a sampling error in the collected ground truth data. As we do not know the true mean of the pixel without error, we also do not know the prediction error (the difference between the predicted pixel-mean and the true pixel-mean) without error. Substituting the 'estimated' errors for the true errors in the estimators does not lead to bias in the estimated mean error, but the estimated spatial means of the squared and absolute errors will be biased.

For categorical soil maps on block support, the sub-sample is used to estimate the areal proportions of the soil classes within the selected pixels. These estimated areal proportions are then used to determine the dominant soil class. This sub-sampling may cause bias in the estimated map quality measures (Foody, 2009).

Type of sampling design

Stehman (1999) evaluated the suitability of the five basic types of sampling design for validation: simple random sampling (SI), stratified simple random sampling (STSI), systematic random sampling (SY), cluster random sampling (CL), and two-stage random sampling (TS).

In SI sampling units are selected fully randomly. In STSI, the mapped area is sub-divided into sub-areas referred to as strata, and from each stratum an SI sample is selected. Typically, the map units serve as strata. In SY, a grid of sampling units is randomly placed over the mapped area. Commonly applied grid configurations are square, triangular and hexagonal. In CL, pre-defined sets of sampling units, referred to as clusters, are selected. If one unit of a cluster is selected, then all other units of that cluster are also observed. In general, not all clusters have an equal number of units and they can therefore be of unequal size. In that case, it is convenient to select the clusters with probabilities proportional to their size (pps sampling). A common cluster shape is a transect. In TS, 'clusters' of units are also selected, but in contrast to CL now the selected 'clusters' are sub-sampled and only the units of these sub-samples are observed. The clusters, referred to as primary sampling units (PSU), typically are contiguous sub-areas, for instance the polygons of a map unit. If the PSUs have unequal size, then again these PSUs can best be selected by pps sampling. Note that validation of a soil map on point support by randomly selecting pixels first, followed by random selection of point locations within these pixels, is an example of TS sampling.

Statistical estimation

To estimate the quality measures described earlier, the usual estimators presented in standard textbooks on sampling theory (Cochran, 1977; Lohr, 1999) can be applied. These estimators are based on the inclusion probabilities as determined by the sampling design, and are referred to as the Horvitz–Thompson or π-estimator. This theorem has been extended to point sampling from continuous universes (infinite populations) by Cordy (1993). The π-estimator of the global mean prediction error of a finite population equals:

$$\hat{\bar{e}} = \frac{1}{N}\sum_{i=1}^{n}\frac{e_i}{\pi_i}, \tag{8}$$

where n denotes the sample size (number of selected units), e_i denotes the error of the i^{th} sampling unit and π_i is defined as the inclusion probability of this unit. The π-estimator of the infinite population mean can simply be obtained by replacing A for N in Equation (8), and noting that the inclusion probability now is an inclusion density. As an example, for SI without replacement (of finite populations), the inclusion probability for any unit equals n/N, leading to the un-weighted sample mean $\hat{\bar{e}} = \frac{1}{n}\sum_{i=1}^{n} e_i$ as the π-estimator of the global mean. The sampling variance of the π-estimator can be estimated by (Särndal *et al*., 1992):

$$\hat{V}\left(\hat{\bar{e}}\right) = \frac{1}{N^2}\left\{\sum_{i=1}^{n}(1-\pi_i)\left(\frac{e_i}{\pi_i}\right)^2 + \sum_{i=1}^{n}\sum_{j\neq 1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}\right\}, \tag{9}$$

where π_{ij} denotes the density that both unit i *and* unit j are included in the sample. Inserting $\pi_{ij} = \frac{n}{N}\frac{n-1}{N-1}$ for SI in Equation (9) leads, after some re-arrangement, to the variance estimator for SI without replacement (Brus, 2000):

$$\hat{V}_{\mathrm{SI}}\left(\hat{\bar{e}}\right) = \frac{1-f}{n}s^2(e), \tag{10}$$

where f denotes the sampling fraction, $f = n/N$, and $s^2(e)$ is defined as the estimated spatial variance of the error, $s^2(e) = \frac{\sum_{i=1}^{n}(e_i-\hat{\bar{e}})^2}{n-1}$. A requirement of the variance estimator (Equation (9)) is that $\pi_{ij} > 0$ for all pairs in the area. This is not the case for SY, and this is the reason that for SY no unbiased estimator of the sampling variance exists.

Complications in estimation arise if the surface area of a region for which we want to estimate a spatial mean or a proportion is unknown. An example is estimation of the class representation, which is defined in terms of the surface area where in reality a given soil class occurs. This surface area is unknown. The

spatial mean (areal proportion) can then be estimated by the ratio estimator:

$$\hat{\bar{e}}_{\text{ratio}} = \frac{\sum_{i=1}^{n} \frac{e_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}, \tag{11}$$

where x_i denotes an indicator defined as:

$$x_i = \begin{cases} 1 & \text{if} \quad i \in \mathcal{A}_d \\ 0 & \text{else} \end{cases}, \tag{12}$$

where $\mathcal{A}_d$ denotes the sub-area of interest, hereafter referred to as the domain. The numerator in Equation (11) is the π-estimator of the total error, that is the global mean error multiplied by the surface area of the domain, the denominator is the π-estimator of the surface area of this domain. The ratio estimator is not unbiased but the bias is negligible in samples of moderate size (Cochran, 1977; Lohr, 1999).

The ratio estimator given in Equation (11) is also recommended in situations were the sample size is not fixed but varies between samples selected with the sampling design. For instance, this occurs with estimation of the mean error or purity of a given map unit with simple random sampling. The number of selected sampling units within the map unit is uncontrolled and varies between simple random samples. In this case, we can either estimate the total by the Horvitz–Thompson estimator and divide it by the known surface area, or estimate the mean (proportion) as a ratio. Interestingly, by dividing by the estimated surface area of the map unit instead of its known surface area, the estimator becomes more precise (Särndal *et al*., 1992).

Table 2 shows in which cases the Horvitz–Thompson or ratio estimator is most appropriate.

Use of ancillary information. Collecting sample data for validation is usually expensive, and therefore it is important to make the validation survey as efficient as possible. Efficiency is largely determined by the type of sampling design. However, given the type of sampling design, we may try to increase the efficiency further by exploiting ancillary information on the errors in the soil map in the estimation stage. This can be done by applying a regression estimator, as proposed by Stehman (1996, 1997b) for validation of land-cover maps derived from remote sensing images.

Table 3 Suitability of basic sampling design types for validation (after Stehman, 1999)

	Simplicity design	Simplicity estimation	Precision	Cost	Validity
SI	++	++	−	+/−	+
STSI	+	+	+	+/−	+
SY	+	−	++	+/−	−
CL	−	+/−	−	+	+
TS	+/−	+/−	−	+	+

++, very suitable; +, suitable; +/−, not very suitable; −, unsuitable.

Evaluation of sampling designs

Table 3 shows how we evaluated the five types of sampling designs. The importance of choosing a sampling design that can be implemented easily and that leads to straightforward statistical estimation procedures cannot be over-emphasized. SI, STSI and SY are easy to implement. Estimation of global quality measures such as the global mean (squared, absolute) error, the SCDF of the (squared, absolute) error and the overall purity, is simple and straightforward for these designs. Moreover, for SI and STSI, estimation of the sampling variance is simple. However, for SY no unbiased estimators of the sampling variance exists, and approximate variance estimators are far from simple (D'Orazio, 2003). Estimation of local quality measures such as the mean (squared, absolute) error for map units and map unit purities and their sampling variances is easy for STSI if the map units are used as strata. For SI and SY, as well as for STSI with strata not coinciding with the map units, estimation is complicated because of the random sample sizes within the map units. With random sample sizes the quality measures can best be estimated with a ratio estimator.

Cluster random sampling does not score well on simplicity of design implementation. In practice, clusters are often selected by selecting one sampling unit, and identifying the remaining sampling units of these selected clusters. With this selection procedure, it is essential that the same cluster is selected, regardless of which of its sampling units is selected as a starting unit. For instance, in the case of random sampling of N-S oriented transects, an invalid selection procedure is to select a starting sampling unit and to identify the remaining sampling units such that the position of the starting unit in the transect is fixed, for

Table 2 Estimators of several map quality measures for five basic sampling design types

		STSI				
	SI	Map units as strata	Other strata	SY	CL	TS
Global mean error	HT	HT	HT	HT	HT	HT
Mean error domain	Ratio	HT or ratio[(a)]		Ratio	Ratio	Ratio
Overall purity	HT	HT	HT	HT	HT	HT
Map unit purity	Ratio	HT	Ratio	Ratio	Ratio	Ratio
Class representation	Ratio	Ratio	Ratio	Ratio	Ratio	Ratio

[(a)]If domain of interest equals stratum, then HT estimator, else ratio estimator. HT = Horvitz–Thompson estimator.

example the most southern unit. With this selection procedure, the set of sampling units that will be selected depends on the unit selected as a start. A valid method is to draw a line through the selected unit, and to select all equidistant units on this line, until the transect crosses the boundary of the mapped area. In general, this will lead to transects of different size (numbers of sampling units), some of which can be very large. The size of the clusters can be controlled by sub-dividing the study area into blocks, for instance, stripes perpendicular to the direction of the transects or square blocks where the clusters are grids. In this case, the remaining units are identified by extending the transect or grid to the boundary of the block. With irregularly shaped areas, blocking will not eliminate variation in cluster sizes entirely. With the selection procedure described above, clusters must be selected with pps and with replacement.

With pps sampling, the unequal cluster sizes are accounted for in the selection, so that the estimation of the global quality measures becomes very simple. For instance, the global mean error can be estimated unbiasedly by the un-weighted average of the mean error per cluster, and its sampling variance by the variance of the mean error per cluster divided by the number of selected clusters. If clusters of unequal size are selected with equal probability, quality measures can best be estimated as a ratio, which is slightly more complicated.

In a similar way to CL, in TS, it is convenient to select PSUs with pps and with replacement. Estimation is then very simple. If the sampling fraction of the PSUs is large, then the probability of selecting a PSU more than once becomes considerable, possibly leading to undesirable spatial clustering of secondary sampling units, and as a consequence less precise estimates. In that case pps sampling *without* replacement can be an attractive alternative design, but implementation of this design is complicated (Särndal *et al*., 1992).

The precision of the estimated map quality measures can often be improved by stratifying the validation area. The stronger the correlation (association) between the stratification variable and the error, the larger will be the gain in precision. Also, in general, the precision of estimates can be increased by spreading the validation units throughout the validation area, either by geographical stratification (Walvoort *et al*., 2010), or by SY (Stehman, 1992). CL and TS lead to spatial clustering of validation units. If the errors show strong spatial correlation at short distances, spatial clustering of units may lead to a considerable increase of the sampling variance.

Precision and cost effectiveness generally are in conflict. For instance, increasing precision by spreading the validation units throughout the validation area leads to larger travel distances and travel costs. If travel distances really count, because of a large spatial extent of the validation area or poor accessibility, selecting spatial clusters of validation units as in CL and TS can be advantageous. However, as indicated before, with strong spatial correlation of errors the savings of travel costs are counter-balanced by a decrease of the precision.

The estimates obtained with all sampling design types except SY have acceptable validity properties. This is because the estimates are not based on a model of the spatial variation of the errors. For SY, the un-weighted sample average is an unbiased estimator of the global mean. However, as noted earlier no unbiased estimator of the sampling variance exists for SY.

Required number of sampling units

After the type of sampling design and the estimator have been decided, we must also decide on the sample size. In SI, this is the total number of sampling units to be selected. In STSI, besides the total number, we must decide on the distribution of the sampling units among the strata. For SY, we must decide on the grid-spacing, which determines the expected total number of sampling units. For cluster sampling, the number of clusters and sampling units per cluster and their spatial pattern must be decided. In two-stage random sampling, we must decide on the numbers of primary and secondary sampling units. As an example, we will consider the situation where the sampling is constrained by a requirement on the quality of the estimated map accuracy. This means computing the minimum number of sampling units leading to at least the required quality: we will give estimators for SI only. For the other sampling design types, the required sample size can either be computed by making use of the design-specific sampling variance estimators or by multiplying the required sample size as computed for SI by an empirical factor, the so-called 'design-effect', to account for the greater (STSI, SY) or lesser (TS, CL) accuracy (see de Gruijter *et al*., 2006 for more details).

Spatial mean of error, absolute error or squared error. The simplest case is when a minimum precision is required for the spatial mean of the error, absolute error or squared error in predictions of a quantitative soil property. Suppose that we have a prior estimate of the spatial variance of the (absolute or squared) error, $\breve{s}^2(.)$, then for SI the required number of sampling units can be computed as:

$$n = \frac{\breve{s}^2(.)}{se^2_{\max}}, \tag{13}$$

where $se_{\max}$ denotes the maximum allowable standard error. Alternatively, the quality constraint on the estimated spatial mean of the (absolute or squared) error can be formulated in terms of the probability of occurrence of an absolute or relative error in the estimated spatial mean. In this case, the first step is to compute the maximum allowable standard error of the estimated spatial mean from this quality constraint, and proceed using Equation (13).

Overall purity. In this case, we wish to estimate with given precision the proportion of the area that is correctly classified. This areal proportion is estimated by introducing an indicator variable, as in our case study of Drenthe. For an indicator variable, the variance is related to the mean, and consequently in this case we can

start from a prior estimate of the areal proportion correctly classified, $\breve{p}$, to compute the required number of sampling units, given a requirement on the standard error of the estimated overall purity:

$$n = \frac{\breve{p}(1-\breve{p})}{se^2_{\max}} + 1. \tag{14}$$

The same procedure can be used at the level of map units where this is a requirement for the standard error of the map unit purity.

Case studies

Soil class map of province of Drenthe, Netherlands

Figure 1 shows the categorical soil map of the province of Drenthe (2680 km^2), which was validated. It depicts 10 soil classes. The soil class map was made by multinomial logistic regression (MLR), using explanatory variables derived from the existing 1:50 000 soil map, a digital elevation model and maps with historical land use, recent land use and depth to groundwater table as predictors (Kempen *et al*., 2009). The map units of the existing 1:50 000 soil map were grouped into 10 soil groups, and for each group an MLR model was built. The MLR method results in an estimated probability distribution of soil classes for each 25 m × 25 m pixel on the map. The soil class with the largest probability was used as the predicted soil class. The support of the soil map is a point location. Overall purity, map unit purities and class representations were used as quality measures. These measures were estimated by STSI. The strata were constructed by overlaying a map depicting the 10 soil groups of the existing 1:50 000 soil map, a map of three geographical regions, and a map depicting the areas with an existing soil map at scale of 1:10 000. After generalization, this resulted in 34 strata. This stratification is mainly motivated by differences in expected purity between strata, and related to this, a gain in precision of the estimated quality measures. On the basis of the overall budget, we decided to select 150 validation locations. These were allocated proportionally to the area of the strata, with a minimum of two.

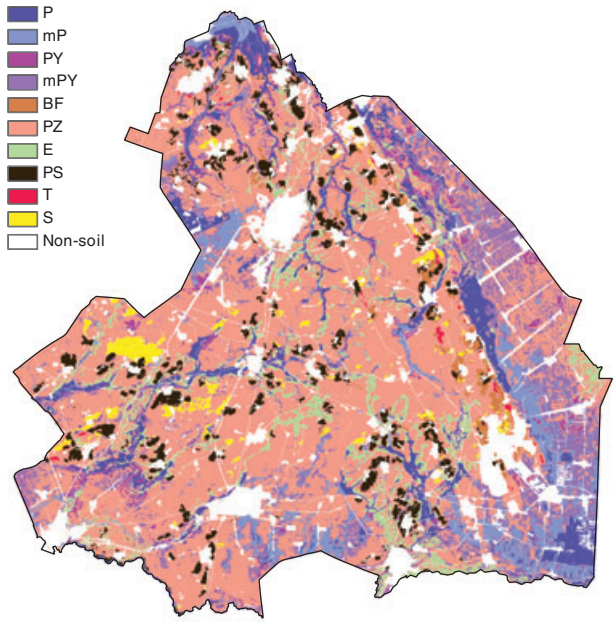

Figure 1 Validated categorical soil map of Drenthe. P = thick peat soils; mP = thick peat soils with mineral surface horizon; PY = thin peat soils; mPY = thin peat soils with mineral surface horizon; BF = brown forest soils; PZ = podzol soils; E = earth soils; PS = plaggen soils; T = till soils; S = sandy vague soils.

Overall purity. For stratified simple random sampling, the overall purity can be estimated as the weighted average of the overall accuracies per stratum $\hat{p}_h$:

$$\hat{p} = \sum_{h=1}^{H} w_h \hat{p}_h = \sum_{h=1}^{H} w_h \frac{\sum_{u=1}^{U} n_{huu}}{n_h}, \tag{15}$$

where w_h denotes the relative size (area) of stratum h, $w_h = A_h/A$, n_{huu} denotes the number of sampling locations in stratum h correctly mapped as u (the counts in cell (u,u) of sample error matrix for stratum h, Table 4), and n_h denotes the total number of sampling locations in stratum h. Its standard error was estimated by:

$$se(\hat{p}) = \sqrt{\sum_{h=1}^{H} w_h^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}}. \tag{16}$$

Map unit purities. As explained above, 10 map units of an existing 1:50 000 polygon soil map were used as strata, not the map units of the final soil map for which we want to estimate the purity. This implies that the number of sampling locations per map unit of the soil map is not controlled, and is random and thus varies between repeated samples. For that reason, we estimated the map unit purity by the ratio estimator:

$$\hat{p}_u = \frac{\sum_{h=1}^{H} A_h \frac{n_{huu}}{n_h}}{\sum_{h=1}^{H} A_h \frac{n_{hu+}}{n_h}}, \tag{17}$$

Table 4 Sample error matrix for a stratum

		Field				
		1	2	⋯	U	$\sum$
Map	1	n_{h11}	n_{h12}	⋯	n_{h1U}	n_{h1+}
	2	n_{h21}	n_{h22}	⋯	n_{h2U}	n_{h2+}
	.	.	.	⋯	.	.
	.	.	.	⋯	.	.
	.	.	.	⋯	.	.
	U	n_{hU1}	n_{hU2}	⋯	n_{hUU}	n_{hU+}
	$\sum$	n_{h+1}	n_{h+2}	⋯	n_{h+U}	n_h

n_{hij} = Number of sampling locations in stratum h mapped as soil class c_i with observed soil class c_j.

where A_h denotes the surface area of stratum h, and n_{hu+} denotes the number of sampling locations in stratum h mapped as u (counts in the row marginal for map unit u of the sample error matrix of stratum h, Table 4). The standard error of this estimate was estimated by:

$$se(\hat{p}_u) = \sqrt{\frac{1}{(A_{u+})^2}\sum_{h=1}^{H} A_h^2 \frac{s_h^2(d)}{n_h}}, \tag{18}$$

where A_{u+} denotes the surface area of map unit u, and $s_h^2(d)$ denotes the estimated residual variance in stratum h:

$$s_h^2(d) = \frac{\sum_{i=1}^{n_h}\left(d_{hui} - \overline{d}_{hu}\right)^2}{n_h - 1}, \tag{19}$$

where

$$d_{hui} = y_{hui} - \hat{p}_u \cdot x_{hui} \tag{20}$$

and $\overline{d}_{hu} = \frac{1}{n_h}\sum_{i=1}^{n_h} d_{hui}$. In Equation (20) y_{hui} and x_{hui} are indicators defined as:

$$y_{hui} = \begin{cases} 1 & \text{if} \quad \hat{c}_{hi} = c_{hi} = u \\ 0 & \text{else} \end{cases}, \tag{21}$$

($\hat{c}_{hi}$ and c_{hi} are the predicted and observed soil class at sampling location i in stratum h, respectively), and:

$$x_{hui} = \begin{cases} 1 & \text{if} \quad \hat{c}_{hi} = u \\ 0 & \text{else} \end{cases}. \tag{22}$$

Class representations. The class representations were estimated by the ratio estimator:

$$\hat{r}_u = \frac{\sum_{h=1}^{H} A_h \frac{n_{huu}}{n_h}}{\sum_{h=1}^{H} A_h \frac{n_{h+u}}{n_h}}, \tag{23}$$

where n_{h+u} denotes the counts in column marginal u of the sample error matrix for stratum h. The standard error of this estimate was estimated by (Cochran, 1977):

$$se(\hat{r}_u) = \sqrt{\left(\frac{1}{\hat{A}_{+u}}\right)^2 \sum_{h=1}^{H} A_h^2 \frac{s_h^2(\delta)}{n_h}}, \tag{24}$$

where $\hat{A}_{+u} = \sum_{h=1}^{H} A_h \frac{n_{h+u}}{n_h}$ denotes the estimated surface area of the domain, and $s_h^2(\delta)$ denotes the estimated residual variance in stratum h:

$$s_h^2(\delta) = \frac{\sum_{i=1}^{n_h}\left(\delta_{hui} - \overline{\delta}_{hu}\right)^2}{n_h - 1}, \tag{25}$$

where $\delta_{hui} = y_{hui} - \hat{r}_u \cdot z_{hui}$ and $\overline{\delta}_{hu} = \frac{1}{n_h}\sum_{i=1}^{n_h}\delta_{hui}$, with z_{hui} an indicator defined as:

$$z_{hui} = \begin{cases} 1 & \text{if} \quad c_{hi} = u \\ 0 & \text{else} \end{cases}. \tag{26}$$

Results. The estimated overall purity of the soil map of Drenthe was 58%, with a standard error of 4%. The overall purity increased only slightly (6%) as compared with the existing soil map. The map unit purities for soil classes P, BF, PZ, PS and S were the largest, all larger than 67%, soil classes mP, mPY, E had map unit purities smaller than 50%, while those of PY and T were close to 0 (Table 5). The class representation of PZ was very large (90%), showing that almost all actual podzols in Drenthe were depicted on the map. However, only two thirds of the surface area depicted as podzols on the map are podzols in reality, as shown by the map unit purity for this map unit. Five soil classes (P, mP, BF, PS and S) have class representations around 50%, while the remaining soil classes (mPY, E) are very poorly represented on the map. Of special interest are the accuracies of the peat soils (P, mP, PY, mPY) as these soil types strongly contribute to the total carbon stocks in the soil of an area. Only about 50% of the thick peat soils (P, mP) were detected by the MLR model, and for the thin peat soils (PY, mPY) the class representations were even smaller. However, if we aggregate the four peat soil classes, then the map unit purity and class representation increases to 80 and 72%, respectively. This is substantially larger than the pooled map unit purities and class representations for the four soil classes separately. Apparently, the MLR model has problems in distinguishing between thick and thin peat soils, and whether an anthropogenic sandy surface horizon is present or not. The large standard errors for the map unit purities and class representations (Table 5), which result from to the few validation locations per soil class.

Interestingly, Kempen *et al*. (2009) reported a larger theoretical overall purity as computed with the model (internal accuracy)

Table 5 Estimated map unit purities and class representation for 10 soil classes as depicted on soil map of Drenthe (Netherlands)

	Map unit purity	Class representation
P	77 (14)	51 (12)
mP	28 (11)	55 (23)
PY	4.6 (9.1)	5.7 (5.7)
mPY	37 (11)	24 (9.6)
BF	71 (17)	40 (28)
PZ	67 (5.2)	90 (3.4)
E	46 (14)	21 (7.2)
PS	80 (16)	47 (14)
T	0.0 (0.0)	—
S	94 (4.1)	52 (22)

Numbers in parentheses = standard error; – = not estimated.

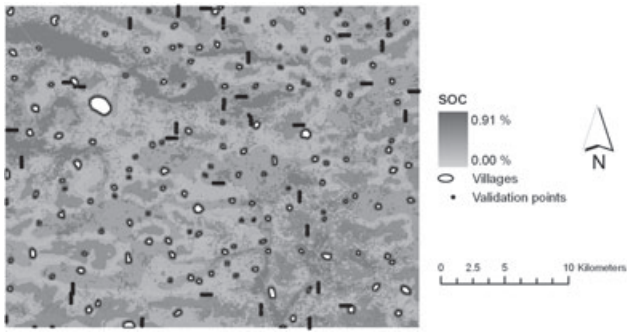

Figure 2 Validated quantitative soil map of the Nioro study area, Senegal, depicting soil organic carbon content of the topsoil (0–20 cm) (Stoorvogel *et al*., 2009).

than the empirical overall purity as estimated from the probability sample (67 vs. 58% overall purity).

Soil map of organic carbon content of the Nioro study area, Senegal

Figure 2 shows the quantitative soil map of the SOC content in the top soil (0–20 cm) which was validated. The soil map was constructed by a classification tree model (Stoorvogel *et al*., 2009). The support of the soil map was a pixel of 30 m × 30 m. The soil map was validated by cluster random sampling. Clusters were transects of five pixels of 30 m × 30 m, North-South or East-West oriented. Because of the available time for fieldwork, it was decided to select 32 clusters. These clusters were selected by simple random sampling without replacement from a map showing all clusters in the population. The average carbon content of every selected pixel was estimated by taking a centred composite sample of five aliquots (Figure 3).

SCDF of the error, the absolute error and the squared error. As quality measures we used the SCDFs of the error, absolute error and squared error, and the SCDF parameters of mean, median, 90th percentile and standard deviation. These SCDFs and parameters were estimated for the mapped areas as a whole, and for two domains defined in terms of land-use (agriculture, uncultivated).

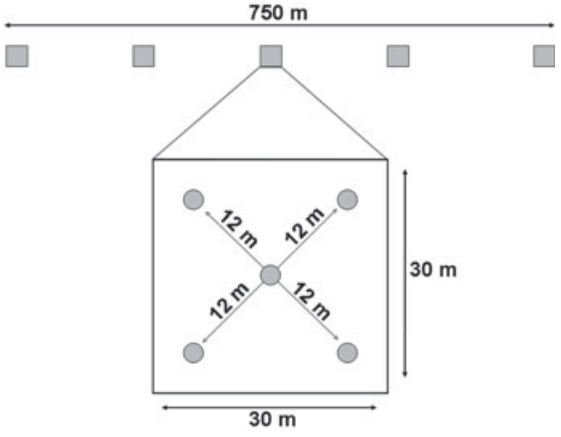

Figure 3 Cluster of pixels and sampling pattern used for estimating mean soil organic carbon content of a pixel (Stoorvogel *et al*., 2009).

The SCDF was obtained by defining a series of indicators, one for each observed error in the sample:

$$y_{t,ij} = \begin{cases} 1 & \text{if} \quad e_{ij} \leq t \\ 0 & \text{else} \end{cases}, \tag{27}$$

where e_{ij} denotes the error for sampling unit (pixel) j in cluster i, and t denotes a threshold error. So, for the threshold error we take one by one the different error values observed in the sample. The SCDF can then be estimated as the spatial means of these indicators:

$$\hat{F}(t) = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{n_i} y_{t,ij}, \tag{28}$$

where C denotes the number of clusters, and n_i denotes the number of sampling units (pixels) in cluster i. The sum at the right-hand side is the number of sampling units in cluster i with an error equal or smaller than a threshold error t. Note that in this case n_i is constant, so that Equation (28) can be simplified to

$$\hat{F}(t) = \frac{1}{Cn}\sum_{i=1}^{C}\sum_{j=1}^{n} y_{t,ij}. \tag{29}$$

The estimator Equation (28) is more general and is appropriate for pps sampling of clusters of unequal size (n_i not constant).

SCDF for a domain. The SCDF for the two domains 'agriculture' and 'uncultivated' were estimated by the ratio estimator:

$$\hat{F}_d(t) = \frac{\sum_{i=1}^{C}\sum_{j=1}^{n_i}\acute{y}_{t,ij}}{\sum_{i=1}^{C}\sum_{j=1}^{n_i} x_{ij}}, \tag{30}$$

with

$$\acute{y}_{t,ij} = \begin{cases} y_{t,ij} & \text{if} \quad ij \in \mathcal{A}_d \\ 0 & \text{else} \end{cases}, \tag{31}$$

and

$$x_{ij} = \begin{cases} 1 & \text{if} \quad ij \in \mathcal{A}_d \\ 0 & \text{else} \end{cases}, \tag{32}$$

Note that $\sum_{j=1}^{n_i} x_{ij}$ equals the number of pixels in cluster i falling in the domain of interest.

Percentiles of the error, absolute error and squared error. Percentiles can be estimated by inverse use of the estimated SCDF. As the SCDF is a stepwise function, this is not straightforward and several methods exist for computing the percentile from the estimated SCDF. We estimated percentiles by linear interpolation of the estimated cumulative frequencies.

Mean error, mean absolute error and mean squared error. The global mean (absolute or squared) error can be estimated by replacing $y_{t,ij}$ in Equation (28) by the (absolute or squared) error e_{ij}. The standard error of this estimate was estimated by

$$se(\hat{\overline{e}}) = \sqrt{\frac{1}{C(C-1)}\sum_{i=1}^{C}(\hat{\overline{e}}_i - \hat{\overline{e}})^2}, \tag{33}$$

where $\hat{\overline{e}}_i$ denotes the sample mean of cluster i. Similarly, the mean (absolute or squared) error in domains can be estimated by replacing $\acute{y}_{t,ij}$ in Equation (30) by $\acute{e}_{ij}$ which has value e_{ij} if sampling unit j of cluster i is in the domain of interest, and 0 else. The standard error of this ratio estimate of the domain mean can be estimated by:

$$se(\hat{\overline{e}}_d) = \sqrt{\frac{1}{C}\left(\frac{1}{\overline{n}_i}\right)^2 \frac{\sum_{i=1}^{C}\left(\delta_i - \overline{\delta}\right)^2}{C-1}}, \tag{34}$$

with $\overline{n}_i = \frac{1}{C}\sum_{j=1}^{n_i} x_{ij}$, $\delta_i = (\sum_{j=1}^{n_i} \acute{e}_{ij}) - \hat{\overline{e}}_d \cdot n_i$, and $\overline{\delta} = \frac{1}{C}\sum_{i=1}^{C}\delta_i$.

Results. Figure 4 shows the estimated SCDFs of the error, absolute error and squared error. These graphs are much more informative than a few parameters of the SCDF only. Any percentile

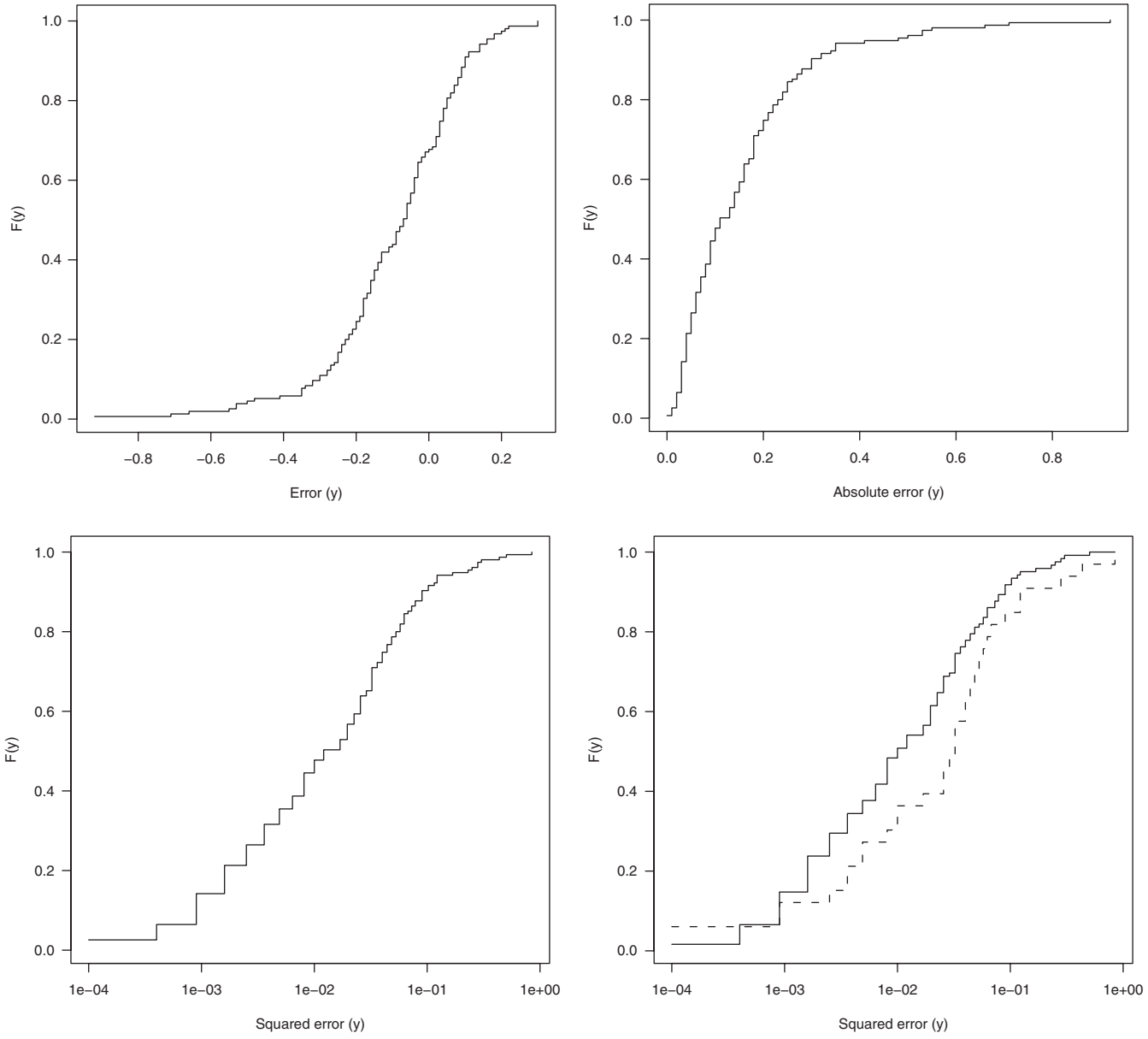

Figure 4 Estimated SCDF of the error, absolute error and squared error of SOC predictions for the Nioro study area (Senegal). Lower right: SCDF of squared error for two land use units. Solid line = agricultural soils; dashed line = uncultivated soils.

Table 6 Estimated means (standard error), medians and 90th percentile of prediction error, absolute prediction error, and squared prediction error, for the Nioro study area (Senegal) as a whole, and for two land use units

	Mean	Median	P90
	Nioro		
Error	−0.096 (0.019)	−0.072	0.096
Absolute error	0.155 (0.014)	0.109	0.298
Squared error	0.044 (0.009)	0.012	0.089
	Agriculture		
Error	−0.074 (0.018)	−0.057	0.104
Absolute error	0.141 (0.013)	0.097	0.285
Squared error	0.035 (0.007)	0.009	0.081
	Uncultivated		
Error	−0.176 (0.047)	−0.175	0.067
Absolute error	0.205 (0.041)	0.172	0.349
Squared error	0.077 (0.034)	0.029	0.121

can be read from the graph visually. The SCDF of the error shows that the error has strong negative skew because there are several strongly negative outliers. The SCDFs of the squared error showed strong positive skew. By taking the logs of the squared error the SCDFs become symmetric (note the log-transformed horizontal axes in the lower two graphs of Figure 4). The SCDF of the squared error in uncultivated soils is less than that of agricultural soils, indicating that the percentiles of the squared error in uncultivated soils will be larger than the corresponding percentiles in agricultural soils. With an estimated global mean SOC of 0.54%, the estimated global mean error of about −0.1% shows that the SOC predictions are seriously biased (Table 6). The bias in uncultivated soils is much larger than in agricultural soils. The spatial standard deviation of the prediction error of 0.188 (obtained by subtracting the squared bias from the mean squared error, and then taking the square root) shows that also the random error is considerable. For the agricultural and uncultivated soils this standard deviation equals 0.184 and 0.217, respectively. On the basis of the validation, Stoorvogel *et al*., (2009) concluded that the classification tree model performed poorly. The classification tree model did not provide a theoretical estimate of the classification error.

Discussion

Ancillary information on the errors in soil maps can be either used in the design of a sample or in estimating the quality measures. For instance, for soil maps obtained by spatial interpolation, the error generally increases with distance to the observation locations. This prior knowledge can be used at the design stage by stratifying the area according to this distance. Brus *et al*. (1996) defined two strata: close to the nearest observation point (<100 m), and further away. The gain in precision compared to simple random sampling was small, but could be increased by optimal allocation of the validation locations to the strata by choosing the stratum sample sizes such that the sampling variance of the estimated mean squared error is minimized. This leads to larger sampling densities in strata with large within-stratum spatial variance of the prediction error. With kriging, the variance can be used to stratify the area and choose the optimal stratum sample sizes. Ancillary information on the error can also be utilized at the estimation stage, for instance by using a post-stratification estimator or regression estimator (Stehman, 1996a). In fact, using ancillary information at the estimation stage is more flexible than at the design stage.

If no budget is available for validation by an additional probability sample, we recommend leave-one-out cross-validation of the soil map, which is obviously to be preferred to leaving the map unvalidated. In practice, the un-weighted mean of the prediction (classification) cross-validation errors at the calibration locations is often taken as an estimate of the spatial mean error. It leads to optimistic estimates of overall purity when easy-to-classify areas are over-represented in the sample, and to pessimistic estimates when difficult-to-classify areas are over-represented. Moreover, the prediction (classification) errors can be spatially auto-correlated. Especially, when validation units show strong spatial clustering it is important to account for spatial auto-correlation in order to reduce the weight of these clusters on the estimated mean error. This can be done by estimating a variogram of the errors, and using it in block-kriging the spatial mean error. A heuristic alternative is to employ spatial declustering, and to use the declustering weights to compute a weighted average of the cross-validation error. Note, however, that accounting for spatial auto-correlation does not guarantee elimination of a possible sampling bias.

Steele *et al*. (2003) describe an alternative method for assessment of the accuracy of land-cover maps when budget or time for validation by additional probability sampling is lacking. Their method estimates the probability of correct classification using the classified polygons rather than the training observations. They argue that their method is superior to re-sampling methods (cross-validation or bootstrapping) when training data are spatially clustered. However, the need for a probability sample is not entirely eliminated, as the estimated probabilities still must be calibrated.

In this paper, we assumed that errors in the observations of the soil attribute or soil class are negligibly small, justifying use of the term 'ground truth'. In practice, observations on the soil will always contain errors to some degree. In the case of validating soil maps on block support, there will also be a sub-sampling error in the estimated block-means of quantitative soil properties or in the estimated dominant soil type within the blocks. Observation and sub-sampling errors affect the quality of the estimated map quality measures. If the observation and sub-sampling methods are unbiased, the estimated mean of the prediction errors $\hat{\bar{e}}$ will also be unbiased. However, a random observation or sub-sampling error in a quantitative soil attribute inflates the spatial variance of the prediction error, as well as the mean of the squared prediction errors, $\hat{\overline{e^2}}$. For SI, a corrected estimate of the variance of the prediction error and mean squared error can simply be obtained by subtracting the observation error variance or sub-sampling error variance from the variance of the prediction error. More research is needed on how the entire SCDF of the error can be corrected for

observation and sub-sampling error. Furthermore, more studies are needed on methods to reduce the negative impacts of ground-truth error on the quality of the estimated confusion matrix and derived quality measures of categorical maps (van Oort, 2005; Foody, 2009). This is particularly relevant for soil science, because soil classification is not a trivial exercise and is not guaranteed to give a unique answer, even when several soil classification experts are gathered around the same soil pit.

Conclusions

We conclude that for thematic soil maps depicting a quantitative soil property the entire SCDF of the (squared, absolute) error is more informative than one or several parameters of the SCDF, such as the mean (squared) error. For categorical soil maps, the overall purity of the map, the map unit purities and the soil class representations are the most interesting quality measures. Further, we conclude that, when the calibration data are a non-probability sample, then validation by the additional probability sampling is superior to the validation by data-splitting or cross-validation using the calibration data, because unbiased and valid estimates of the quality measures and their associated estimation errors can then be obtained. If no budget or time is available for additional probability sampling, the best option is leave-one-out cross-validation. If the calibration data show strong spatial clustering, then we recommend accounting for spatial auto-correlation of the errors, for instance by predicting the mean error by block-kriging. However, this does not guarantee elimination of possible sampling bias.

When choosing between design types, stratified simple random sampling generally is a good option. Estimation of the map unit purities is straightforward when the soil map units are used as strata. Prior knowledge about the error can be used to define the strata, which may increase the precision of the estimated map quality measures. When the costs of travel to the validation units form a large part of the total costs, selecting spatial clusters of validation units as in CL and TS can be efficient. However, there are some pitfalls in the implementation of CL, which makes this design type less suitable for soil surveyors with little experience in probability sampling.

Digital soil maps in raster format have either point support or block support. In sampling for validation, the basic sampling units should have the same support as the digital soil map. Soil maps on point support must be validated with point observations, and soil maps on block support must be validated with observations or estimates on block support. In the case of maps on point support, all points within a pixel should have positive inclusion density, and not just the centre point.

Acknowledgements

This research was co-funded by the Research DG of the European Commission within the RTD activities of the FP7 Thematic Priority Environment, under the Collaborative Projects iSOIL - Interactions between soil related sciences - Linking geophysics, soil science and digital soil mapping (Grant Agreement number 211386), e-SOTER (Grant Agreement number 211578), and the strategic research program 'Sustainable spatial development of ecosystems, landscapes, seas and regions' of the Dutch Ministry of Agriculture, Nature Conservation and Food Quality (BAS-number KB-01-001-018-ALT).

References

Beckett, P.H.T. & Bie, S.W. 1978. *Use of Soil and Land-system Maps to Provide Soil Information in Australia*. Division of Soils Technical Paper, Volume 33, CSIRO, Collingwood, Victoria, Australia.

Bie, S.W. & Beckett, P.H.T. 1971. Quality control in soil survey. Introduction: 1. the choice of a mapping unit. *Journal of Soil Science*, **22**, 32–49.

Brus, D.J. 2000. Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science*, **51**, 159–172.

Brus, D.J. & de Gruijter, J.J. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma*, **80**, 1–59.

Brus, D.J. & Heuvelink, G.B.M. 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, **138**, 86–95.

Brus, D.J., Bogaert, P. & Heuvelink, G.B.M. 2008. Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. *European Journal of Soil Science*, **59**, 166–177.

Brus, D.J., de Gruijter, J.J., Marsman, B.A., Visschers, R., Bregt, A.K., Breeuwsma, A. *et al*. 1996. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. *Environ metrics*, **7**, 1–16.

Bui, E.N. & Moran, C.J. 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling Basin of Australia. *Geoderma*, **111**, 21–44.

Burrough, P.A., Beckett, P.H.T. & Jarvis, M.G. 1971. The relation between cost and utility in soil survey. (I-III). *Journal of Soil Science*, **22**, 368–394.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, **158**, 419–466.

Cochran, W.G. 1977. *Sampling Techniques*. Wiley, New York.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, 35–46.

Cordy, C.B. 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, **18**, 353–362.

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

de Gruijter, J.J. & Marsman, B.A. 1985. Transect sampling for reliable information on mapping units. In: *Proceedings of the SSSA-ISSS Workshop on Spatial Variability, Las Vegas, U.S.A., Nov. 30 - Dec. 1, 1984*. (eds. D.R. Nielson & J. Bouma), pp. 150–163. Pudoc, Wageningen.

de Gruijter, J.J. & ter Braak, C.J.F. 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, **22**, 407–415.

D'Orazio, M. 2003. Estimating the variance of the sample mean in two-dimensional systematic sampling. *Journal of Agricultural, Biological & Environmental Statistics*, **8**, 280–295.

Efron, B. & Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, London.

Foody, G.M. 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, 185–201.

Foody, G.M. 2009. The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, **30**, 3275–3281.

Friedl, M.A., Woodcock, C., Gopal, S., Muchoney, D., Strahler, A.H. & Barker-Schaaf, C. 2000. A note on procedures used for accuracy assessment in land cover maps derived from avhrr data. *International Journal of Remote Sensing*, **21**, 1073–1077.

Grinand, C., Arrouays, D., Laroche, B. & Martin, M.P. 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, **143**, 180–190.

Grunwald, S. 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, **152**, 195–207.

Kempen, B., Brus, D.J., Heuvelink, G.B.M. & Stoorvogel, J. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. *Geoderma*, **151**, 311–326.

Krzanowski, W.J. 2001. Data-based interval estimation of classification error rates. *Journal of Applied Statistics*, **5**, 585–595.

Lark, R.M. 1995. Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. *International Journal of Remote Sensing*, **16**, 1461–1480.

Lewis, H.G. & Brown, M. 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, **22**, 3223–3235.

Lohr, S.L. 1999. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, USA.

Marsman, B.A. & de Gruijter, J.J. 1986. *Quality of Soil Maps: A Comparison of Survey Methods in a Sandy Area*. Soil Survey Papers 15, Soil Survey Institute, Wageningen.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J. & Heuvelink, G.B.M. 2008. Small scale digital soil mapping in Southeastern Kenya. *Catena*, **76**, 44–53.

Muchoney, D. & Strahler, A. H. 2002. Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, **81**, 290–299.

Mueller, T.G., Pusuluri, N.B., Mathias, K.K., Cornelius, P.L. & Barnhisel, R.I. 2004. Site-specific soil fertility management: a model for map quality. *Soil Science Society of America Journal*, **68**, 2031–2041.

Pontius, R.G. & Cheuk, M. 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, **20**, 1–30.

Särndal, C-E., Swensson, B. & Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer Verlag, New York.

Steele, B.M., Patterson, D.A. & Redmond, R.L. 2003. Toward estimation of map accuracy without a probability test sample. *Environmental and Ecological Statistics*, **10**, 333–356.

Steers, C.A. & Hajek, B.F. 1979. Determination of map unit composition by a random selection of transects. *Soil Science Society of America Journal*, **43**, 156–160.

Stehman, S.V. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, **58**, 1343–1350.

Stehman, S.V. 1996. Use of auxiliary data to improve the precision of estimators of thematic map accuracy. *Remote Sensing of Environment*, **58**, 169–176.

Stehman, S.V. 1997a. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, 77–89.

Stehman, S.V. 1997b. Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, **60**, 258–269.

Stehman, S.V. 1999. Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, **20**, 2423–2441.

Stehman, S.V. 2000. Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, **72**, 35–45.

Stehman, S.V. & Czaplewski, R.L. 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, **64**, 331–344.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M. & de Bruin, S. 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, **149**, 161–170.

Taylor, J.A., Coulouma, G., Lagacherie, P. & Tisseyre, B. 2009. Mapping soil units within a vineyard using statistics associated with high-resolution apparent soil electrical conductivity data and factorial discriminant analysis. *Geoderma*, **153**, 278–284.

Thompson, J.A. & Kolka, R. K. 2005. Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. *Soil Science Society of America Journal*, **69**, 1086–1093.

van Kuilenburg, J., de Gruijter, J.J., Marsman, B.A. & Bouma, J. 1982. Accuracy of spatial interpolation between point data on soil moisture capacity, compared with estimates from mapping units. *Geoderma*, **27**, 311–325.

van Oort, P.A.J. 2005. Improving land cover change estimates by accounting for classification errors. *International Journal of Remote Sensing*, **26**, 3009–3024.

Walvoort, D.J.J., Brus, D.J. & de Gruijter, J.J. 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*, **36**, 1261–1267.

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

Webster, R. & Oliver, M. A. 2007. *Geostatistics for Environmental Scientists*, 2nd edn. Wiley, Chichester.

White, E.M. 1966. Validity of the transect method for estimating compositions of soil-map areas. *Soil Science Society of America Proceedings*, **30**, 129–130.

Woodcock, C.E. & Gopal, S. 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, **14**, 153–172.