Published May, 2011

Using Additional Criteria for Measuring the Quality of Predictions and Their Uncertainties in a Digital Soil Mapping Framework

Brendan P. Malone*
Faculty of Agriculture
Food & Natural Resources
The Univ. of Sydney
S207 John Woolley Building
NSW 2006
Australia

Jaap J. de Gruijter
Wageningen Univ. and Research Centre
P.O. Box 47
6700 AA Wageningen
The Netherlands

Alex B. McBratney
Faculty of Agriculture
Food & Natural Resources
The Univ. of Sydney
S224 John Woolley Building
NSW 2006
Australia

Budiman Minasny
Faculty of Agriculture
Food & Natural Resources
The Univ. of Sydney
S222 John Woolley Building
NSW 2006
Australia

Dick J. Brus
Wageningen Univ. and Research Centre
P.O. Box 47
6700 AA Wageningen
The Netherlands

In this paper we introduce additional criteria to assess the quality of digital soil property maps. Soil map quality is estimated on the basis of validating both the accuracy of the predictions and their uncertainties (which are expressed as a prediction interval [PI]). The first criterion is an accuracy measure that is different in form to the usual mean square error (MSE) because it accounts also for the prediction uncertainties. This measure is the spatial average of the statistical expectation of the mean square error of a simulated random value (MSES). The second criterion addresses the quality of the uncertainties which is estimated as the total proportion of the study area where the $(1-\alpha)$–PI covers the true value. Ideally, this areal proportion equals the nominal value $(1 - \alpha)$. In the Lower Hunter Valley, NSW, Australia, we used both criteria to validate a soil pH map using additional units collected from a probability sample at five depth intervals: 0 to 5, 5 to 15, 15 to 30, 30 to 60, and 60 to 100 cm. For the first depth interval (0–5 cm) in 96% of the area, the 95% PI of pH covered the true value. The root mean squared simulation error (RMSES) at this depth was 1.0 pH units. Generally, the discrepancy between the nominal value and the areal proportion in addition to the RMSES increased with soil depth, indicating largely a growing imprecision of the map and underestimation of the uncertainty with increasing soil depth. In exploring this result, conventional map quality indicators emphasized a combination of bias and imprecision particularly with increasing soil depth. There is great value in coupling conventional map quality indicators with those which we propose in this study as they target the decision making process for improving the precision of maps and their uncertainties. For our study area we discuss options for improving on our results in addition to determining the possibility of extending a similar sampling approach for which multiple soil property maps can be validated concurrently.

Abbreviations: ATV, all terrain vehicle; CCC, Lin's concordance correlation coefficient; CI, confidence interval; DSM, digital soil mapping; GPS, global positioning satellite; IMP, imprecision; ME, mean error; MSE, mean square error; MSES, mean squared simulation error; PI, prediction interval; PIC, cluster prediction interval; PICP, prediction interval coverage probability; PL, prediction limit; RMSE, root mean square error; RMSES, root mean squared simulation error; SI, simple random sample design; STSI, stratified simple random sample design; SU, sampling units.

Quantifying soil map quality has been an area of sustained research for well over 40 yr, with seminal papers dating back to the 1960s and 1970s; e.g., Webster and Beckett (1968) and Burrough et al. (1971). Brus et al. (2011) demonstrate not only the continual development of methodologies for validating soil maps but also the importance of quantifying soil map accuracy. Grunwald (2009) however, points out that while digital soil mapping (DSM) has become popularized in recent times, it is frequently the case that maps are not validated.

Generally, soil map quality has been related to measures of accuracy (Finke, 2007). These are conventionally based on measures of variance between observed and predicted values (Bishop et al., 2001). In soil attribute mapping, this is quantified by the goodness of fit (R^2) or the mean of the squared prediction error (MSE). For assessing the quality of categorical soil maps measures based on mapping unit purity and user's and producer's accuracies are the most common (Lark, 1995). While useful, these measures are limited because it is possible only to estimate the

Soil Sci. Soc. Am. J. 75:1032–1043
Posted online 28 Apr. 2011
doi:10.2136/sssaj2010.0280
Received 20 July 2010.
*Corresponding author (brendan.malone@sydney.edu.au).

accuracy of the predictions. While a number of tools are available that allow one to express or quantify the level of uncertainty in soil functions generated from uncertain predictions of basic soil properties (Minasny and McBratney, 2002; Brown and Heuvelink, 2005), rarely do we consider the quality or appropriateness of them in a practical sense. For example, given a level of uncertainty regarding a prediction of a soil attribute across a spatial extent, what implications does this have in how we interpret soil phenomena or manage the soil resource in question? Is it possible to optimize inputs such as lime for soil pH management given a prescribed level of uncertainty? It is apparent therefore that there needs to be additional measures of map quality to those conventionally reported which also take into account the quality of the estimates of prediction uncertainty for a particular soil attribute. By incorporating this additional information, we are able to consolidate our limited understanding of soil variability with reciprocal measures of soil map quality.

A useful and empirical approach to estimating the uncertainty of model outputs was proposed by Shrestha and Solomatine (2006) where uncertainty was expressed in the form of two quantiles (constituting a PI) of the underlying distribution of prediction errors (residuals). The PI explicitly takes into account all sources of uncertainty and circumvents attempts to separate out the contribution of each source (Shrestha and Solomatine, 2006; Solomatine and Shrestha, 2009). In Malone et al. (2011) the empirical uncertainty approach of Shrestha and Solomatine (2006) was adapted within a DSM framework to map PIs across a study area. To that end, the soil map quality criteria we propose in this study are largely based on the empirical coverage of PIs.

In terms of quality measures for validating soil maps, Brus et al. (2011) provide a comprehensive review of those used for both categorical and quantitative digital soil maps. Map validation by probability sampling involves the random selection of additional test units (observations at locations not used for model calibration) from a study area. In probability sampling all units within a study area have a positive probability of being selected; where the probabilities are determined by the sampling design and can be derived from this design (de Gruijter et al., 2006). Brus et al. (2011) concluded that probability sampling is the more superior validation method (in comparison with data splitting or cross-validation) because unbiased estimates of the measures of soil map quality can be obtained by 'design-based' inference and thus are free of model assumptions (Brus and de Gruijter, 1997). This is generally not the case for data splitting or cross-validation (Brus et al., 2011).

Overall the aim of this research therefore is to present and illustrate new criteria (in addition to those conventionally reported) for evaluating both the quality of predictions and the quality of quantifications of the uncertainty in a DSM. Validation is performed on the basis of a probability sample to collect additional sampling units from the study area.

SOIL MAP TO BE VALIDATED

Study Area

The area selected for this study is an approximately 220 km^2 area north of the town of Cessnock (32°49′48″S 151°20′59″E) in the Lower Hunter Valley, approximately 140 km north of Sydney, NSW, Australia (Fig. 1). Topographically, this area consists mostly of undulating hills that ascend to low mountains to the south-west. This area is part of the Sydney Basin where parent materials are composed mostly of Mesozoic sandstones and shales (Thackway and Cresswell 1995). The dominant soil types are Typic Natrudalfs and, on topographic rises, Typic Calciudolls (Odgers, 2010). In terms of landuse, dryland agricultural grazing systems are predominant, followed by an expansive viticultural industry. While most of the land has been dedicated for these uses, tracts of remnant natural vegetation (dry forest) are apparent, particularly toward the south-western (Broken Back Range), eastern (Werakata National Park) and northern margins of the study area. See Bell (2004) for further details of the environmental setting of this area.

For this study we chose soil pH as the target variable. The reason for this selection was because viticulture is an extensive industry across the study area and many management decisions are centered on the nutrient status of the high value wine grape crops, which is generally dependent on soil pH (White, 2003). Soil nutritional status affects all parts of the grape vine, from root growth and distribution through to shoot growth and grape composition.

The soil dataset we used contains 994 sites where pH was recorded at each horizon and/or at specific depths to at least 1 m. Three hundred of the soil samples were collected in 2004 using the Conditioned Latin Hypercube sampling method (Minasny and McBratney 2006) where compound topographic index, parent material, and Normalized Difference Vegetation Index were used as auxiliary variables to provide environmental information. The remaining 694 samples were collected sporadically between 2001 and 2009 mainly using toposequence sampling design (Odgers et al., 2008).

MATERIALS AND METHODS

Digital Soil Mapping of the Predictions

A soil map depicting the lateral and vertical distribution of soil pH across the study area was generated following the procedure of Malone et al. (2009) which uses an amalgam of soil depth spline functions and DSM techniques (Fig. 2b) (only 0–10 cm shown). The spatial entity of the map is point support where point estimates were made on a 25-m regular grid. The vertical resolution at each grid node is 1 cm.

This map was generated using a regression kriging approach where the predictions were based on the calibration dataset of 994 soil profile descriptions distributed across the area (Odgers, 2010). To standardize prediction depths, mass preserving splines were fitted individually to each profile before mean estimates of pH were taken at 0 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50, 50 to 70, 70 to 80, and 80–100 cm. The deterministic component (of regression kriging) used a neural network where the target variables were the mean observations at each of the standardized depths. These were modeled against a suite of environmental covariates derived from 25 m rasters of a digital elevation model (DEM) and various derived terrain attributes; Landsat 7 ETM+ band data and various band derivatives. The model formulae derived from this

Fig. 1. Lower Hunter Valley study area with respect to location in New South Wales (large box) and Australia (small box).

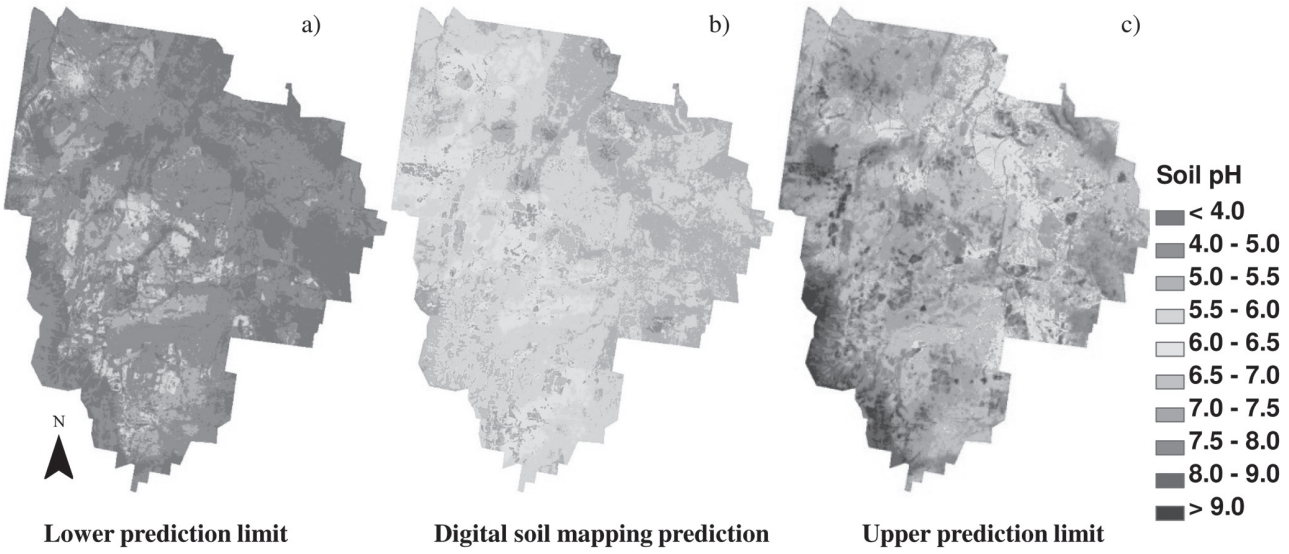

Fig. 2. Soil pH map to be validated (displaying only the 0- to 10-cm interval). Soil map depicts the (b) digital soil mapping prediction and the (a) upper and (c) lower prediction limits which constitute a prediction interval.

procedure were then applied across the extent of the study area, where covariate information existed only.

For the stochastic component, model residuals were independently mapped for each standardized depth with kriging using localized variograms (exponential function) of the 100 nearest neighbors to a prediction point. Both the deterministic and stochastic components were summed to arrive at a final prediction for each standard depth at each point. The standard depths at each grid node were then used as parameters to construct continuous profiles using the mass-preserving spline function (see Malone et al., 2009).

Digital Soil Mapping of the Uncertainties

First, a PI should be distinguished from a confidence interval (CI). A CI is an interval estimate of a fixed quantity/parameter (e.g., a spatial mean in a design-based approach), whereas a PI is an interval prediction of a random quantity (e.g., the value of the target variable at a given location or the spatial mean in a model-based, geostatistical approach (de Gruijter et al., 2006). A PI is bounded by its upper and lower limits and by definition, the prediction limits for a given confidence level (1−a)%, enclose the true but unknown target value (1−a)%, of times on average (Shrestha and Solomatine, 2006).

Calculating the PI for a prediction node following the method of Shrestha and Solomatine (2006) is performed independently of the model building or calibration process and is referred to as the "uncertainty estimation based on local errors and clustering" (UNEEC) method (Solomatine and Shrestha, 2009). The purpose of the UNEEC is to derive the upper and lower prediction limits based on the model error, and since it is estimated through an empirical distribution, it is not necessary to make any assumption about residuals (Solomatine and Shrestha, 2009). Using the calibration data, the adaption of the UNEEC for DSM (Malone et al., 2011) first uses fuzzy k-means with extragrades to partition the environmental covariate data into clusters which share similar environmental attributes (McBratney and de Gruijter, 1992). Fuzzy k-means with extragrades is used on the basis that extragrades are considered the outliers of the data, and as Tranter et al. (2010) also point out, they are also spatially in regions of low density calibration data, which confers a low reliability of prediction in these areas. Ultimately each observation is given a grade of membership to all clusters including the extragrade cluster.

Next, model output error is determined. The regression kriging approach similar to that used for the DSM prediction procedure was used except final predictions were evaluated by leave-one-out-cross-validation (Hastie et al., 2009). The only output required here is the residual between the observed value of the target variable and the corresponding final prediction at each calibration point.

Once the clusters and model output error have been identified, the PIs for each cluster are computed from the empirical distributions of the corresponding residuals. Therefore, one underlying assumption of the empirical uncertainty method is that particular areas (clusters) within a landscape will have similar residuals or distribution of residuals and ultimately share a similar range of uncertainty (Malone et al., 2011). To construct a cluster 95% PI, first we assign each calibration point to the cluster it has the highest membership to. Then for each cluster, the 2.5 and 97.5 percentile values are taken from the empirical distribution of residuals for the lower and upper prediction limits, respectively. Those points that belong to the extragrade cluster are handled differently in that we impose a penalization by way of a multiplier, as proposed by Tranter et al. (2010):

$$\mathrm{PIC}_{ej}^{L} = 2 \times q_{2.5}$$
$$\mathrm{PIC}_{ej}^{U} = 2 \times q_{97.5} \qquad [1]$$

where PIC_{ej}^{L} and PIC_{ej}^{U} are the lower and upper prediction limits of the extragrade class, and q is the quantile value of the extragrade cluster error distribution at each depth increment.

Computing the PI for each prediction node across the study area first requires each node to be assigned membership grades to each of the clusters characterized in the calibration procedure. These are determined on the basis of the environmental covariates at each node and the predetermined cluster centroids. From this, we determine the prediction node PI using the weighted mean of the PI of each cluster (Shrestha and Solomatine 2006). This can be defined mathematically as:

$$\mathrm{PI}_{j}^{L} = \sum_{i=1}^{c} m_{ij} \mathrm{PIC}_{j}^{L}$$
$$\mathrm{PI}_{j}^{U} = \sum_{i=1}^{c} m_{ij} \mathrm{PIC}_{j}^{\mathrm{U}} \qquad [2]$$

where PI_{i}^{L} and PI_{i}^{U} correspond to the weighted lower and upper PI for the *i*th prediction node, PIC_{j}^{L} and PIC_{j}^{U} are the lower and upper PIs for each cluster *j* (including extragrade cluster) as determined from the calibration sites, and m_{ij} is the membership grade of *i*th prediction node to cluster *j*. Finally, the lower and upper prediction limits (PL_{i}^{L} and PL_{i}^{U} respectively) are then derived for each prediction node by adding the prediction (from the DSM procedure) to PI_{i}^{L} and PI_{i}^{U}.

All the steps after the clustering process are repeated for each of the standardized depths. Figure 2a and 2c show the lower and upper prediction limits of pH across the study area for the 0- to 10-cm depth increment. Mass preserving splines are also used to construct continuous representations of the upper and lower prediction limits of pH by using the estimated PL_{i}^{U} and PL_{i}^{L} values at each depth respectively as parameters.

Concepts of Soil Map Quality

In this study we use a design-based approach to derive external accuracy estimates of soil map quality in terms of both the predictions and their uncertainties. This involves the use of additional data collected from a probability sample for which the specifics will be discussed later in more detail. Using such a design-based approach means that our estimates of map quality are model-free and unbiased (de Gruijter et al., 2006; Brus et al., 2011).

For this study, the mapped predictions and their uncertainties are on point support. This means that the additional independent sampling units should also have point-support. As the first criteria of our proposed soil map quality indicators we use a measure which assesses the accuracy of predictions (but also taking into account the uncertainties). As discussed previously, the usual measure for the accuracy of quantitative map predictions is the mean squared prediction error (MSE) which is defined as:

$$\mathrm{MSE}=\frac{1}{\|A\|}\int_{s\in A}\left\{z_p(s)-z(s)\right\}^2\,\mathrm{d}s \qquad [3]$$

where A is the mapped area and $z_p(s)$ and $z(s)$ are the predicted and true values at location s of the target variable, here pH. Additionally bias, expressed as the mean error (ME) can then also be derived, which has the general formula:

$$\mathrm{ME}=\frac{1}{\|A\|}\int_{s\in A}\left\{z_p(s)-z(s)\right\}\mathrm{d}s \qquad [4]$$

These conventional measures quantify only the error and bias of the predictions, it does not account for the uncertainty of the local predictions. To this end we introduce the 'Mean Squared Error of Simulation' (MSES). The underlying idea can be simply explained by imagining that a value is simulated randomly at a given location, using the local prediction and the local prediction variance given by the map as parameters of a local error distribution. The measure that we propose is the spatial average of the statistical expectation of the mean squared error of a simulated random value:

$$\mathrm{MSES}=\frac{1}{A}\int_{s\in A}E_s\left\{z_s(s)-z(s)\right\}^2\mathrm{d}s \qquad [5]$$

where $z_s(s)$ and $z(s)$ are the simulated and true values, respectively, at location s and E_s is the statistical expectation over the error distribution. For estimation of the MSES, it is equated as the summation of two components; a spatially averaged (squared) local bias component, identical to the MSE and a spatially averaged local precision component:

$$\begin{aligned}\mathrm{MSES}&=\frac{1}{\|A\|}\int_{s\in A}E_s\left\{z_p(s)-z(s)\right\}^2\mathrm{d}s+\\&\frac{1}{\|A\|}\int_{s\in A}E_s\left\{z_p(s)-z_s(s)\right\}^2\mathrm{d}s\\&=\mathrm{MSE}+\overline{\sigma^2}\end{aligned} \qquad [6]$$

where $\sigma^2(s)$ is the variance of the prediction error as given by the map at location s. Note that the second component can be simply obtained by averaging the σ^2 over the entire map grid. An estimate of MSES can therefore be obtained by estimating MSE from the sample data, and adding $\overline{\sigma^2}$ to it:

$$\widehat{\mathrm{MSES}}=\widehat{\mathrm{MSE}}+\overline{\sigma^2} \qquad [7]$$

The second criterion is expressed as the areal proportion of the mapped area where the measured value at a specified depth interval fits within the bounds of its estimated $(1-\alpha)$ % PI (the PI covers the measured value), shortly referred to hereafter as the areal proportion correctly predicted (APCP):

$$\mathrm{APCP}=\frac{1}{\|A\|}\int_{s\in A}i(s)\,\mathrm{d}s \qquad [8]$$

where $i(s)$ equals 1 if the true pH value at location s is covered by the PI given by the map at s, and 0 otherwise.

Conceptually, both criteria form an agreeable pair and ideally should be reported together. The first explicitly deals with the accuracy of the simulations, while the second signals either an under- or over-coverage (PIs too narrow or too wide) for a given confidence level. As a case in point and the necessity for reporting on both criteria; it is possible to have an inaccurate map (wide prediction intervals) with the same coverage proportion (APCP) as an accurate map. Alone, the second criterion cannot tell the difference between the two. Therefore coupling the APCP with the MSES allows penalization to be given where predictions are found to be inaccurate.

Probability Sampling Design

A stratified simple random sampling design (STSI) (de Gruijter et al., 2006) was used to select the sampling locations at which soil pH was to be laboratory analyzed at specified depth increments. This design was chosen over a simple random sample (SI) design on the basis that greater efficiency can be expected in terms of smaller sampling variance of the estimated map quality measures from the same number of samples (Brus et al., 2011). We used two stratification variables: the depth-averaged whole-profile prediction of pH and an uncertainty measure; the depth-averaged whole-profile difference between the upper and lower prediction limits. The averaged prediction and the uncertainty measure for each mapped point location were plotted on a graph from which four equal-area strata were generated empirically by shifting the threshold values of each stratification variable (Fig. 3a). The total number of prediction nodes equaled 353,316 making the size of each stratum 88,329 nodes. The characteristics or threshold values for each stratum are summarized in Table 1. Figure 3b shows the spatial extent of the strata across the study area where it is worth noting that the strata are not from contiguous areas. From knowledge of the soil landscape across the study area, the strata which have the highest uncertainty (C and D; widest PIs) appear to cover areas of mountainous terrain (south-west), densely forested areas (east), and areas of drainage, for example, creek lines and areas of low relief. A low density of prediction sites coupled with complex terrain and for most part inaccessible, it is quite logical that uncertainty is high in these landscape settings. Conversely, the strata that have a lower uncertainty (A and B) exist predominantly where the relief is less complex, the land has been cleared and where most agricultural pursuits are concentrated in the area such as grazing and viticultural production. For example, Stratum B, is a good indicator where viticulture is practiced as the soils are a little higher on the landscape and are generally characterized by higher pH levels.

A total of 100 sampling units (SUs) were used to validate the map in this study. These were allocated proportionally to the strata, so that from each stratum 25 nodes were randomly selected fully randomly from the potential 88,329 nodes. A handheld global positioning satellite (GPS) receiver was used to locate the positions of the SUs within the field. The SUs were soil cores of between 100 and 120 cm in length and a diameter of 5 cm. These were taken using a hydraulic geoprobe soil corer mounted on the back of a truck/all terrain vehicle (ATV).

In the laboratory, each SU was subsampled corresponding to the depth intervals of: 0 to 5, 5 to 15, 15 to 30, 30 to 60, and 60 to 100 cm. Once mixed, a small aliquot from each depth interval was analyzed using the 1:5 soil/water suspension method to determine soil pH

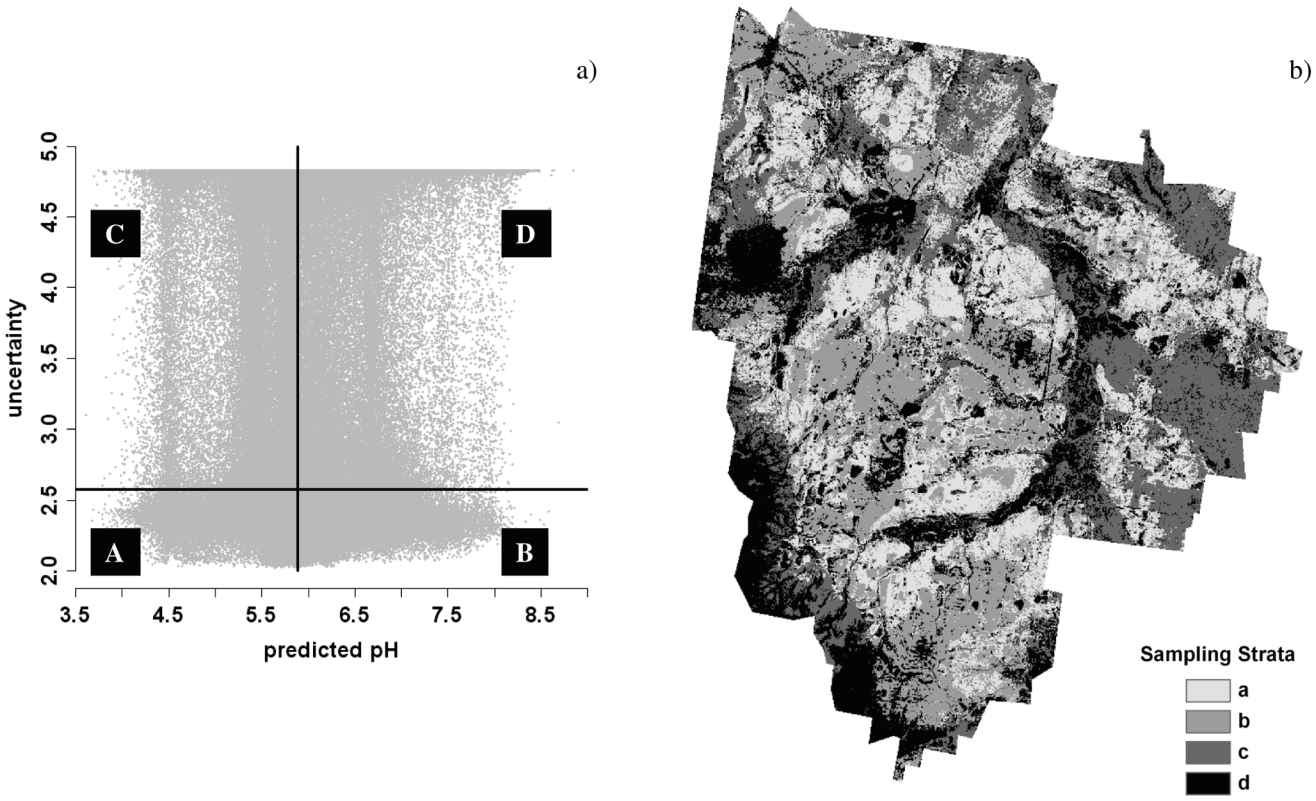

Fig. 3. Determination of sampling strata and their subsequent spatial extent. (a) Plot illustrating the process for constructing equal-area strata where the stratification variables were the depth-averaged whole-profile pH prediction and depth-averaged whole-profile difference between upper and lower prediction limits. Black lines indicate the threshold values for demarcation of each Stratum A, B, C, and D. (b) Spatial coverage of the equal-area strata across the study area.

(Rayment and Higginson, 1992). Randomly selected, duplicate aliquots totaling 50 were also analyzed to estimate measurement error for reasons described further on.

Design-Based Estimation of Soil Map Quality

The statistical inference for estimating the areal proportions and the standard errors of point locations which fit within their PI is given in de Gruijter et al. (2006). The indicator value for each SU at each depth was evaluated by:

- First, the PL_i^L and PL_i^U estimates corresponding to each of the eight prediction depths were used as the spline parameters for the construction of continuous representations of the prediction limits;
- The continuous PL_i^L and PL_i^U estimates were then queried to derive the mean of the prediction limits for each of the five sampling depths at the validation points;

Table 1. Threshold values determined empirically of the stratification variables: depth-averaged whole-profile prediction of pH and depth-averaged whole-profile difference between the upper and lower prediction limits for each stratum.

Stratum	pH prediction	Uncertainty (95% PI)
A	Low (≤5.9)	Narrow (≤2.6)
B	High (>5.9)	Narrow (≤2.6)
C	Low (≤5.9)	Wide (>2.6)
D	High (>5.9)	Wide (>2.6)

- Analysis was performed to determine whether the observed pH value fitted within its corresponding PI. Indicator values of either a one (1) indicating a fit within the bounds of the PI, or zero (0) not fit within the bounds of the PI was assigned for each observation.

From Eq. [8] the design-based estimator for APCP in this study is:

$$\widehat{\mathrm{APCP}}=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A\|_h}{n_h}\sum_{j=1}^{n_h}I_{hj} \qquad [9]$$

where A_h and n_h are the surface area and sample size (number of selected validation points, here 25) of Stratum h, respectively, and I_hj is the indicator value as determined at the j-th sample point of Stratum h. The reason why we use I_{kj} instead of i_{kj} is because we do not possess *true* values of pH and thus need to work from the *measured* pH values, meaning that the indicator values are thus subject to random error. The measurements that are used for validation may be accurate enough that one can safely assume that the effects of measurement error on the validation results are negligible. However, that assumption cannot be made in this study; therefore we need to consider the effects of measurement error. Contained within Appendix A is an explanation of accounting for measurement error in map validation. Also in Appendix A are the detailed calculations for estimation of the error variance of $\widehat{\mathrm{APCP}}$ due to both the sampling and measurement errors.

For estimation of the MSES, Eq. [7] is followed. As only $\widehat{\mathrm{MSE}}$ is subject to sampling error, the sampling variance and standard error of $\widehat{\mathrm{MSES}}$ equal those of $\widehat{\mathrm{MSE}}$. Finally, an estimate at the scale of the

mapped variable, the root mean squared error of simulation (RMSES), is obtained by taking the square root of $\widehat{\text{MSES}}$. Similarly for the limits of a confidence interval.

The methodology for estimation of the APCP is evaluated on the basis of a 95% PI. One way to assess both the performance and sensitivity of the PIs is to estimate the APCP using a range of confidence levels (besides 95%). Variants of this type of validation, termed the prediction interval coverage probability (PICP) have previously been shown to be an important validation criterion in other studies for example, Shrestha and Solomatine (2006) and Tranter et al. (2010). Akin to the estimation of a 95% PI, we performed the same procedures previously described to construct PIs with confidence levels ranging from 5 to 90% and 99%. For each confidence level, estimation of the APCP was derived in the same way as described for a 95% confidence level. The PICP is a valuable indicator of the validity of the uncertainty model where it is said to be optimal when the PICP value is close to the range of corresponding 100 $(1-\alpha)$% confidence levels.

Conventional Measures of Map Quality

For conventional measures of map quality, we compared the predicted values of pH at each depth for each SU with the corresponding observed value. The method (using mass-preserving splines) for evaluating the prediction at each sampling depth was the same as for deriving the lower and upper prediction limits.

We derived spatial estimates of accuracy, bias, and imprecision at each depth interval, where accuracy is stated in terms of the root mean squared error (RMSE) which is estimated by taking the square root of the predicted MSE (Eq. [3]). As we cannot assume that the measurements used for validation are free of the effects of measurement errors, we need to make allowances for them within our calculation. Appendix A provides unbiased estimates for the MSE and variance of MSE in the presence of measurement errors. Confidence limits for the RMSE can be calculated as the square roots of the confidence limits for MSE, whereby $\widehat{\text{MSE}}$ is assumed normally distributed. For calculating the variance of the ME (bias) there is no need to correct for measurement error, because no nonlinear transformation is involved which means it is automatically accounted for in the standard estimation of the variance.

Imprecision was simply calculated as the square root of the difference between the $\widehat{\text{MSE}}$ and the squared $\widehat{\text{ME}}$ defined as:

$$\text{IMP}=\sqrt{\widehat{\text{MSE}}-\left(\widehat{\text{ME}}\right)^2} \qquad [10]$$

RESULTS AND DISCUSSION

As a preamble to interpreting the output of our proposed map quality criteria, there are four general outcomes to consider. The first is that the APCP is equal to or very near (above or below) the nominal confidence level and the RMSES is small. For this result the predictions and simulations are accurate but more importantly, what uncertainty there is, is adequately accounted for and covered by the PI. This result is ideal, however the map user will need to consider the level of uncertainty they are willing to accept for the given purpose of the map. In the case of this study, the purpose may be the optimization of liming where viticulture is practiced. Nevertheless, given that the accuracy is quite high for this first outcome, logic will generally indicate that the uncertainties will also be quite low (narrow prediction intervals). The second outcome is the situation where the APCP is as the first but the RMSES is large. In this result the mapped predictions are inaccurate, but the PIs manage to account for most if not all the known sources of uncertainty. This is not an ideal result, because ultimately the map user wants an accurate map and the PIs could be considered to be too wide, thus making precise management decisions difficult for example, the optimization of lime inputs. It should be avoided however to interpret a map as being either high or low quality without consideration of the intended purpose of the map. Nevertheless, this second result is more ideal than the third whereby the difference between the APCP and the nominal confidence level is large, and the RMSES is high. This result indicates an inaccurate map in addition to the presence of other sources of uncertainty that were not accounted for and/or bias. The fourth outcome is where the RMSES is small but the APCP is far from the nominal confidence level. Given the occurrence of any one of the four outcomes, it then becomes necessary to investigate reasons why for example the RMSES is large but the APCP is ideal (second outcome) etc. It is therefore worth stressing at this point that conventional map quality indicators should also be used as a companion to our proposed criteria because they address explicitly issues of bias and imprecision.

In terms of the results from this study, the PI was computed for a 95% confidence level. With this basis, the ideal APCP for any of the specified depth intervals should be 95%. Taking into account the standard errors of the estimated areal proportions, this condition was met at the 0- to 5-cm and 5- to 15-cm depth intervals where proportions of 96 ± 4% and 88 ± 7% were estimated, respectively (Table 2). This means that at least to 15 cm, we are 95% confident the true value of pH falls within the specified PI at

Table 2. Results of the proposed soil map quality indicators: The areal proportion of the map within the specified prediction interval or correctly predicted (APCP) and root mean square error of simulation (RMSES) at each depth increment. Additionally, corresponding measures of accuracy (RMSE), bias (ME) and imprecision (IMP) of the given map taking into account only the quality of the predictions.

Depth, cm	APCP	RMSES	RMSE	ME	IMP
0–5	96 ± 5%	1.0 ± 0.1	0.6 ± 0.1	0.2 ± 0.1	0.6
5–15	88 ± 7%	1.0 ± 0.1	0.7 ± 0.1	0.3 ± 0.1	0.6
15–30	81 ± 8%	1.0 ± 0.1	0.8 ± 0.1	0.2 ± 0.1	0.7
30–60	74 ± 9%	1.3 ± 0.1	1.0 ± 0.1	0.1 ± 0.1	1.0
60–100	81 ± 8%	1.6 ± 0.1	1.1 ± 0.1	0.0 ± 0.1	1.1

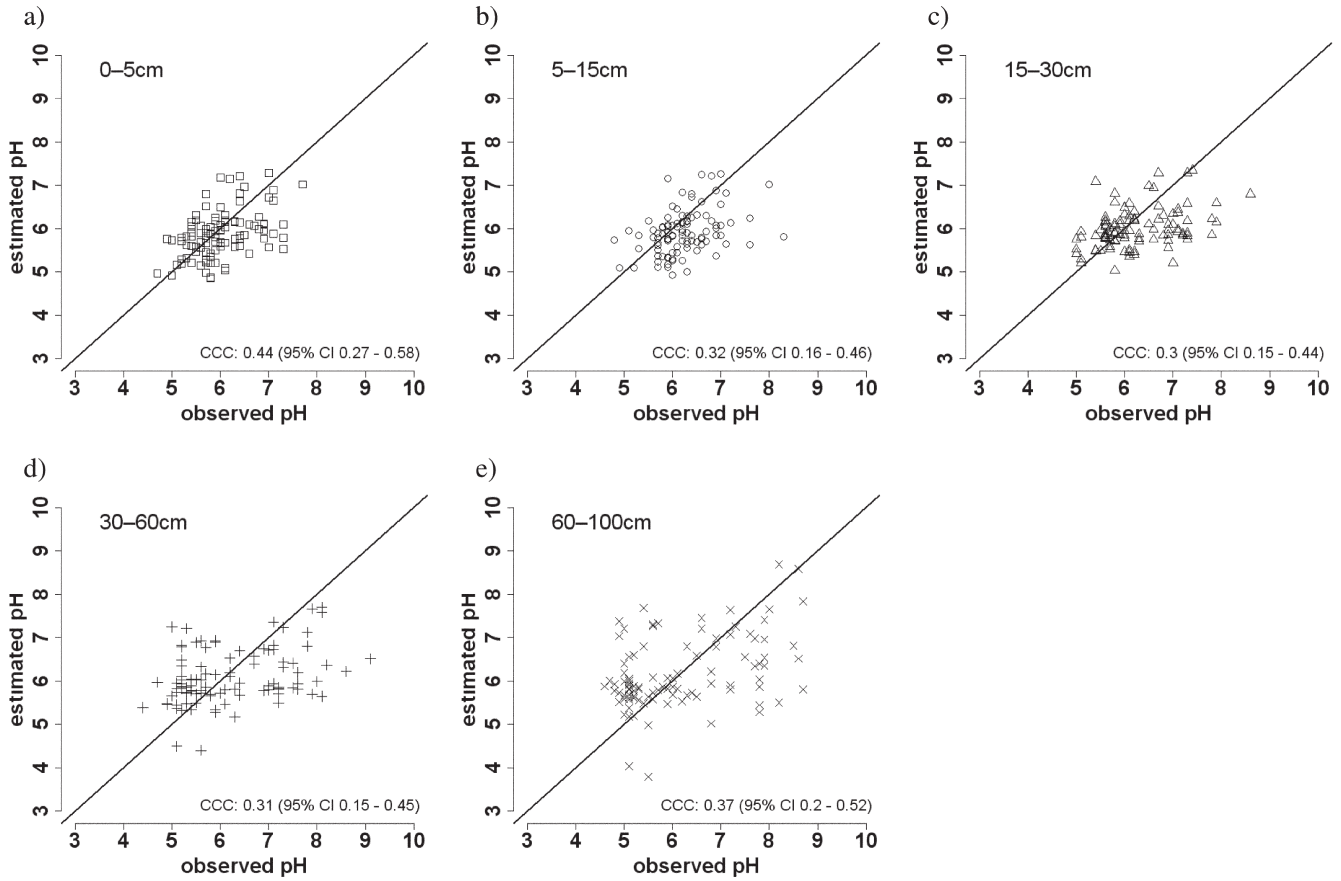

Fig. 4. Prediction interval coverage probability plots for the areal proportion correctly predicted at (a) 0 to 5, (b) 5 to 15, (c) 15 to 30, (d) 30 to 60, and (e) 60 to 100 cm.

each prediction node. The RMSES at these depths was found to be 1.0 ± 0.1 units of pH, indicating that there may be some issues of accuracy. However, relative to the RMSES values from 30 cm, the predictions toward the soil surface are more accurate. This issue of inaccuracy is corroborated with the conventional map quality indicators where it can be seen just at the first two depth increments that there is evidently some positive bias (indicating systematic under prediction) within the predictions. More detail of the conventional indicators is discussed further on.

With increasing depth down the profile an underestimation of the uncertainty is observed. For example, at 15 to 30 cm the APCP indicates that on average 81 ± 8% of map nodes have PIs that cover the true value of pH at that depth interval. At 30 to 60 cm, the APCP decreased to 74 ± 9%, then increased marginally at the 60- to 100-cm depth interval (81 ± 8%). At these three depth increments the RMSES increased from 1 ± 0.1, to 1.3 ± 0.1, and finally 1.6 ± 0.1, respectively. Taken as a whole, the picture that is quite evident is that with increasing soil depth, the accuracy of the predictions decreases along with a growing level of uncertainty that is not accounted for.

Observed vs. fitted plots provide a visual guide of the deviation of the predicted values of pH from the colocated measured values at each depth interval (Fig. 4a-e). Lin's concordance correlation coefficient (CCC) derives a quantitative measure and considers how well the relationship between the measurements (predicted and observed) is represented by a line through the origin at an angle of 45°, as would be generated if the two measurements generated identical results (Lin, 1989). The plots (Fig. 4a-e) show a moderate agreement between the observed and fitted values where CCC ranged between 0.44 and 0.30, with the strongest predictions at 0 to 5 cm (Fig. 4a). Similarly the accuracy at 0 to 5 cm was 0.6 ± 0.1 and gradually decreased with depth to 0.7 ± 0.1 (5–15 cm), 0.8 ± 0.1 (15–30 cm), 1.0 ± 0.1 (30–60 cm), and 1.1 ± 0.1 (60–100 cm).

What can also be observed from the plots is that at higher pHs (>7) there is a systematic under prediction, particular at 15 to 30 cm, 30 to 60 cm, and 60 to 100 cm (Fig. 4c-e). Bias estimates corroborate this observation approximately where strong positive bias (under prediction) was observed for both 5 to 15 cm (0.3) and 15 to 30 cm (0.2 ± 0.1). For the depth increments of 30 to 60 cm and 60 to 100 cm, bias is smaller relative to the other depths; however the low accuracy can be attributed mainly to the higher level of imprecision at these subsoil depths.

For further analysis, mainly in assessing the quality of the uncertainty estimation, Fig. 5a-e shows the deviation of the areal proportion of the map correctly predicted at the corresponding confidence levels for each depth increment. The PICP plots demonstrate a significant degree of sensitivity with change in confidence level. Between the 90 and 40% confidence levels it can be observed that the areal proportions demonstrate a pattern

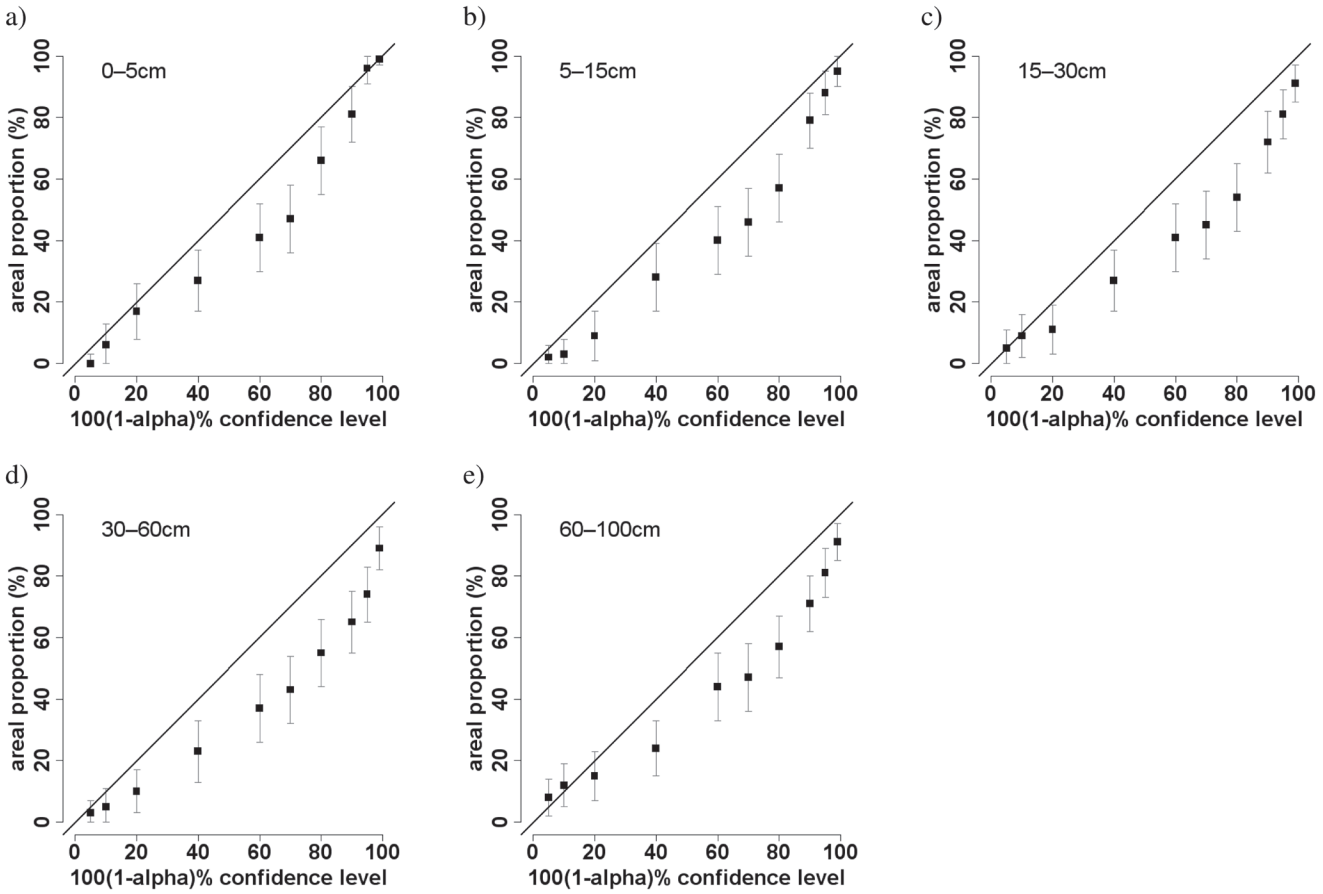

Fig. 5. Plots of the observed soil pH vs. the corresponding digital soil mapped prediction of soil pH and resultant Lin's Concordance Correlation Coefficient (CCC) at (a) 0 to 5, (b) 5 to 15, (c) 15 to 30, (d) 30 to 60, and (e) 60 to 100 cm.

of increasing deviation from the desired result. This outcome is likely attributed to the bias observed within the predictions which has resulted in the bias being transferred to the PIs because they are empirically derived from the prediction errors. At higher confidence levels, the PI performs as it would be expected to, and also has the added benefit of being able to buffer the bias in the predictions. For predictions that have significant bias, it is with decreasing confidence levels that it becomes evident that there is some misspecification of the PIs at these confidence levels. The fact that the observed deviations are below the 1:1 line, indicates not only some misspecification of PIs, but more importantly is that the predictions of uncertainty are underestimated. In this regard this is better than the alternative where the areal proportions for a given confidence level are above the lines which would indicate the PIs are unnecessarily wide or in other words, an overestimation of the uncertainty.

As expressed in Malone et al. (2011), the aim of calculating uncertainties is to account for all perceived or known sources, including those associated with our poor understanding of soil patterns and processes, and those associated with the model and their parameters and model inputs (covariates). Because the approach to quantifying the uncertainties is evaluated empirically (Malone et al., 2011), bias in the predictions will inadvertently mean bias will be present also in the estimates of uncertainty. This phenomenon was apparent in this study; the problem is that we used legacy soil data to generate the map which ultimately resulted in prediction bias. Consequently, this bias is reflected in the uncertainties as well. Because a probability sample was used to validate the pH map and their uncertainties we are able to discover such bias and subsequently an underprediction of uncertainties. It is not always the case that independent data from a probability sample can be used for validation. Data splitting was used in Malone et al. (2011). In that study, at confidence levels from 5 to 99% there was a near matching proportion of observations which fitted within their PI for both available water and organic C. While a better result per se, it is clear to see that indicators of map quality are more valid when using an independent probability sample for validation.

From a map producer's perspective, there is significant value in coupling the proposed criteria with those conventionally reported for soil map quality. First, the APCP, RMSES, and RMSE all indicated an increased uncertainty with soil depth. We are more confident in the quality of the soil map at the soil surface where the predictions themselves are more accurate, but more importantly, the uncertainties of those predictions are also adequately accounted for. With increasing soil depth however, map quality decreases; there is decreased accuracy and precision of the actual predictions, coupled with a systematic underestimation of the uncertainties.

Overall (at all depths) the issues of bias and imprecision need to be addressed to improve the accuracy. Ideally this would be done directly by improvement in the modeling of the spatial distribution of the soil properties, which would result in a decrease of the RMSES. The PI's would naturally adjust themselves accordingly and become narrower in the process.

There are difficulties however with existing methods to make improved predictions within a DSM, particularly in the subsoil. One difficultly is that soil pH is evidently a dynamic soil property, which is likely to change as a result of human intervention such as agronomic practices (Bastida et al., 2008). From field knowledge of the study area, we expect however that significant change in pH as a result of intervention would only be minor. Rather, the difficultly is more the paucity of well known and available covariates that are able to describe soil attribute variability in the subsurface.

A question that then remains is whether other soil properties can be validated using the same sampling units used in this study? In short, it is not optimal to do so. The procedures for validation in this study were optimized based on the predictions and their uncertainties of soil pH. Because of this, the stratification may be inappropriate for other soil attributes (maps) that could be validated in this area. The intention of this project was not to validate multiple soil attributes; it was more the presentation of additional criteria for quantifying map quality. Nevertheless, for the sake of efficiency there may be a requirement that multiple soil attribute maps need to be validated concurrently. One alternative to stratification on the basis of the predictions and their uncertainties is compact geographical stratification (Brus et al., 1999). With this design, the target area is stratified on the basis of spatial coordinates. With such a design, the same design-based inference could be used as that used in this study (de Gruijter et al., 2006). Future studies will obviously need to properly investigate this alternative approach.

CONCLUSIONS

In the course of this paper we first impressed on the need to assess the quality of the quantifications of uncertainty in a DSM framework as one does for assessing the quality of the predictions alone. Subsequently, we then presented two new criteria that collectively address map quality in terms of both the predictions and their uncertainties. These criteria are largely based on the empirical coverage of PIs; our methodology for expressing prediction uncertainty. The MSES explicitly deals with the prediction accuracy in that it is a modification of the MSE, yet includes a measure of the map uncertainty within its formulation. The APCP deals mainly with the quality of the uncertainty component whereby we express with a 95% confidence that the true value of a soil attribute lies between the two interval limits. We used additional samples collected from a probability sample to determine unbiased estimates of these quality measures in addition to conventional quality measures such as MSE, ME, and imprecision. By coupling these two new criteria with conventional measures means more information is gained and for a map producer aids in the efforts to improve precision of the map. For the map user, greater clarity of decision making regarding for example the optimization of inputs (fertilizers etc.), monitoring soil changes or ameliorating soil threats can be made. Regardless of the purported quality, it is up to the map user to determine map's fitness for use. The criteria we have proposed in this study ensure a more objective approach to those decisions.

APPENDIX A
Accounting for Measurement Error in Map Validation

Measurements that are used for validation may be so accurate that one can safely assume that the effects of measurement error on the validation results are negligible. That assumption can not be made in this study, thus we have to consider the effects of measurement error.

The validation criteria used in this study are areal proportion correctly predicted, Root Mean Squared Simulation Error, Root Mean Squared Prediction error, Mean Error, and Imprecision. All of these, except for the Mean Error, involve non-linear transformation of the measured pH: a 0/1 indicator transformation for the APCP and squaring for RMSE and Imprecision. This implies that, without the bias corrections as detailed hereafter, the estimates of these 'non-linear' criteria would be biased. Especially the RMSE would be over-estimated, thus punishing the map for errors in the validation data, and all error variances would be underestimated.

Estimation of the Areal Proportion Correctly Predicted

As defined previously the criterion is defined as:

$$\mathrm{APCP} = \frac{1}{\|A\|} \int_{s \in A} i(s)\,\mathrm{d}s \qquad [A1]$$

where A is the mapped area, and $i(s)$ equals 1 if the true pH value at location s is covered by the PI given by the map at s, and 0 otherwise. As we do not possess true pH values, we have no error-free *i's* either. Instead we have to work with indicator values determined from the measured pH values. These indicator values are thus subject to random error and denoted here by $I(s)$. Each $I(s)$ follows a Bernoulli distribution with expectation $\pi(s)$, being the probability that a randomly measured pH value at s is covered by the PI at s. (The pH measurements are assumed to have no systematic error.)

The usual design-based estimator for APCP is:

$$\widehat{\mathrm{APCP}} = \frac{1}{\|A\|} \sum_{h=1}^{H} \frac{\|A_h\|}{n_h} \sum_{j=1}^{n_h} I_{hj} \qquad [A2]$$

where A_h and n_h are the surface area and sample size of Stratum h, respectively, and I_{hj} is the indicator value as determined at the *j-th* sample point of Stratum h. If the pH measurements were error-free, $I(s)$ would equal $i(s)$ for all s, and $\widehat{\mathrm{APCP}}$ would be an unbiased estimator. This is not so in the present study. To investigate the bias of $\widehat{\mathrm{APCP}}$ we take its expectation over the process of measuring, conditional on a given sample $\boldsymbol{S}$:

$$E_m\left\{\widehat{\mathrm{APCP}}|\boldsymbol{S}\right\}=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}E_m\left\{I_{hj}\right\}$$

$$=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}\pi_{hj} \qquad [A3]$$

The conditional bias due to measurement error thus equals:

$$E_m\left\{\widehat{\mathrm{APCP}}|\boldsymbol{S}\right\}-\widehat{\mathrm{APCP}}_t=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}\left\{\pi_{hj}-i_{hj}\right\} \qquad [A4]$$

where $\widehat{\mathrm{APCP}}_t$ is the (hypothetical) estimate based on error-free measurements. We could assess this bias numerically by simulation, but in the present case we may assume that it is small enough to be neglected. The reason is that the terms $\pi_{hj}-i_{hj}$ tend to be small and, more importantly, being both positive and negative they will largely cancel out.

Estimation of the error variance of $\widehat{\mathrm{APCP}}$

The error variance of $\widehat{\mathrm{APCP}}$ due to sampling and measurement error is estimated by:

$$\hat{V}\left(\mathrm{APCP}\right)=\frac{1}{A^2}\sum_{h=1}^{H}\frac{A_h^2}{n_h\left(n_h-1\right)}\sum_{j=1}^{n_h}\left(i_{hj}-\bar{i}_h\right)^2 \qquad [A5]$$

where $\bar{i}_h$ is the mean of the indicator values in stratum h.

Estimation of the Mean Square Error

The mean squared prediction error is defined as:

$$\mathrm{MSE}=\frac{1}{\|A\|}\int_{s\in A}\left\{z_p\left(s\right)-z(s)\right\}^2\,\mathrm{d}s \qquad [A6]$$

where $z_p(s)$ and $z(s)$ are the predicted and true values at location s of the target variable, here pH. The MSE is estimated from stratified random sample data by:

$$\widehat{\mathrm{MSE}}=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}\left(z_{phj}-z_{hj}\right)^2 \qquad [A7]$$

Applying this estimator to measured values (z_m) instead of true ones gives:

$$\mathrm{MSE}=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}\left(z_{phj}-z_{mhj}\right)^2 \qquad [A8]$$

Assuming that the prediction error and the measurement error are spatially uncorrelated, this can be rewritten as:

$$\widehat{\mathrm{MSE}}=\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}\left\{\left(z_{phj}-z_{hj}\right)^2+\left(z_{mhj}-z_{hj}\right)^2\right\} \qquad [A9]$$

Taking the expectation over both sampling and measuring gives:

$$E_pE_m\widehat{\mathrm{MSE}}=\mathrm{MSE}+\frac{1}{\|A\|}\sum_{h=1}^{H}\frac{\|A_h\|}{n_h}\sum_{j=1}^{n_h}E_pE_m\left(z_{mhj}-z_{hj}\right)^2$$

$$=\mathrm{MSE}+\sigma_m^2 \qquad [A10]$$

where σ_m^2 is the variance of the measurement error. It follows from [A10] that the usual estimate [A8] should be diminished with σ_m^2 to make it unbiased.

Estimation of the Error Variance of $\widehat{\mathrm{MSE}}$

The error variance of $\widehat{\mathrm{MSPE}}$ due to sampling and measurement error is estimated by:

$$\hat{V}\left(\mathrm{MSE}\right)=\frac{1}{A^2}\sum_{h=1}^{H}\frac{A_h^2}{n_h\left(n_h-1\right)}\sum_{j=1}^{n_h}\left(d_{hj}^2-\bar{d}_h^2\right)^2 \qquad [A11]$$

where $\boldsymbol{d}_{\boldsymbol{hj}}^2$ is the squared difference between the predicted and the measured value at location j in Stratum h, and $\bar{d}_h^2$ is the stratum mean of the squared differences.

ACKNOWLEDGMENTS

The authors express appreciation to the landholders in the Lower Hunter Valley whom were happy to participate in our regional soil survey. Thanks also to Michael Short whose technical expertise in soil sampling was invaluable. We thank the Australian Research Council for supporting this research. This work is part of the *GlobalSoilMap.net* effort.

REFERENCES

Bastida, F., A. Zsolnay, T. Hernandez, and C. Garcia. 2008. Past, present and future of soil quality indices: A biological perspective. Geoderma 147:159–171. doi:10.1016/j.geoderma.2008.08.007

Bell, S.A.J. 2004. Vegetation of Werakata National Park, Hunter Valley, New South Wales. Cunninghamia 8:331–347.

Bishop, T.F.A., A.B. McBratney, and B.M. Whelan. 2001. Measuring the quality of digital soil maps using information criteria. Geoderma 103:95–111. doi:10.1016/S0016-7061(01)00071-4

Brown, J.D., and G.B.M. Heuvelink. 2005. Assessing uncertainty propagation through physically based models of soil water flow solute transport. p. 1181–1195. *In* M. Anderson (ed.) Encyclopaedia of hydrological sciences. John Wiley & Sons, Chichester.

Brus, D.J., and J.J. de Gruijter. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80:1–44. doi:10.1016/S0016-7061(97)00072-4

Brus, D.J., L. Spatjens, and J.J. de Gruijter. 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89:129–148. doi:10.1016/S0016-7061(98)00123-2

Brus, D.J., B. Kempen, and G.B.M. Heuvelink. 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. doi:10.1111/j.1365-2389.2011.01364.x

Burrough, P.A., P.H.T. Beckett, and M.G. Jarvis. 1971. Relation between cost and utility in soil survey (I-III). J. Soil Sci. 22:359–394. doi:10.1111/j.1365-2389.1971.tb01624.x

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

Finke, P.A. 2007. Quality assessment of digital soil maps: Producers and users perspectives. p. 523–542. *In* P. Lagacherie et al. (ed.) Digital soil mapping, an introductory perspective. Developments in soil science, Vol. Vol. 31. Elsevier, Amsterdam.

Grunwald, S. 2009. Multi-criteria characterization of recent digital soil mapping and modelling approaches. Geoderma 152:195–207. doi:10.1016/j.geoderma.2009.06.003

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of statistical learning: Data mining, inference and prediction. 2nd ed. Springer, New York.

Lark, R.M. 1995. Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. Int. J. Remote Sens. 16:1461–1480. doi:10.1080/01431169508954488

Lin, L.I. 1989. A concordance correlation-coefficient to evaluate reproducibility. Biometrics 45:255–268. doi:10.2307/2532051

Malone, B.P., A.B. McBratney, B. Minasny, and G.M. Laslett. 2009. Mapping

continuous depth functions of soil carbon storage and available water capacity. Geoderma 154:138–152. doi:10.1016/j.geoderma.2009.10.007

Malone, B.P., A.B. McBratney, and B. Minasny. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160:614–626. doi:10.1016/j.geoderma.2010.11.013

McBratney, A.B., and J.J. de Gruijter. 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. J. Soil Sci. 43:159–175. doi:10.1111/j.1365-2389.1992.tb00127.x

Minasny, B., and A.B. McBratney. 2002. Uncertainty analysis for pedotransfer functions. Eur. J. Soil Sci. 53:417–429. doi:10.1046/j.1365-2389.2002.00452.x

Minasny, B., and A.B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32:1378–1388. doi:10.1016/j.cageo.2005.12.009

Odgers, N. 2010. Bottom-up digital soil mapping. Geoderma.

Odgers, N.P., A.B. McBratney, and B. Minasny. 2008. Generation of kth-order random toposequences. Comput. Geosci. 34:479–490. doi:10.1016/j.cageo.2007.05.018

Rayment, G.E., and F.R. Higginson. 1992. Australian laboratory handbook of soil and water chemical methods. Inkata Press, Melbourne.

Shrestha, D.L., and D.P. Solomatine. 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19:225–235. doi:10.1016/j.neunet.2006.01.012

Solomatine, D.P., and D.L. Shrestha. 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resour. Res. 45:W00B11. doi:10.1029/2008WR006839

Thackway, R., and I.D. Cresswell. 1995. An Interim biogeographic regionalisation for Australia: A framework for establishing the National System of Reserves Australian Nature Conservation Agency, Canberra.

Tranter, G., B. Minasny, and A.B. McBratney. 2010. Estimating pedotransfer function prediction limits using fuzzy k-means with extragrades. Soil Sci. Soc. Am. J. 74:1967–1975. doi:10.2136/sssaj2009.0106

Webster, R., and P.H.T. Beckett. 1968. Quality and usefulness of soil maps. Nature 219:680–682. doi:10.1038/219680a0

White, R.E. 2003. Soils for fine wines. Oxford Univ. Press, New York.