5.3 Validation

Bas Kempen, Dick J. Brus, Gerard B.M. Heuvelink

5.3.1 What is validation?

No map is perfect. All maps, including soil maps, are representations of reality that are often based on an underlying model. This means that there will always be a deviation between the phenomenon depicted on the map and the phenomenon observed in the real world, i.e. each map will contain errors. The magnitude of the errors determine the quality of the map. If a map matches reality well (the error is small), the quality or accuracy of the map is high. On the other hand, if a map does not match reality well, map accuracy is low.

Soil maps are used for many purposes. For example to report on (changes in) soil organic carbon stocks, as input in agro-environmental models, to determine land use suitability or for decision- and policy-making. It is therefore, important that the quality of a map is determined and quantified. This is achieved through (statistical) validation.

Validation is defined here as an activity in which the soil map predictions are compared with observed values. From this comparison, the map quality can be quantified and summarized using map quality measures. These measures indicate how accurate the map is on average for the mapping area, i.e. what is the expected error at a randomly selected location in the mapping area. This means that map quality measures obtained through validation are global measures: each quality measure gives one value for the entire map. Note that this is different from results obtained through uncertainty assessment. Such assessment provides local, location-specific (i.e. for each individual grid cell) estimates of map quality as we saw in the previous sections. Another important difference between validation and uncertainty assessment is that validation can be done using a model-free approach. We saw in section 5.2 that uncertainty assessment takes a model-based approach by defining a geostatistical model of the soil property of interest and deriving an interpolated map and the associated uncertainty from that, or by constructing a geostatistical model of the error in an existing map. The approach yields a complete probabilistic characterisation of the map uncertainty, but such characterisation is only valid under the assumptions made. For instance the stationarity assumptions required for kriging. Validation, when done properly as explained hereafter, does not assume a geostatistical model of the error, and hence is model- or assumption-free. This is an important property of validation since we do not want to question the objectivity and validity of the validation results.

We distinguish internal and external map accuracy. Statistical methods typically produce direct estimates of map quality, for instance the kriging variance or the coefficient of determination (R^2) of a linear regression model. These we refer to as internal accuracy measures since these rely on model assumptions and are computed from data that are used for model calibration. Preferably, validation is done with an independent dataset not used in map making. Using such dataset gives the external map accuracy. One will often see that the external accuracy is poorer than the internal accuracy.

In section 5.3.2 we will present the most common accuracy measures used to quantify map quality of quantitative (continuous) soil maps and qualitative (categorical) soil maps. In section 5.3.3 we will introduce three commonly used validation methods and show how to estimate the map quality measures from a sample. This chapter is largely based on Brus et al. (2011), for details we refer to this paper.

5.3.2 Map quality measures

Quality measures for quantitative soil maps

All map quality measures considered here are computed from the *prediction error*. For quantitative soil maps of continuous soil properties (e.g. organic carbon content, pH, clay content) the prediction error is defined as the difference between the predicted value at a location and the true value at that location (which is the value that would be observed or measured by a preferably errorless measurement instrument) (Brus et al., 2011):

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

where $\hat{z}(\mathbf{s})$ is the predicted soil property at validation location $\mathbf{s}$, and $z(\mathbf{s})$ is the true value of the soil property at that location. We consider six map quality measures that are computed from the prediction error here: the mean error, the mean absolute error, the mean squared error and root mean squared error, the model efficiency and the mean squared deviation ratio.

Before we introduce the map quality measures and show how to estimate these, it is important to understand the difference between the *population* and a *sample* taken from the population. The population is the set of all locations in a mapping area. For digital soil maps, this is the set of all pixels or grid cells of a map. A sample is a subset of locations, selected in some way from the set of all locations in the mapping area. With validation we want to assess the map accuracy for the entire population, i.e. for the map as a whole; we are not interested in the accuracy at the sample of locations only. For instance, we would like to know the prediction error averaged over all locations of a map and not merely the average prediction error at a sample of locations. Map quality measures are therefore, defined as population means. Because we cannot afford to determine the prediction error at each location (grid cell) of the mapping area to calculate the population means, we have to take a sample of a limited number locations in the mapping area. This sample is then used to *estimate* the population means. It is important to realize that we are uncertain about the population means, because we estimate it from a sample. Ideally this uncertainty is quantified and reported together with the estimated map quality measures.

In this section we will introduce the definitions of the map quality measures. In the next section we show how we can estimate these measures from a sample.

Mean error

The mean error (ME) measures bias in the predictions. The ME is defined as the population mean (spatial mean) of the prediction errors:

$$\mathrm{ME} = \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e(\mathbf{s}_i)$$

where $\boldsymbol{i}$ indicates the location, $i = 1,2,\ldots,N$, and N is the total number of locations or grid cells/pixels in the mapping area. The mean error should be (close to) zero, which means that predictions are unbiased meaning that there is no systematic over- or under-prediction of the soil property of interest.

Mean absolute error and (root) mean squared error

The mean absolute error (MAE) and mean squared error (MSE) are measures of map accuracy and indicate the magnitude of error we make on average. The MAE is defined by the population mean of the absolute errors:

$$\mathrm{MAE} = |\bar{e}| = \frac{1}{N}\sum_{i=1}^{N} |e|(\mathbf{s}_i)$$

and the MSE by the population mean of the squared errors:

$$\mathrm{MSE} = \overline{e^2} = \frac{1}{N}\sum_{i=1}^{N} e^2(\mathbf{s}_i)$$

Many authors report the root mean squared error (RMSE) instead of the MSE, which is computed by taking the square root of the MSE. The RMSE can be a more appealing quality measure since it has the same unit of measurement as the mapped property and can therefore more easily be compared to it. If the squared error distribution is strongly skewed, for instance when several very large errors are present, then this can severely inflate the (R)MSE. In such case, the (root) median squared error is a more robust statistic for the 'average' error (Kempen et al., 2012).

Brus et al. (2011) argue that instead of using a single summary statistic (the mean) to quantify map quality measures, one should preferably express quality measures for quantitative soil maps through cumulative distribution functions (CDFs). Such functions provide a full descriptions of the quality measures from which various parameters can be reported, such as the mean, median or percentiles. Furthermore, they argue that it can be of interest to define CDFs or its parameters for sub-areas, for

instance geomorphic units, soil or land cover classes. Brus et al. (2011) give examples of estimating CDFs for validation of digital soil maps.

Amount of variance explained

The model efficiency, or Amount of Variance Explained (AVE) (Angelini et al., 2016; Samuel-Rosa et al., 2015), quantifies the fraction of the variation in the data that is explained by the prediction model. It measures the improvement of the model prediction over using the mean of the data set as predictor and is defined as follows (Krause et al., 2005):

$$AVE = 1 - \frac{\sum_{i=1}^{N}\left(\hat{z}(\mathbf{s_i}) - z(\mathbf{s_i})\right)^2}{\sum_{i=1}^{N}(z(\mathbf{s_i}) - \bar{z})^2}$$

where $\bar{z}$ is the population mean of soil property z. The quantity in the numerator is the sum of the squared prediction errors (for each location the prediction error is computed and squared; the squared prediction errors are summed over all locations in the area). In linear regression this quantity is known as the *residual sum of squares* (RSS). The quantity in the denominator is also a sum of squared prediction errors, but here the mean of the area is used as predictor. In linear regression this quantity is known as the *total sum of squares* (TSS). Note that if we would divide the quantity in the denominator by the number of locations in the mapping area *N* we would obtain the population variance (spatial variance) of the soil property z.

If the numerator and denominator are equal, meaning the AVE is zero, then the model predictions are no improvement over using the mean of the data set as predictor for any location in the mapping area. An AVE value larger than zero (RSS smaller than TSS) means that the model predictions area an improvement over using the mean as predictor (this is what we hope for). In case the AVE is negative, then the mean of the data set is a better predictor than the prediction model.

Mean squared deviation ratio

Finally, we introduce the mean squared deviation ratio (MSDR) as a map quality measure (Kempen et al., 2010; Lark, 2000; Voltz and Webster, 1990; Webster and Oliver, 2007). Contrary to the quality measures discussed so far, the MSDR assesses how well the prediction model estimates the prediction uncertainty (expressed as the prediction error variance). The MSDR is defined as:

$$MSDR = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\hat{z}(\mathbf{s_i}) - z(\mathbf{s_i})\right)^2}{\sigma^2(\mathbf{s_i})}$$

where $\sigma^2(\mathbf{s_i})$ is the prediction error variance at location $\mathbf{s_i}$, $i = 1,2,\ldots,N$. The numerator is the squared error at location $\mathbf{s_i}$. The fraction represents the squared Z_{score}. In case of kriging, the prediction error variance is the kriging variance. In case of linear regression, the prediction error variance is the prediction variance of the linear regression predictions that can be obtained by the statistical software `R` by running the `predict` function with argument `se.fit=TRUE`. This function returns for each prediction location the standard error of the predicted value as well as the residual standard deviation (the `residual.scale` value). By squaring both values and then summing these, the prediction error variance is obtained. If the prediction model estimates the error variance well, then the MSDR should be close to one. A value smaller than one suggests that the prediction error variance overestimates the variance; a value larger than one suggests that the prediction error variance underestimates the variance.

Lark, (2000) notes that outliers in the prediction data will influence the squared Z_{score} and suggests to use the median squared Z_{score} instead of the mean since it is a more robust estimator. A median squared Z_{score} equal to 0.455 suggests that the prediction model estimates the prediction uncertainty well.

Quality measures for qualitative soil maps

Like the quality measures for quantitative soil maps, the quality measures for qualitative or categorical soil maps (e.g. soil classes) are defined for the population, i.e. all locations in the mapping area. The basis for map quality assessment of qualitative maps is the error matrix (Brus et al., 2011; Lark, 1995). This matrix is constructed by tabulating the observed and predicted class for all locations in the mapping area in a two-way contingency table (Figure 1). The population error matrix is a square matrix of order

U, with *U* being the number of soil classes observed and mapped. The columns of the matrix correspond to observed soil classes and the rows to predicted soil classes (the map units). *N* is the total number of locations of the mapping area. Elements N_{ij} are the number of locations mapped as class *i* with observed class *j*. The row margins N_{i+} are the locations mapped as class *i*, and column margins N_{+j} the locations for which the observed soil class is *j*. Note that the elements of the population error matrix can also be interpreted as surface areas. In that case element N_{ij} is the surface area mapped as class *i* with observed class *j*.

		Observed					
		1	2	.	.	*U*	Σ
Mapped	1	*N11*	*N12*	.	.	*N1U*	*N1+*
	2	*N21*	*N22*	.	.	*N2U*	*N2+*
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	U	*NU1*	*NU2*	.	.	*NUU*	*NU+*
	Σ	*N+1*	*N+2*	*0*	*0*	*N+U*	*N*

Figure 1. Population error matrix.

From the population error matrix several quality measures can be summarized, though it is strongly recommended that the error matrix is included in a validation assessment. Brus et al. (2011) follow the suggestion by Stehman (1997) that quality measures for categorical maps should be directly interpretable in terms of the probability of a misclassification and therefore recommend the use of three map quality measures: the overall purity, the map unit purity and class representation. We follow this recommendation here. Note that the map unit purity often is referred to as *user's accuracy*, and class representation as *producer's accuracy* (Stehman, 1997). Lark (1995) however, questions the appropriateness of these terms since both quality measures can be important for users as well as producers. He proposes to use map unit purity and class representation instead, which is adopted by Brus et al. (2011) and followed here.

A fourth frequently used group of quality measures are Kappa indices, which adjust the overall purity measure for hypothetical chance agreement (Stehman, 1997). How this chance agreement is defined differs between the various indices. Some authors however, conclude that Kappa indices are difficult to interpret, not informative, misleading and/or flawed and suggest to abandon their use (Pontius and Millones, 2011). These authors argue that Kappa indices attempt to compare accuracy to a baseline of randomness, but randomness is not a reasonable alternative for map construction. We therefore do not consider kappa here.

Overall purity

The overall purity is the fraction of locations for which the mapped soil class equals the observed soil class and is defined as (Brus et al., 2011):

$$p = \sum_{i=1}^{U} N_{uu}/N$$

which is the sum of the principal diagonal of the error matrix divided by the total number of locations in the mapping area. The overall purity can be interpreted as the areal proportion of the mapping area that is correctly classified.

Alternatively, an indicator approach can be used to compute the overall purity. A validation site gets a '1' if the observed soil class is correctly predicted and a '0' otherwise. The overall purity is then computed by taking the average of the indicators.

Map unit purity

The map unit purity is calculated from the row marginals of the error matrix. It is the fraction of validation locations with mapped class *u* for which the observed class is also *u*. The map unit purity for class *u* is defined as (Brus et al., 2011):

$$p_u = \frac{N_{uu}}{N_{u+}}$$

The map unit purity can be interpreted as the proportion of the area of the map unit that is correctly classified. The complement of p_u, $1 - p_u$, is referred to as the error of commission for mapped class *u*.

Class representation

The class representation is calculated from the column marginals of the error matrix. It is the fraction of validation locations with observed class *u* for which the mapped class is *u*. The class representation for class *u* is defined as (Brus et al., 2011):

$$r_u = \frac{N_{uu}}{N_{+u}}$$

The class representation can be interpreted as the proportion of the area where in reality class *u* occurs that is also mapped as class *u*. The complement of r_u, $1 - r_u$, is referred to as the error of omission for mapped class *u*.

Estimating the map quality measures and associated uncertainty

In validation, we estimate the population means of the map quality measures from a sample taken from a limited number of locations in the mapping area. After all, we cannot afford to sample all locations, i.e. each grid cell of our soil map. Because the map quality measures are estimates, we are uncertain about these: we infer the quality measures from only a limited number of observations taken from the population. We do not know the true population means. The estimation uncertainty can be quantified with the sampling variance. From the variance, the lower and upper boundary of a confidence interval, typically the 95%, can be computed using basic statistical theory:

$$CI = (\hat{\bar{x}} - 1.96 \times \frac{\sigma}{\sqrt{n}}; \hat{\bar{x}} + 1.96 \times \frac{\sigma}{\sqrt{n}})$$

where $\hat{\bar{x}}$ is the estimated map quality measure, for instance the ME, MSE or overall purity, σ is the estimated standard deviation of the map quality measure and n is the validation sample size.

Quantified information about the uncertainty associated to map quality measures is useful and required for statistical testing. For instance, if one wants to test if one mapping method performs better than the other method one needs quantified information about uncertainty. Because we are uncertain about the estimated quality measures, an observed difference in map quality between two methods does not necessarily mean that one method is better than the others, even when there is a substantial difference. The difference might be attribute to chance because we infer the quality measures from a limited sample from the population. With statistical hypothesis testing we can calculate how large the probability is that observed difference is caused by chance. Based on the outcome we can accept or reject the hypothesis that there is no difference between the performance of two mapping methods (this would be the null hypothesis for statistical testing) for a given significance level, usually 0.05.

Graphical map quality measures

In addition to quantifying map accuracy statistically, one can also present validation results obtained from a sample graphically. This can be done by creating scatter plots of predicted against observed values and spatial bubble plots of validation errors. Figure 2 shows an example of a scatterplot and bubble plot. Both plots can be easily made with `R` (R Development Core Team, 2016). Use the function `plot(x,y)` to generate a scatter plot. The 1:1 line (black line in Figure 2) can be added to the plot with the command `abline(0,1)`. The spatial bubble plot can be generated with the `bubble` function of the `sp` package (Pebesma and Bivand, 2005).

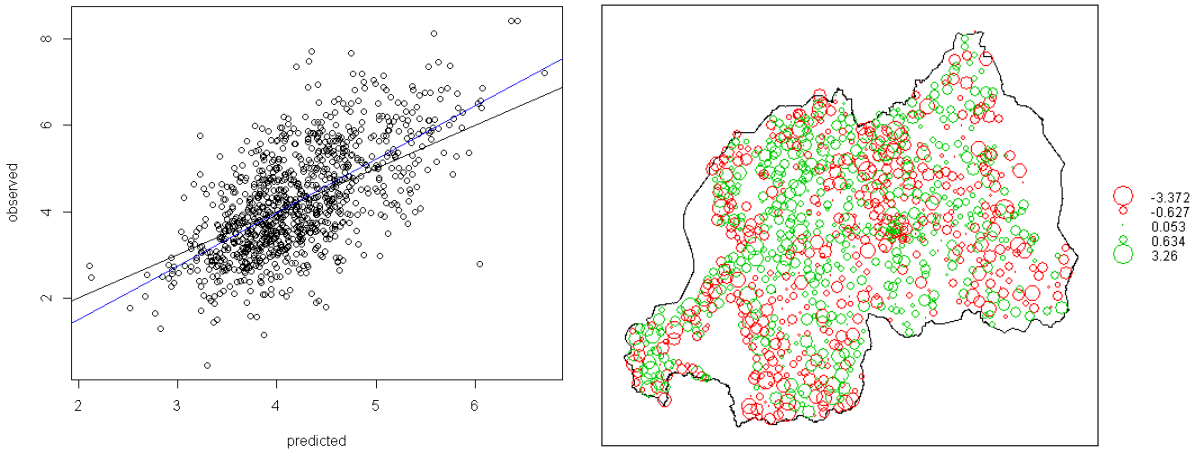

Figure 2. Scatterplot of predicted versus observed soil organic matter content for Rwanda (left) and spatial bubble plot of cross-validation error for soil organic matter (right) (Kempen et al., 2015). The black line in the scatter plot represents the 1:1 line of prediction versus observed, the blue line represents the regression between observed and predicted values (see section 5.3.3).

5.3.3 Validation methods and statistical inference

Following Brus et al. (2011), we introduce and discuss three common validation methods: *additional probability sampling*, *data-splitting* and *cross-validation*, and show how to estimate the map quality measures introduced in previous section from a sample.

With additional probability sampling an independent dataset is collected from the sampling population (all grid cells of a digital soil map) for the purpose of validation. This dataset is used in addition to a dataset that is used to calibrate a prediction model. Such dataset is often a legacy dataset collected with a purposive sampling design.

Data-splitting and cross-validation are applied in situations where one has only one data set available for prediction model calibration and validation. This can be a dataset collected with probability sampling, but in practice this typically is a legacy dataset collected with some purposive sampling design.

We warn here that if one uses data-splitting or cross-validation with a dataset collected with purposive sampling, then this has severe implications on the validity and interpretation of the estimated map quality measures as we will explain below.

Additional probability sampling

The most appropriate approach for validation is by additional probability sampling. This means that an independent validation dataset is collected in the field on basis of a probability sampling design. Validation based on probability sampling ensures one obtains *unbiased* and *valid* estimates of the map quality measures (Brus et al., 2011; Stehman, 1999). Additional probability sampling has several advantages compared to data-splitting and cross-validation using non-probability sample data. These are:

- no model is needed for estimating map quality estimates. We can apply *design-based estimation*, meaning that model-free unbiased and valid estimates of the map quality measures can be obtained;
- discussions on the validity of the estimated map quality are avoided;
- model-free, valid estimates of the variance of the map quality measures can be obtained that allow for hypothesis testing, e.g. for comparison of model performance.

Disadvantages can be extra costs involved in collecting an additional sample or terrain conditions that make it difficult to access all locations in the mapping area.

Probability sampling is random sampling such that:

- all locations in the mapping area have a probability larger than 0 of being selected
- the inclusion probabilities are known but need not be equal.

It should be noted that random sampling is often used for arbitrary or haphazard sampling. Such sampling is not probability sampling because the inclusion probabilities are not known. Design-based, model-free estimation of map quality measures is not possible in this case. All probability samples are random samples but not all random samples are probability samples. The term *probability sampling* should therefore only be used for random sampling with known inclusion probabilities.

There are many different probability sampling designs: simple, stratified, systematic, two-stage, clustered random sampling. We will not give an exhaustive overview here of all these designs. A good resource is de Gruijter et al. (2006). We focus here on two designs: *simple random sampling* and *stratified random sampling*.

Simple random sampling

Design. In sample random sampling no restrictions are imposed on random selection of sampling sites except that the sample size is fixed and chosen prior to sampling (de Gruijter et al., 2006). All sampling locations are selected with equal probability and independently from each other. This can for instance be done as follows (de Gruijter et al., 2006):

1. Determine the minimum and maximum X and Y coordinates of the mapping area (the *bounding box*).
2. Generate two independent random coordinates X and Y from a uniform probability distribution on the interval (x_{min}, x_{max}) and (y_{min}, y_{max})
3. Check if the selected sampling site falls within the mapping area. Accept the sampling site if it does; discard the sampling site if it does not.
4. Repeat steps 2 and 3 until the n locations have been selected.

If a sampling location cannot be visited because of inaccessibility for instance, then this location should be discarded and be replaced by a location chosen from a reserve list. Always the location at the top of the list should be selected for this purpose; not an arbitrarily chosen location from the list, for instance the closest one. It is not allowed to shift an inaccessible sampling location to a location nearby that can be accessed. Irregularity, clustering and open spaces characterise the simple random sampling design (de Gruijter et al., 2006).

Estimation of quantitative map quality measures. For each validation location we compute the error, $e(\mathbf{s}_i) = \hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)$, the absolute error, $|e|(\mathbf{s}_i) = |\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)|$, or squared error, $e^2(\mathbf{s}_i) = (\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i))^2$. The spatial mean of the mapping area for map quality measure x is then estimated by:

$$\hat{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x(\mathbf{s}_i)$$

where i indicates the validation location, $i = 1,2,\ldots,n$, n the validation sample size, and $\hat{\bar{x}}(\mathbf{s}_i)$ the estimated population mean of map quality measure x at location $\mathbf{s}_{i,}$. x is the prediction error in case of the ME, absolute error in case of the MAE, squared prediction error in case of the MSE. Note that the estimator is the unweighted sample mean. This unweighted mean is an unbiased estimator because all sampling locations were selected with equal probability.

The MSDR is estimated by:

$$\widehat{MSDR} = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)\right)^2}{\sigma^2(\mathbf{s}_i)}$$

and the AVE by:

$$\widehat{AVE} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)\right)^2}{\sum_{i=1}^{n}(z(\mathbf{s}_i) - \hat{\bar{z}})^2}$$

where $\hat{\bar{z}}$ is the mean of the target soil property estimated from the validation sample.

One should be careful when assessing the proportion of variance explained by computing the R^2 from a linear regression of the predicted value on the observed value (Krause et al., 2005), as is often done in practice. The R^2 quantifies the dispersion around the regression line; not around the 1:1 line in which we are interested in validation. So it does not directly compare the predicted with observed value as does

the AVE; i.e. it is not based on the prediction error. A high R^2-value therefore, does not automatically mean a high AVE. For instance, in case of strongly biased predictions the R^2 can be high but the AVE will be low. The blue line in Figure 2 is the regression line that one obtains when regression the observed value on the predicted value. This line slightly differs from the 1:1 line. In this example the R^2 of the regression is 0.42 while the AVE is 0.40.

The uncertainty associated to the estimated map quality measures is quantified with the sampling variance, which for the ME, MAE and MSE is estimated by:

$$Var(\hat{\bar{x}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x(\mathbf{s_i}) - \hat{\bar{x}})^2$$

and the 95% confidence interval (CI) of $\hat{\bar{x}}$ is given by:

$$CI_{95} = \hat{\bar{x}} \pm 1.96 \times \sqrt{Var(\hat{\bar{x}})}$$

We should warn here that the calculation of the CI is based on the assumption that the estimated map quality measure means have a normal distribution (the central limit theorem). For the squared errors this assumption can be unrealistic, especially for small sample sizes.

Estimation of qualitative map quality measures. For validation of qualitative soil maps, a sample error matrix is constructed from the validation data (Figure 3). n is the total number of validation locations in the sample. Element n_{ij} of the matrix corresponds to the number of validation locations that have been predicted as class i, $i = 1,2,\ldots,U$ and belong to class j, $j = 1,2,\ldots,U$ (Lark, 1995). The matrix summarizes correct predictions and incorrect predictions within the validation data.

		Observed					
		1	2	.	.	*U*	Σ
Mapped	1	*n11*	*n12*	.	.	*n1U*	*n1+*
	2	*n21*	*n22*	.	.	*n2U*	*n2+*
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	U	*nU1*	*nU2*	.	.	*nUU*	*nU+*
	Σ	*n+1*	*n+2*	*0*	*0*	*n+U*	*n*

Figure 3. Sample error matrix.

From the sample error matrix the overall purity, map unit purity and class representation are estimated by:

$$\hat{p} = \sum_{i=1}^{U} n_{uu}/n$$

$$\hat{p}_u = \frac{n_{uu}}{n_{u+}}$$

$$\hat{r}_u = \frac{n_{uu}}{n_{+u}}$$

Alternatively, the overall purity can be estimated by defining a purity indicator variable for each validation location that takes value 1 if the mapped soil class equals the observed soil class at that location and 0 else. The overall purity is then estimated by:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{s}_i)$$

where $d(\mathbf{s}_i)$ is the indicator variable at validation location $\mathbf{s}_i$. The variance of the estimated overall purity is estimated by:

$$Var(\hat{p}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (d(\mathbf{s}_i) - \hat{p})^2$$

Alternatively, the variance is estimated by:

$$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$$

which is the variance of a binomial probability distribution. The 95% confidence interval of $\hat{p}$ is given by:

$$CI_{95} = \hat{p} \pm 1.96 \times \sqrt{Var(\hat{p})}$$

We warn that the CI as calculated here is a rough approximation which only holds when $n \times \hat{p}$ and $n \times (1 - \hat{p})$ are large (5 as a rule of thumb). Otherwise the binomial distribution should be used to compute the CI.

Figure 4 shows an hypothetical example of a sample error matrix for soil class map. For this example, the overall purity is estimated by: $(19 + 33 + 25 + 42 + 19)/240 = 0.575$, meaning that for an estimated 57.5% of the mapping area the mapped soil class is equal to the true soil class.

		Observed					
		Anthrosol	Cambisol	Gleysol	Luvisol	Podzol	Σ
Mapped	Anthrosol	19	5	3	0	1	28
	Cambisol	5	33	9	13	5	65
	Gleysol	2	8	25	3	5	43
	Luvisol	3	15	9	42	2	71
	Podzol	1	3	8	2	19	33
	Σ	30	64	54	60	32	240

Figure 4. Sample error matrix for a hypothetical soil class map.

Table 1 gives the map unit purities and class representations for this example. The map unit purity of the Gleysol is 0.581, meaning that at 58.1% of the validation locations for which a Gleysol is predicted, a Gleysol is observed. Assuming the validation data were collected by simple random sampling, we could conclude that for 58.1% of the area mapped as Gleysol we would find a Gleysol in the field. The class representation of the Gleysol is 0.463, meaning that for 46.3% of the validation locations classified as Gleysol, we map a Gleysol. The majority of the Gleysol locations is thus mapped as a different soil class. Again, assuming the validation data were collected by probability sampling, we would estimate that 22.5% ($\frac{54}{240} \times 100\%$) of our mapping area is covered by Gleysols. We map Gleysols for 17.9% of the area ($\frac{43}{240} \times 100\%$). It can happen that a soil class has a high map unit purity and a low class representation. This means that if we map a Gleysol we will likely find a Gleysol there, but that a large extent of the true Gleysol area is not mapped as such.

Table 1. Map unit purity and class representation statistics for the hypothetical example given in Figure 4.

	map unit purity	class representation
Anthrosol	0.679	0.633
Cambisol	0.508	0.516
Gleysol	0.581	0.463
Luvisol	0.592	0.700
Podzol	0.576	0.594

Stratified random sampling

Design. In stratified random sampling the area is divided in sub-areas, called *strata* in each of which simple random sampling is applied with sampling sizes chosen prior to sampling. Three attributes should be chosen for this design: 1) the definition of strata; 2) the total sample size; 3) the allocation of sample sizes to the strata (de Gruijter et al., 2006).

Strata can for instance be soil or land cover classes derived from a map, or accessibility classes if terrain accessibility is an issue. Strata can also be formed on basis of so called *compact geographical strata* (Brus, Spatjens, de Gruijter, 1999). In this case the mapping area is divided into *K* geographical strata, for instance by *k*-means clustering, of equal size. This ensures good spatial coverage over the mapping area. Creation of geographic strata using *k*-means clustering can be implemented in `R` with the `spcosa`

package (Walvoort et al., 2010). Selection of sampling sites within each stratum is done with the same routine as for simple random sampling.

Estimation. The spatial mean of the area for map quality measure x is estimated by (de Gruijter et al., 2006):

$$\hat{\bar{x}} = \sum_{h=1}^{H} a_h \hat{\bar{x}}_h$$

where H is the number of strata, a_h is the relative area of stratum h, and $\hat{\bar{x}}_h$ is the estimated mean for stratum h:

$$\hat{\bar{x}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x(\mathbf{s_i})$$

where n_h is the number of validation locations in stratum h. The sampling variance of the estimated map quality measure is estimated by:

$$Var(\hat{\bar{x}}) = \sum_{h=1}^{H} a_h^2 Var(\hat{\bar{x}}_h)$$

where $Var(\hat{\bar{x}}_h)$ is stratum variance that is the estimated by:

$$Var(\hat{\bar{x}}_h) = \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (x_h(\mathbf{s_i}) - \hat{\bar{x}}_h)^2$$

Note that the inclusion probabilities are accounted for by using the relative areas of the strata. Also note that substituting $d(\mathbf{s_i})$ for $x(\mathbf{s_i})$ and p_h for x_h gives the estimates of the overall purity and its variance.

Data-splitting

In data-splitting the sample data set is split in two subsets. One subset is used to calibrate the prediction model. The other subset is used for validation. A frequently used splitting criterion is 70-30, where 70% of the sample data are used for calibration and 30% for validation. The choice of a splitting criterion however, is arbitrary and it is not evident how to split a data set in such a way that unbiased and valid estimates of the map accuracy can be obtained. For sparse data sets, data-splitting can be inefficient since the information in the data set is not fully exploited for both calibration and validation.

It is important to note here that a random subsample of (legacy) data that are collected with a purposive (non-probability) design, is *not* a probability sample of the study area. This means that design-based estimation of map quality measures is not possible.

If a validation (sub)sample is a non-probability sample of the mapping area, then we must account for possible spatial autocorrelation of the prediction errors when estimating the map quality measures. One can imagine that when two validation locations are close together and the prediction errors are correlated that there is less information in these two locations (there is information redundancy because of autocorrelation) than in two isolated locations. This information redundancy has to be accounted for when estimating map quality measures and implies that we have to rely on model-based estimation: a model for the spatially autocorrelated prediction error has to be assumed. Thus, we will not obtain model-free, unbiased and valid estimates of the quality measures from non-probability sample validation data. In a case study, Knotters and Brus (2013) showed that model-based predictions of producer's accuracies from two models differed strongly, indicating that with the model-based approach the validation results strongly depend on model assumptions.

In most studies however, spatial correlation is not accounted for when estimating map quality measures using the estimators presented above under 'Simple random sampling' from non-probability sample data. In such case, the quality measures cannot be considered unbiased and valid estimates of the population means of the map quality measures. In addition, the estimated variance of the map quality measures is not valid and statistical testing of mapping methods to assess which method gives the most accurate predictions cannot be done.

In other words, if the simple random sampling estimators are used to estimate map quality measures then these are only valid for the validation data points. The map quality measures do not give a valid estimate of the quality of the map as a whole (the population). For instance, the overall purity cannot be interpreted as an areal proportion of correctly mapped soil classes, only as the proportion of the validation data points for which the soil class is correctly predicted.

Cross-validation

In *K*-fold cross-validation (CV), the dataset is split into *K* roughly equal sets. One of these sets is set aside for validation. The model is then calibrated using the data from the *K*-1 sets and used to predict the target variable for the data points set aside. From this prediction the prediction error is calculated. This procedure is repeated *K* times, each time setting a different set aside for validation. In this way we obtain *K* estimates of the prediction error: one for each validation sample site. In this way, all data are used for validation and model calibration. It is thus much more efficient than data-splitting.

K is typically chosen as 5 or 10, or as *N* the number of data points. The latter is referred to as leave-one-out cross-validation (LOOCV) in which only one validation site is set aside in each iteration. The model is then calibrated with *N*-1 observations. Some repeat *K*-fold cross-validation a number of times and average the results to obtain a more robust estimate of the map quality measures.

Note that the problem of spatially correlated errors remains when data are non-probability sample data. Cross-validation using a non-probability sampling dataset suffers from the same drawbacks with respect to unbiasedness and validity of the estimates of the map quality measures as data-splitting. The estimates cannot be interpreted as being valid for the mapping area, but only for the validation locations.

In `R`, the `caret` package (Kuhn, 2015) offers functionality for data-splitting and cross-validation.

References

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., Morrás, H.J.M., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281, 102-118.

Brus, D.J., Spätjens, L.E.E.M., de Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89(1–2), 129-148.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for natural resource monitoring. Springer.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., De Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci Soc Am J 76(6), 2097-2115.

Kempen, B., Heuvelink, G.B.M., Brus, D.J., Stoorvogel, J.J., 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. European Journal of Soil Science 61, 333-347.

Kempen, B., Vereijken, P., Keizer, P., Ruiperez Gonzalez, M., Bindraban, P., Wendt, J., 2015. Preliminary evaluation of the feasibility of using geospatial information to refine soil fertility recommendations., VFRC Report 2015/6. Virtual Fertilizer Research Centre, Washington D.C.

Knotters, M., Brus, D.J., 2013. Purposive versus random sampling for map validation: a case study on ecotope maps of floodplains in the Netherlands. Ecohydrol. 6(3), 425-434.

Krause, P., Boyle, B.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5, 89-97.

Kuhn, M., 2015. A Short Introduction to the caret Package. Available at: https://cran.r-project.org/web/packages/caret/index.html.

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(8), 1461-1480.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51, 137-157.

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5(2).

Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32(15), 4407-4429.

R Development Core Team, 2016. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244, 214-227.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 62(1), 77-89.

Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. International Journal of Remote Sensing 20(12), 2423-2441.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. Journal of Soil Science 41(3), 473-490.

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 & Geosciences 36, 1261-1267.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists. Statistics in practice. Second edition ed. John Wiley & Sons, Chichester