Assessing the thematic accuracy of area-class soil maps

D G Rossiter

PREPRINT prepared on August 31, 2001

Soil Science Division, ITC
P.O. Box 6, 7500AA Enschede
the Netherlands.

phone: +31 53 487-4499
email: `rossiter@itc.nl`
home page: `http://www.itc.nl/~rossiter/`

Contents

List of Tables

Abstract

This paper is meant as a guide for data users who need to quantify the accuracy of area-class maps of soil classes or properties, and for data producers who need to provide quantified accuracy measures in metadata for digital datasets derived from such maps. The thematic accuracy of a soil map is the degree to which the attribute information in a map agrees with reality, and is assessed by statistics of a partial field sample or a complete reference map. Accuracy may be assessed on the basis of legend categories, taxonomic classes, single or multiple soil properties, interpretations, or interpretive classes. Binomial tests may be used to determine the error frequency, and multinomial tests to also identify which classes or properties are confused. Multinomial tests are based on the confusion matrix, and include naïve measures of agreement and the *kappa* and *tau* indices. Partial credit for mapping errors may be given by weighted variants of these tests, with weights being assigned by utility functions or by the producer's subjective judgement of class similarity. A χ^2 test may be used to simultaneously verify the overall accuracy and the agreement with stated proportions of a compound map unit. Field sampling for accuracy assessment may be random or stratified by legend category or major landscape. The required sample size depends on prior estimates of class proportions, required relative accuracy, and acceptable chance of Type I error. A reference map to be used for comprehensive sampling should be discretized to a grid cell size equal to one quarter of a minimum legible delineation.

Keywords

Soils maps; accuracy; multivariate analysis; sampling; legends

1 Thematic accuracy as an aspect of adequacy

Maps of the geographic distribution of soil classes and properties are valuable inputs to environmental studies (e.g. Bouma et al. [7], Lilly et al. [43]), land management plans, land evaluation [21, 22, 56], and land use planning. Soil maps are available at reconaissance to detailed scales over large areas of the world. They are increasingly presented in digital form, incorporated into multi-theme Geographical Information Systems, and made available on the World Wide Web [57]. In this context, their *adequacy*, defined as 'fitness for use', is a key issue.

Forbes et al. [29] classify the adequacy of a soil resource inventory according to four aspects : map scale and texture, map legend, base map quality, and ground truth. This paper deals only with the last aspect, that is, the degree to which the map agrees with reality. This is also called *thematic accuracy*.

Modern metadata standards provide a standardised method to inform data users about accuracy. For example, §2 of the USA's federal government standard [25] is dedicated to 'data quality information', which may include a report on attribute (thematic) accuracy, preferably a quantitative assessment (§2.1.2), defined as "a value assigned to summarize the accuracy of the identification of the entities and assignments of values in the data set and the identification of the test that yielded the value". The relevance to soil maps is obvious; unfortunately these maps rarely provide such information. Digital soil surveys from the USA at both county and state levels [e.g. 70, 69] make no mention of accuracy assessment in this metadata field; instead they state only that "attribute accuracy is tested by manual comparison of the source with hard copy plots ...", that is, the digital dataset is only as good as the original soil survey, which itself has never been tested for accuracy.

Soil maps have been produced over a long period of time, by a wide variety of survey organisations, and with a wide variety of methods, standards, legends, and classifications. Even within one survey organisation, a variety of methods may have been used, and of course each soil landscape has its own mapping problems. Thus soil maps provide special challenges for accuracy assessment, especially in the categorical interpretation of accuracy, that is, what exactly is meant by "agreement" between map and reality.

This paper explains methods of accuracy assessment from the point of view of the soil map user, and provides a guide for selecting and applying appropriate assessment techniques. Producers should also use these methods and to provide objective accuracy measures to their clients, especially as metadata for digital datasets. In addition, it explains methods from the point of view of the map producer, to determine how successful the mapper has been in representing the soil landscape on the map; these can be used for quality control.

There are four general questions that may be asked relating to the thematic accuracy of a map:

1. What is the error *frequency*: how often (i.e., over what proportion of the mapped area) does the map not agree with reality?

2. What is the *nature* of the errors: which soil classes or properties are not mapped correctly, and with which other classes or property values are they confused?

3. What is the *magnitude* of errors: how serious are they for a decision-maker?

4. What is the *source* of the errors: why did they occur?

Here we deal with the first three questions; the map producer can use this information to investigate the fourth.

This paper is structured as follows. First, we discuss conceptual models of spatial variability and limit the further discussion to area-class soil maps (§1.2). Second, we consider the cartographic treatment of boundaries and spot symbols (§1.3). Third, we explain the different possible measures of accuracy (§1.4) and the treatment of legend categories where the map admits that there are contrasting soils in a polygon (§2.2). Fourth, we introduce the statistical assessment of accuracy by simple binomial tests (§3), followed a lengthy discussion of multinomial tests using confusion matrices for simple map units (§4) and validation matrices for compound map units (§5). Fifth, we discuss sampling schemes for accuracy assessment (§6), and finally summarize with a checklist of points to consider when designing an accuracy assessment (§7).

1.1 Background

The systematic study of soil map quality began with Webster [76] and continued with a group at Oxford, although most of their effort was focused on optimum methods of producing soil maps (e.g. Beckett and Burrough [4]), not assessing existing ones. A group at Cornell University worked for several years on aspects of soil survey adequacy, including accuracy assessment [63, 29]. A group at the Staring Centre (now Alterra) in the Netherlands [32, 45] developed methods for quantifying map unit composition and thematic quality. The Soil Science Society of America held a symposium on soil spatial variability [52] where attention was given to this topic. Several authors have reported on transect studies and their statistical treatment [79, 78, 8].

Accuracy assessment has been extensively studied for various kinds of thematic maps, most notably land use and land cover maps produced by satellite remote sensing [16, 17, 59], and a wide variety of statistical techniques have been proposed for this purpose [18, 44, 50, 30].

1.2 Conceptual models of soil spatial variability

There are two conceptual models of the soil landscape that are used in soil mapping: the *discrete* and *continuous* models of spatial variation (DMSV and CMSV, respectively) [35]. In this paper we are only concerned with the DMSV, which hypothesizes that the variation in soil classes (and therefore properties) across the landscape can be partitioned by sharp boundaries into 'homogeneous' areas. In the extreme, all locations within a polygon would be identical, but the DMSV is still successful if the within-polygon variance is less than the between-polygon variance. The DMSV does not account for spatial dependence either within polygons or across boundaries. The geometric representation is a 2-dimensional simplicial complex (if boundaries are made up of line segments) or, more generally, a 2-dimensional cell complex [39]. The usual physical representation in a GIS is as a set of polygons in the so-called 'vector' data model of entities [11, p. 22] with a linked relational database; the map unit name is the key into the legend table. These soil maps may be termed *area-class polygon maps*. A variation is when delineations are known to have several contrasting soils, with stated proportions and perhaps with known landscape relations within the delineation; however, the DMSV can provide no information on the location of these components of such 'compound' map units.

Almost all maps produced to date by soil survey organisations are area-class maps, and it is these maps that are being converted from analog to digital format and integrated into multi-theme GIS. It may be hoped that soil mapping in the future will increasingly use pedometric techniques [47], including continuous class maps [46, 33, 74], which generally provide include internal measures of accuracy as part of their production, but this does not help the situation with existing surveys.

The remainder of this paper deals only with area-class maps. These impose a hierarchical division of the mapped area into classes (legend categories), and then into individual delineations (corresponding to map polygons). Each delineation belongs to exactly one legend class. Therefore each mapped location falls into one delineation, and hence into one legend class. In soil mapping, a legend class may be considered *simple* ('consociation' or 'undifferentiated group'), referring to only one soil type or interpretive group, or *compound* ('association' or 'complex'), referring to several contrasting soil types [64, pp. 27-29]. Simple map units are often called *homogeneous*, but this is misleading since it is quite rare in nature that a delineation contain only one soil type. That is, the natural soil-landscape segments that can be reliabily delineated at reasonable scales on airphotos or in the field by experienced surveyors contain more than one narrowly-defined soil class, and these classes often have contrasting properties and interpretations. For this reason, soil survey organizations have developed standards for map unit *purity*, using the concepts of *similar* and *dissimilar* soils [64, pp. 29-30]. If a map unit is sufficiently pure, it is called 'homogeneous' (or better, 'simple'), even though in reality there are contrasting soils in its mapped delineations. This has important implications for accuracy assessment, as will be discussed in §2.2.

1.3 Geometric accuracy, symbols, and boundaries

Accuracy can be divided into *geometric* and *thematic* aspects [54]; here we deal only with the thematic question: to what degree do the predictions made by the map correspond to reality? In practice, it is difficult to separate these aspects. Any geometric errors in the placement of boundaries or georeference will contribute to apparent thematic error, since ground sampling will be displaced with respect to the map. The map producer can then examine an inaccurate map to see if a geometric adjustment would improve its quality.

Some soil maps show point ('spot') or line symbols to represent areas that are too small or thin to be drawn as polygons at the map scale [64, p. 236]. The soil class is either given directly in the legend or can be inferred from the symbol type (e.g. wetness, stoniness) and adjacent polygons. For purposes of accuracy assessment, spot symbols may be considered to be circles of the size of the MLD (see §6.4) centered on the point. Line symbols may be considered to be polygons with rounded ends of a specified width, converted to ground units, centered on the line. The width corresponds to the minimum delineation width allowed by the cartographic standards of the mapping agency. For example, in the Netherlands this width is 3mm on the map [36]. In this way, spot and line symbols become polygons for the purposes of accuracy assessment.

Boundaries between delineations of area-class soil maps are modelled as having no width, both in the conceptual DMSV and in the vector GIS representation. In reality they represent zones at least as wide as the maximum location accuracy, and at most as wide as the narrowest mappable delineation, at map scale. Conceptually, these may be considered either as belonging to both bounding polygons or as belong to neither. These zones may be of several types: (1) a transition between related soils (as in a hillslope catena or drainage sequence towards a waterway); (2) a zone of uncertainty between unrelated soils (as in the boundary between hilland and terrace); (3) a different soil body that can not be mapped at this scale (as in a small colluvial footslope between hilland and terrace). Lagacherie et al. [41] present an interesting classification of the nature of soil boundaries.

The width of the zone of uncertainty depends on the soil bodies on either side, and may be set by the mapper; however it can not be any wider than minimum width allowed by the cartographic standards of the mapping agency nor any narrower than the maximum location accuracy, which also depends on the mapping agency's standards; this latter is typically an order of magnitude smaller than the former, e.g. 0.3mm vs. 3mm.

For purposes of accuracy assessment, there are four ways to deal with boundaries, in descending order of strictness. First, the boundary can indeed be considered to have no width, and a sample point falling near the boundary is compared to its mapped class. Second, a sample point within the zone of uncertainty may be compared to the classes on either side of the boundary, and considered to be in agreement with the map if it matches either one. Third, a sample point in the zone of uncertainty may be assigned to either of the bounding polygons at the discretion of the assessor. Finally, the zone of uncertainty around each boundary can be excluded from the accuracy assessment. The second method seems to provide a good compromise which recognizes the cartographic uncertainty.

1.4 Types of thematic accuracy

Intuitively, the ground truth of a soil map is the proportion of the mapped area in which the soil class, property, or behaviour as predicted by the map is in fact encountered. The difficulty is in making this idea precise and operational.

We distinguish five kinds of thematic accuracy: (1) legend, (2) taxonomic, (3) attribute, (4) interpretive, and (5) interpretive classification.

The *legend* accuracy is a comparison of actual with mapped legend classes. It measures how well the soil map predicts the legend categories on the landscape. If the classes are from a soil classification system such as the World Reference Base [24, 20] or Soil Taxonomy [65], this is also called *taxonomic* accuracy. In a hierarchical classification system such as these just mentioned, taxonomic accuracy can be assessed at any level; presumably the accuracy at the higher (more general) levels will be greater than those at the lower (more specific) ones. If the legend classes include taxonomic names but also other information, for example phases, taxonomic accuracy is less demanding than legend accuracy.

The *attribute* accuracy is a comparison of actual with mapped values or classes for a selected set of soil properties (as opposed to legend or taxonomic clases). Typically, the selected properties are important for land evaluation [56]. These attributes may be used as input to some model of soil performance (e.g., rating tables or process models), and the observed and predicted performances compared; this is *interpretive* accuracy. In practice, it is difficult to determine interpretive accuracy in this way, because performance for a given use is not generally observed over the entire mapped area. An exception in some areas is the yield of a widely-grown crop under a single management system [23]. Therefore, interpretive accuracy is often evaluated by means of the attribute accuracy for a set of diagnostic land characteristics that are considered important to performance.

Taxonomic and interpretive assessments may be combined by evaluating the agreement with an *interpretive classification*, that is, a re-classification of the soil types for a specific purpose, e.g. hydrological response [6] or management group [58, 60, 40]. This has the advantage, for the map user, that soils are grouped into classes that have been designed to relate directly to use.

2 What is 'agreement' between map and ground?

The previous section used the term 'agreement' without defining what this means and how it is determined in practice. In this section we investigate this question for the different kinds of accuracy.

The concept of 'agreement' between mapped and ground classes is clear for a single or multiple attributes, as long as the attribute classes completely cover feature space. Then a single ground observation matches exactly one of the classes, with an uncertainty equal to the measurement error.

For taxonomic classes, this concept is not so clear, because the set of classes in the legend is in general not complete, i.e. does not classify all possible pedons, for two reasons. First, there may be minor areas of defined taxa that are not included in this survey's legend. These may be explicitly correlated with taxa that are in the legend, or they may not have been recognized during mapping. Second, there may be pedons which do not fit in any existing taxonomic class. Modern soil classification systems [e.g. 24, 65, 1, 61, 2] usually classify all soils at higher categorical levels by means of a 'miscellaneous' or 'other' class, but not at the detailed levels used in most mapping. This is especially true at the series level.

There are several ways to deal with an observation that does not match any legend category for purposes of accuracy assessment:

1. Assign it to the legend category which it most closely resembles.

   This can be pre-defined for defined taxa which were explicitly excluded from this area's mapping legend, i.e. 'correlated out' of the survey. Any pedons of these classes are considered correctly mapped if they occur in map units of their correlated class. In general the interpretations for use of these 'mis-classifications' are almost the same as the taxa to which they are correlated.

   For pedons that do not fall into any pre-defined class, the assignment to a legend class requires expert judgement. The guiding principle is that the interpretations be almost the same.

2. Assign it to a separate 'other' class, which is treated as a mapping failure.

3. As (1), but if the point is too dissimilar to all classes in the legend, then as (2). The decision as to what is 'too dissimilar' is usually made based on interpretations.

4. Assign the point to more than one class, with different degrees of agreement. Then the accuracy must be assessed with weighted methods (§4.2.3).

For interpretive accuracy, we could require that the map and ground agree exactly, but this does not distinguish between small mistakes that don't seriously affect the use and gross errors. Forbes et al. [29] proposed a measure of interpretive accuracy based on a series of attribute accuracies for classified attributes. Each classified attribute is measured either on an *ordinal* scale, for example depth to root-limiting layers or structure grade, or on a *nominal* scale, for example landform or structure type. Ordinal attributes have an implicit concept of 'adjacency' in attribute space, whereas for nominal attributes, the evaluator must define an *adjacency matrix* which states which class pairs are considered 'adjacent'. Each observation is scored on a scale of 1 to 4:

1. All attributes are in the mapped attribute classes;

2. Less than some specified proportion of the attributes are in adjacent classes, and none are in non-adjacent classes;

3. More than the specified proportion of the attributes are in adjacent classes, or exactly one attribute is in a non-adjacent class, or both;

4. More than one attribute is in a non-adjacent class.

Accuracy is then assessed with respect to a threshold defined by the evaluator. A strict assessment would consider only observations with score 1 to be 'true'; a weaker assessment would consider observations with either scores 1 or 2 to be 'true', and the weakest assessment would consider any observation except 4 to be 'true'.

For accuracy based on an interpretive classification, a strict assessment is preferable, since the process of assigning pedons to interpretive classes has already simplified the taxonomic legend.

For any type of accuracy, but especially taxonomic, there is the complication that the map units are of several types, that is, they make different statements about the number and relation of soil types within a legend category. We consider first the case when the map predicts that only one type of soil is present (perhaps with some allowable proportion of impurities), and second the case when the map predicts several contrasting soils in one map unit.

2.1 Simple map units

Simple map units, usually called 'homogeneous', might better be called 'homogeneous enough for detailed interpretations'. The most common type of unit is the *consociation*, which contains one dominant soil type, perhaps with some other soils that are consistenly associated with the dominant soil and have similar interpretations. The map unit has some allowable level of *impurities*, i.e. contrasting soils. Consociations may be recognized in the legend by the form of their name [64, 72]: they have only one named soil at any categorical level, e.g. 'Cazenovia silt loams' or 'Lithosols'.

If the map unit contains several similar soil types which are not consistently associated geographically, they may be mapped as an *undifferentiated group*. The legend entry for these is a list of the possible similar soils, separated by 'and', e.g. 'Cazenovia and Cayuga silt loams'. This means that in any delineation one may encounter either or both of the named soils.

Field experience shows that simple map units almost always contain pedons of several classes. The Soil Survey Manual [64] acknowledges that "[i]n all soil surveys, virtually every delineation of a map unit includes areas of soil components . . . that are not identified in the name of the map unit". The descriptive legend should acknowledge this reality, and give an estimate of the composition of each map unit, including the estimated proportion of each *inclusion*, and whether it is similar or dis-similar to the named soil class.

For example, a typical soil survey report from the USA [71] includes the following statement about the 'Tioga silt loam' consociation:

> "Included with this soil in mapping are areas of moderately well drained . . . Middlebury soils on the lower part of the flood plain. Also included are soils that have more silt in the subsoil than this Tioga soil and that do not have underlying sand or gravel. Some areas have a surface layer of gravelly silt loam."

The first inclusion is another named soil which is limiting for most uses, with respect to the dominant soil in the map unit. The second and third are similar soils but not within the range of the taxon. The mapper has recongized and acknowledged all three impurities, but could not map them. Implicit in this text is that the inclusions occupy less than proportional areas specified by the survey organisation's purity standards for simple ('homogeneous') map units. In this case, the relevant standard is the USDA's concept of 'consociation' [64, 72], which requires $\geq 75\%$ in the taxon, similar taxa, or their taxadjuncts, $<15\%$ dis-similar, limiting inclusion with respect to the taxon, and $<10\%$ of any single dis-similar, limiting inclusion.

Purity standards are a promise from the map maker to the map user, therefore they must be validated.

2.2 Compound map units

These are map units which, according to the mapper, contain several contrasting soil types which could not be separately mapped at the chosen scale. The most common is the *association*, which has a regular and predictable landscape relation between its components, and could be mapped at a larger scale. These are given names such as 'Ontario-Cazenovia-Lakemont association'. The legend should give the relation between the components which allows them to be recognized in the field by the map user; this is the so-called 'two-stage' survey of Beckett [3]. The other common type is the *complex*, where the contrasting soils could not be mapped at a realistic scale, either because the landscape relations are unpredictable or because, while predictable, they occur at too fine a scale. The limiting scale has been conventionally defined by the USDA as 1:24 000 [64, Ch. 2], but in view of increasingly-detailed mapping projects, this limit should be raised to 1:12 000 or even 1:6 000. These are given names such as ‘Farmington-rock outcrop complex’. The legend should give an estimate of the proportion of each component, and, if possible, the landscape relations. In both cases, it is expected that the reported proportion of components applies to each delineation separately, as well as to the entire map unit.

A less useful, and less common, type of compound map unit is a list of *unassociated soils*. Delineations of these contain any one or more of the several soil names joined by ‘or’ in the map unit name, e.g. ‘Cazenovia or Cayuga soils’. Any statement on the proportion of components applies to the entire map unit, not to individual delineations.

A compound map unit without stated proportions of its constituents is not very useful. Its accuracy must be assessed by a very lenient standard: any observation that matches any of the named classes is counted as a success. So the map may be ‘accurate’ but is not adequate for most uses.

The map legend commonly states the relative proportion for the various named components of each map unit. Transect sampling [32] or detailed photo-interpretation of sample areas may have been used to establish these statements, but in general they are usually based on the surveyor’s impression of the landscape during mapping. Still, they are what is stated to the user. No matter how the estimates were made, the map can be considered accurate only if both the soil names and their proportion are correct; a method to assess this is given in §5.

Similar inclusions are grouped with their corresponding component for this purpose. That is, a statement such as ‘60% A, 40% B’ is to be read as ‘60% A and similar, 40% B and similar’.

3 Binomial tests of accuracy

The simplest method of accuracy assessment is a binomial test based on the number of agreements (‘true’, ‘success’) or disagreement (‘false’, ‘failure’) of a series of sample points with the mapped class, attribute, or interpretation. Confidence intervals can be determined for the accuracy of the whole map or of any stratum, for example a single legend category. This method only answers the first question posed in §1, namely ‘What is the error frequency?’

An important concept in both binomial and multinomial tests is the *null hypothesis* H_0, which here is that the true or population statistic, e.g. the overall map accuracy, is within a range that we estimate from the sample. We fix a *confidence level* $(1-\alpha)$ for this hypothesis, and this in turn allows us to compute the range. The parameter α is the acceptable probability of a *Type I* error, that is, the chance that we are rejecting the null hypothesis when it is in fact true. The lower we set α, the narrower the computed confidence intervals, but the less certain we are that the (unknown) population value is in fact in this interval; in addition, a larger sample is required to establish a given precision (§6.3). In soil survey work, α is often set to 0.1 or even 0.2, as compared to the usual practice in agricultural experimentation, where it is set to 0.05 or even 0.01 to detect significant effects.

The statistics for the binomial test are covered in many sources [e.g. 26, 66]. They have been used in transect studies to estimate the composition or purity of map units [32, 79] but serve equally well to summarize the accuracy of an entire sample for map accuracy assessment [29]. For example, Brannon and Hajek [8] used binomial confidence limits to verify a recompiled soil survey in Alabama, USA with 98 locations over 62 000ha, each with three sampled pedons, in a random sample stratified by map unit. They distinguished between taxonomic accuracy, in their case 79%, and interpretive accuracy, in their case 90 to 98%; 90%-confidence itervals were about 4% wide.

Given an unbiased sample of size n, with n_1 successes, the proportional accuracy $\hat{p}$ and its standard deviation $\hat{s}$ are estimated as:

$$\hat{p} = n_1/n \tag{1}$$

$$\hat{s} = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} \tag{2}$$

The confidence interval of the estimate is then:

$$\hat{p} \pm \left[\hat{s}\cdot Z_{1-\alpha} + \frac{1}{2n}\right] \tag{3}$$

where $Z_{1-\alpha}$ is the two-tailed normal score for the probability of Type I error α. The small-sample correction $1/2n$ is usually ignored for sample sizes $n > 50$; it is then less than 1%. The lower and upper limits as computed by Equation 3 are truncated at 0 and 1, respectively, if necessary.

As an example, we summarize the confusion matrix of Table 3 (below); of the total observations $n = 163$, the diagonals, representing the correctly classified observations, sum to $n_1 = 86$. Then we calculate $\hat{p} = 0.5276$ and $\hat{s} = 0.0391$. For a 5% probability of a Type I error, the corresponding area under the normal curve is $\mathrm{Pr} = 0.95$, which is obtained for the two-tailed test with $Z = 1.96$. Then the 95% confidence interval for $\hat{p}$ is $[0.4479\ldots0.6073]$. This is generally interpreted to mean that if we had repeated the same sampling scheme a large number of times, we expect that in 95% of these experiments the observed accuracy would be somewhere between 44.8% and 60.7%. We are running a 5% risk that the true proportion is $<44.8\%$ or $>60.7\%$.

We can narrow the confidence interval at the expense of a higher risk of Type I error. For example, increasing α to 0.1, we obtain $Z = 1.64$ and an interval for $\hat{p}$ of $[0.4602\ldots0.5950]$, i.e. about 2.5% narrower. But now there is a 10% risk that the true proportion is outside this interval. Increasing α to 0.2, i.e. a one in five chance of the true value being outside our calculated interval, we obtain $Z = 1.28$ and an interval for $\hat{p}$ of $[0.4744\ldots0.5808]$, now 5.3% narrower.

3.1 Comparing the relative accuracy of two maps

The confidence interval allows us to compare two maps, for example made by different mappers, at different scales, or at different times. If the intervals do not overlap, we can assert that the maps have different accuracies.

If we assume that the two samples come from a single underlying population, we can pool the estimated variances to calculate the test Z statistic for the difference between the two measures.

$$Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{s}_1^2 + \hat{s}_2^2}} \tag{4}$$

This Z-score is the ordinate on the cumulative normal distribution, i.e. the number of standard deviations from the mean. The tabulated value is then the one-tailed probability that the difference between the classifications is due to chance. In general the test is two-tailed, since neither classification is preferred *a priori*, so the probability obtained from the table should be doubled.

Continuing the numerical example of the previous section, suppose there were a second map that appears to be more accurate, with $\hat{p}_2 = 0.65$ but with the standard deviation of the estimate somewhat higher, $\hat{s}_2 = 0.045$. The 95% confidence interval for this second classification computed by Equation 3 is [0.5618 ... 0.7382]. The two intervals substantially overlap, so the two maps could not be considered to have different accuracies by this comparaison. However, the Z-score for the difference between the classifications would be, according to Equation 4:

$$Z = \frac{|0.5276 - 0.65|}{\sqrt{(0.0391)^2 + (0.045)^2}} = 2.0532$$

The one-tailed probability corresponding to a Z-score of 2.0532 is ≈ 0.020, so the two-tailed probability that the second map would in fact be better than the first is $\approx (1-(2 \cdot 0.020)) = 96.0\%$. That is, if we decide that it is justified to pool the variances, the test suggests that the second map would in fact be better than the first, with $\mathrm{Pr} = 0.960$, whereas comparing the 95% confidence intervals would not establish this difference.

4 Multinomial tests of accuracy

Binomial tests treat errors for all classes equally, and therefore can only estimate the accuracy of the map or stratum as whole. To investigate the errors associated with individual classes, *multinomial* tests must be used. These answer the first two questions posed in §1, namely 'What is the error frequency?' (as with binomial tests) and also 'What is the *nature* of the errors: which classes or properties are not mapped correctly?' In this section we consider multinomial tests where each legend category to be verified is simple (§2.1); in §5 we consider how these methods may be used to verify compound map units (§2.2) and estimate the proportions of their consituents.

Multinomial tests are based on the so-called *confusion matrix*. This is a table with columns representing the *reference* (observed) classes, and the rows the *classified* (mapped) classes. Each cell in the matrix contains the number of observations in the mapped class of its row that were in fact observed in the class of its column. Thus the diagonals represent agreement between map and ground (as in binomial tests), and off-diagonals represent different mis-classifications. Numerous authors [17, 42, 59, 31] explain confusion matrices.

Table 1 shows the basic notation of confusion matrices, and Table 2 shows the simple statistics that are computed from it.

Symbol	Meaning	Computation
$\mathbf{X}$	confusion matrix	
	$\rightarrow$ rows $[1 \ldots r]$ are *classified* ("mapped") data	
	$\rightarrow$ columns $[1 \ldots r]$ are *reference* ("true") data	
r	number of rows and columns of $\mathbf{X}$	
x_{ij}	number of observations in row i, column j, i.e. in reference class j but mapped as class i	as observed
x_{i+}	marginal sum of row (mapped class) i	$\sum_{j=1}^{r} x_{ij}$
x_{+j}	marginal sum of column (reference class) j	$\sum_{i=1}^{r} x_{ij}$
n	total number of observations	$\sum_{i=1}^{r} \sum_{j=1}^{r} x_{ij}$
	… or	$\sum_{i=1}^{r} x_{i+}$
	… or	$\sum_{j=1}^{r} x_{+j}$
$\mathbf{P}$	proportions matrix	$\mathbf{X}\ ./n$
p_{ij}	proportion of observations in row i, column j	x_{ij}/n
p_{i+}	proportion of mapped data in row (class) i	x_{i+}/n
	… or	$\sum_{j=1}^{r} p_{ij}$
p_{+j}	proportion of reference data in column (class) j	x_{+j}/n
	… or	$\sum_{i=1}^{r} p_{ij}$

Table 1: Notation for the confusion matrix

Symbol	Statistic	Computation
C_i	User's 'accuracy', mapped class i	x_{ii}/x_{i+}
	… or	p_{ii}/p_{i+}
$\bar{C}_i$	Errors of commission, mapped class i	$1 - C_i$
O_j	Producer's 'reliability', reference class j	x_{jj}/x_{+j}
	… or	p_{jj}/p_{+j}
$\bar{O}_j$	Errors of omission, reference class j	$1 - O_j$
A_o	Overall accuracy	$\sum_{i=1}^{r} x_{ii}/n$
	… or	$\sum_{i=1}^{r} p_{ii}$
$\bar{A}_o$	Overall error	$1 - A_o$

Table 2: Statistics computed from the confusion matrix

The traditional terminology for errors is somewhat confusing, but well-established. It is expressed from the point of view of the mapper. Looking across the rows (classes as mapped), an error of *commission* is said to occur when the mapper incorrectly mapped this class at a reference (ground truth) site where it does not exist. That is, the mapper 'committed' the error of over-mapping a class. This leads to a lower user's 'accuracy' C_i. Conversely, looking down the columns (classes as found in the field), an error of *omission* is said to occur when the mapper failed to correctly identify this reference site in its true class. That is, the mapper 'omitted' to map the site in the correct class. This leads to a lower producer's 'reliability' O_j. The overall accuracy A_o is the same as the estimated binomial accuracy $\hat{p}$ from §3, considering all off-diagonals together as classification failures.

These matrices may be evaluated with various statistics, which we now explain: (1)

naïve measures of agreement; (2) *kappa*; and (3) *tau*. These can all be modified to give partial credit for mis-classifications.

4.1 Naïve measures of agreement

The statistics from Table 2 may be used directly as measures of overall accuracy (A_o) or per-class accuracy from the user's or producer's point of view (C_i or O_j, respectively). Individual off-diagonals $p_{ij}, i \neq j$ show the frequency with which mapped class i was in fact class j on the ground. These statistics are easy to interpret. The standard deviation of the overall accuracy is computed as in the binomial test (Equation 2) with the proportion of successes $\hat{p}$ in that equation replaced by the overall accuracy A_o.

The standard deviation of the per-class user's accuracy is calculated similarly, but with the number of successes and the total number of trials limited to a single row:

$$\hat{p}_i = x_{ii}/x_{i+} \tag{5}$$

$$\hat{s}_i = \sqrt{\frac{\hat{p}_i \cdot (1-\hat{p}_i)}{x_{i+}}} \tag{6}$$

That is, $\hat{p}_i = C_i$. The confidence interval for $\hat{p}_i$ is calculated as in Equation 3, except that the finite-population correction refers to the number of observations in the row:

$$\hat{p}_i \pm \left[\hat{s}_i \cdot Z_{1-\alpha} + \frac{1}{2x_{i+}}\right] \tag{7}$$

The per-class producer's reliability is computed similarly, using the column totals (x_{+j}) instead of row totals (x_{i+}).

The problem with these measures is that they do not account for the fact that some of the apparent classification accuracy could be due to chance, and thereby over-estimate the mapper's skill. This is where the *kappa* (§4.2) and *tau* (§4.3) statistics are applicable.

A numerical example

Table 3 is a sample confusion matrix also used by Congalton et al. [18] and Skidmore [59]. The counts from the confusion matrix $\mathbf{X}$ are shown in roman type, and the proportions from the corresponding proportions matrix $\mathbf{P}$ are shown in in italics. Table 4 shows per-class user's accuracies and producer's reliabilities, along with their standard deviations and confidence intervals.

We can examine the matrix and statistics to gain insight on the magnitude and type of errors. Clearly, some classes are better mapped than others. From the map user's point of view, classes A, B, and C are all mapped with an accuracy of about 60%; however, because of the small number of reference sites located in map unit B, its confidence interval is quite wide. From the map producer's point of view, only classes A and C were reliabily mapped. The producer often mis-mapped class B as one of these two.

4.1.1 The weighted confusion matrix

Naïve accuracy statistics may also be computed for a *weighted* confusion matrix, which gives partial credit for incorect classifications. This is attractive for assessing soil sur-

		Reference Class				Total
		A	B	C	D	p_{i+}
	A	35	14	11	1	61
		0.2147	*0.0859*	*0.0675*	*0.0061*	*0.3742*
	B	4	11	3	0	18
Mapped		*0.0245*	*0.0675*	*0.0184*	*0.0000*	*0.1104*
Class	C	12	9	38	4	63
		0.0736	*0.0552*	*0.2331*	*0.0245*	*0.3865*
	D	2	5	12	2	21
		0.0123	*0.0307*	*0.0736*	*0.0123*	*0.1288*
	Total	53	39	64	7	163
	p_{+j}	*0.3252*	*0.2393*	*0.3926*	*0.0429*	*1.0000*

Table 3: Example of a confusion matrix

	User's accuracy			Producer's reliability		
Class	C_i	$\hat{s}_i$	95% C.I.	O_j	$\hat{s}_j$	95% C.I.
A	0.5738	0.0633	0.4415 . . . 0.7061	0.6604	0.0651	0.5234 . . . 0.7973
B	0.6111	0.1149	0.3581 . . . 0.8641	0.2821	0.0721	0.1280 . . . 0.4361
C	0.6032	0.0616	0.4744 . . . 0.7319	0.5938	0.0614	0.4656 . . . 0.7219
D	0.0952	0.0641	0.0000 . . . 0.2446	0.2857	0.1707	0.0000 . . . 0.6918

Overall accuracy: 0.5276; $\hat{s}$: 0.0391; 95% C.I.: 0.4479 . . . 0.6073

Table 4: Naïve statistics for the example confusion matrix

veys, because it is well-known that not all mapping mistakes are equally serious for the map user. This idea was first developed by Cohen [14] as a modification for the *kappa* statistic, but may be applied directly to naïve statistics.

This method requires that the analyst make a second matrix containing the weights for each cell in the confusion matrix. The diagonals must be 1, and the off-diagonals must be in the range $[0 \ldots 1]$. A value of 0 indicates that there is no credit for this mistake, and as the values increase towards 1, the mistake is considered decreasingly serious. A value of 1 means that the two classes are considered identical for accuracy assessment. There is no need for the weights matrix to be symmetric, since an error of commission may be considered more or less serious than an error of omission. If all off-diagonals are 0, this reduces to the unweighted case. In this section we explain the computations and interpretation of results; in the following section we discuss the more interesting question of how to assign weights.

The additional notation and statistics for the weighted confusion matrix are shown in Table 5. The statistics $\overline{w}_{i+}$ and $\overline{w}_{+j}$ are the weights in the mapped (rows) or reference (columns) class, themselves weighted by the proportion of each class in the mapped or reference population, respectively, i.e. from $\mathbf{P}^{\mathrm{T}}$.

Symbol	**Meaning**	**Computation**
$\mathbf{W}$	weights matrix	
	$w_{ii} = 1, \forall i$	$\equiv 1$
	$w_{ij} \in [0 \ldots 1], \forall i \neq j$	assigned by evaluator
$\overline{w}_{i+}$	weighted average of the weights of row i	$\sum_{j=1}^{r} w_{ij} \cdot p_{+j}$
$\overline{w}_{+j}$	weighted average of the weights of column j	$\sum_{i=1}^{r} w_{ij} \cdot p_{i+}$

Table 5: Notation for the weighted confusion matrix

The sums of weighted averages ranges from $[1 \ldots r]$, with 1 representing the unweighted case, and r (the number of classes) representing the (meaningless) case where no misclassification is important. The weighted overall agreement and per-class accuracies are computed as:

$$A_{ow} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} \cdot p_{ij} \tag{8}$$

$$C_{iw} = \frac{1}{p_{i+}} \cdot \sum_{j=1}^{r} w_{ij} \cdot p_{ij} \tag{9}$$

$$O_{jw} = \frac{1}{p_{+j}} \cdot \sum_{i=1}^{r} w_{ij} \cdot p_{ij} \tag{10}$$

If the off-diagonals are zero, these reduce to the formulas from Table 2.

The standard deviations and confidence intervals for the overall accuracy are calculated with Equations 2 and 3, respectively, as in the unweighted case; however, the proportion of successes $\hat{p}$ in those equations is not limited to the diagonals (as in Equation 1), but instead includes the weighted contribution from the off-diagonals, i.e. it is the overall weighted accuracy A_{ow} from Equation 8, above.

Similarly, the standard deviations and confidence intervals for the per-class user's accuracy are calculated with Equations 6 and 7, with the proportion of successes $\hat{p_i}$ in those equations replaced by the per-class weighted user's accuracy C_{iw} from Equation 9, above.

A numerical example for the weighted confusion matrix

This example, shown in Table 6 uses the confusion matrix of Table 3, and in addition a hypothetical (arbitrary) 4x4 weights matrix. The confusion matrix $\mathbf{X}$ is presented here as the proportions matrix $\mathbf{P}$, since these are used in the formulas of the previous section.

	Reference Class						
	A	B	C	D	p_{i+}	$\overline{w}_{i+}$	C_{iw}
A	0.2147	0.0859	0.0675	0.0061	0.3742	0.6312	0.7110
	1	*0*	*0.67*	*1*			
B	0.0245	0.0675	0.0184	0.0000	0.1104	0.2393	0.6111
Mapped	*0*	*1*	*0*	*0*			
Class C	0.0736	0.0552	0.2331	0.0245	0.3865	0.7607	0.8571
	1	*0*	*1*	*1*			
D	0.0123	0.0307	0.0736	0.0123	0.1288	0.5783	0.5305
	0.91	*0*	*0.61*	*1*			
p_{+j}	0.3252	0.2393	0.3926	0.0429	1	2.5938	
$\overline{w}_{+j}$	0.9880	0.1104	0.7158	0.8896	2.2095		
$O_{j\,w}$	0.9211	0.2821	0.8233	1.0000			

Overall weighted accuracy A_{ow}: 0.7332

Table 6: Example weighted confusion matrix

In this example, $r = 4$, which is then the maximum possible for the sums of weighted averages. The producer's sum is ≈ 2.2 and the user's ≈ 2.6, indicating that many classes in the map are similar, especially from the user's point of view.

The weighted per-class user's accuracies range from moderate ($\approx 53\%$ for class D) to high ($\approx 86\%$ for class C). Producer's reliabilities range from very low ($\approx 28\%$ for class B) to perfect (for class D). In most cases, these are significantly higher than the corresponding unweighted figures from Table 4. The exception is class B, which is so unlike the other classes that partial credit was not given; in this case the weighted and unweighed figures are identical.

The overall utility of the map to the user is measured by the overall weighted accuracy A_{ow}. In this example it is quite high, $\approx 73\%$, as compared to the unweighted accuracy $A_o \approx 53\%$.

Per-class weighted accuracies, their standard deviations, and confidence intervals for this example are shown in Table 7, which may be compared to the unweighted equivalents in Table 4.

The standard deviations for the weighted case are in general lower, because the accuracies are in general further from 0.5. Only for Class D's user's accuracy, where the

Class	User's accuracy C_{iw}	$\hat{s}_{iw}$	95% C.I.	Producer's reliability O_{jw}	$\hat{s}_{jw}$	95% C.I.
A	0.7110	0.0580	0.5890 … 0.8329	0.9211	0.0370	0.8391 … 1.0000
B	0.6111	0.1149	0.3581 … 0.8641	0.2821	0.0721	0.1280 … 0.4361
C	0.8571	0.0441	0.7628 … 0.9515	0.8233	0.0477	0.7220 … 0.9245
D	0.5305	0.1089	0.2932 … 0.7677	1.0000	0.0000	0.9286 … 1.0000

Overall accuracy: 0.7332; $\hat{s}$: 0.03464; 95% C.I.: $0.6622 \ldots 0.8042$

Table 7: Naïve statistics for the example confusion and weight matrices

weighted accuracy increases to near 0.5, is the standard deviation higher.

4.1.2 Assigning weights

The key question, of course, is how to assign the weights of $\mathbf{W}$. For accuracy assessment, these should be assigned from the user's point of view, that is, how serious is the mistake to the user? This can be assigned in two ways, which we consider in turn: (1) with an objective utility function, for accuracy assessment and (2) with the producer's judgement of the similarity of the classes, for quality control.

Assigning weights with a utility function Næsset [51] suggests the use of *utility functions* to determine weights. The weight assigned to a mistaken classification is then the user's utility of the incorrect choice as a proportion of the utility of the correct choice. In a soil survey context, this can be viewed as the proportional economic value of using one soil (the reference class) as if it were in fact the soil as mapped. The weights are assigned within rows, from the map user's point of view of the seriousness of each error of commission.

Utility functions are any assignment to each map unit of the relative value to the user of that map unit, compared to a standard. For example, a simple utility function for a soil survey which is mainly intended to support agricultural planning is the proportional yield, compared to a reference map unit, for an important crop in the survey area. For example, Table 8 shows the predicted yields of groundnuts under prevailing management for a selection of mapping units from the soil survey of Edgecombe County, North Carolina, USA [70].

The sequence Ly–GoA–NoA–{NoB or WaB}–{NoC or WaC or GyC}–GyD is typical for a transect on the Sunderland surface of the middle Atlantic coastal plain from the margins of undissected uplands to small drainageways [19, p. 5]. The proportional groundnut yields along this sequence are $\{0, .9, 1, \{.92 \mid .73\}, \{.82 \mid .63 \mid .65\}, .55\}$. Map units Ly–GoA–NoA form a ground-water depth sequence across an essentially flat landscape, and the boundary between them is often difficult to find during routine mapping. The transitions between the slope classes A (0-2%), B (2-6%), C (6-10%) D (10-15%) are fairly clear on the landscape; however, the soils in the B and C positions must be mapped by augering because of the significant differences in family particle-size class (fine loamy to fine) and thickness of a coarser-textured surface horizon (10 to 50cm). These result from different sedimentary parent materials which are exposed by dissection and different intensities of lessivation.

Symbol	**Map Unit Name**	**Yield** ($\mathrm{kg\ ha^{-1}}$)	**% of key soil**
GoA	Goldsboro fsl, 0-2%	1630	90
GyC	Gritney fsl, 6-10%	1180	65
GyD	Gritney fsl, 10-15%	1000	55
Ly	Lynchburg fsl	0	0
NoA†	Norfolk ls, 0-2%	1820	100
NoB	Norfolk ls, 2-6%	1680	92
NoC	Norfolk ls, 6-10%	1500	82
WaB	Wagram ls, 0-6%	1320	73
WaC	Wagram ls, 6-10%	1140	63

†key soil

Table 8: Map units and their groundnut yields, Edgecombe County, NC

Weights w_{ij} are calculated from the map user's perspective, i.e. by row, to reflect the loss in user's utility if the reference class is in fact found in the field instead of the mapped class. Where the soil found in the field has a higher yield than the mapped soil, a weight of 1 is assigned, reasoning that the map user has already made a decision on land use based on the mapped class. The weights are then calculated as:

$$w_{ij} \quad = \quad \min\{1, u_i/u_j\} \tag{11}$$

where u_i is the utility of class i, in this example the relative yield. In this example, we arrive at the weights matrices shown in Table 9.

Mapped	**Reference Class**								
Class	GoA	GyC	GyD	Ly	NoA	NoB	NoC	WaB	WaC
GoA	1	0.72	0.61	0	1	1	0.92	0.81	0.70
GyC	1	1	0.85	0	1	1	1	1	0.97
GyD	1	1	1	0	1	1	1	1	1
Ly	0	0	0	1	0	0	0	0	0
NoA	0.90	0.65	0.55	0	1	0.92	0.82	0.73	0.63
NoB	0.97	0.70	0.60	0	1	1	0.89	0.79	0.68
NoC	1	0.79	0.67	0	1	1	1	0.88	0.76
WaB	1	0.89	0.76	0	1	1	1	1	0.86
WaC	1	1	0.88	0	1	1	1	1	1

Table 9: Weights assigned by a utility function based on groundnut yields

Note the asymmetry of the weights: for example, since the GyD unit has the lowest non-zero yield, it is not considered an error to map this class when in fact any of the others, except Ly, is actually found in the field. By contrast, if GyD is found in the field where the other classes have been mapped, this is considered an error, with weights (partial credit to the mapper) of from 0.55 (if mapped as the key soil, NoA) to 0.88 (if mapped as WaC).

Assigning weights by expert judgement It may be difficult to design a utility function that correctly reflects the relative value of two map units, in particular because there are so many uses for soil. Another method for assigning weights follows from an attempt to formalize the quality control function of the soil correlator or senior surveyor. The idea is to determine the correlator's opinion of seriousness of any mis-mappings in terms of the usefulness of the survey and the competence of the mapper. These concepts are usually clear in the minds of the mappers in any case, since they must group observations during mapping to decide on boundaries and map unit names. This is related to the concepts of *similar* and *dissimilar* soils [64, pp. 29-30], and for dissimilar soils, *limiting* and *non-limiting* for major uses. The concept propsosed here is more flexible than these categories, because it allows the correlator to assign continuous values to particular mis-mappings, rather than considering similar soils as equivalent to the mapped soil, and dissimilar soils as mapping failures. It also allows confusion between similar soils to be considered mapping failures, if the correlator considers that the mapper should never have made such a mistake, e.g. if the soil as mapped is never found in a particular landscape.

For a large legend, rather than assign all weights individually, it is more practical and consistent to develop some rules for assignment of weights, and then adjust them individually for special cases. Errors are classified by an adjacency diagram in several dimensions, based on the major differences between soils that may be confused in mapping and which are used to define mapping classes.

For example, in the North Carolina example discussed above, the main differences between mapped series are ground-water depth, family particle-size class, and thickness of the coarser-textured surface horizon. Map units of some series are further differentiated by slope. Daniels et al. [19] use a relational diagram to show the differences of each series from the series which represents the central concept in the landscape. Rules could be formulated as follows:

> "Observed one class more poorly-drained than mapped $\rightarrow w = .9$"
> "Observed one class better drained than mapped $\rightarrow w = .95$"

Or, the rules could apply to each transition:

> "Observed excessive, mapped well $\rightarrow w = .95$"
> "Observed well, mapped moderately well $\rightarrow w = .85$"
> ...
> "Observed well, mapped excessive $\rightarrow w = .9$"
> "Observed moderately well, mapped well $\rightarrow w = .9$"
> ...
> "Observed very thick sandy surface, mapped thick $\rightarrow w = .7$"
> "Observed thick, mapped very thick $\rightarrow w = .9$ "
> ...

To combine weights computed by r rules for the same mis-classification, the simplest method is to use the minimum of the individual factors $\min_{k=1\ldots r}\{w_k\}$; another possibility is to use the geometric mean:

$$\sqrt[r]{\prod_{i=1}^{r} w_k} \qquad (12)$$

4.2 The *kappa* index of agreement

The *kappa* index of agreement for categorical data was developed by Cohen [13, 14] and associates [27] in the context of psychology and psychiatric diagnosis. The specific question they addressed is whether two classifications of the same group of subjects agree or not, in other words, whether the two persons performing a diagnosis agree in general. If not, it suggests that one or both of the diagnosticians are in error, or that the categories of the classification can not reliably be distinguished by competent professionals.

Kappa was subsequently adopted by the remote sensing community as a useful measure of classification accuracy, first in the obvious analogy to psychiatric diagnosis where the two diagnosticians are two human photo-interpreters or two automatic classifiers. It was then extended to the case where one diagnostician is the automatic classifier (the computer) and the other is the manual classifier (a human), by tradition with the manual classifier considered as the columns or ‘reference’ to which the automatic classifier is to be compared. To evaluate soil surveys, the columns remain the reference samples, and the rows are the mapped classes, whether determined by conventional soil survey methods or by automatic landform classification [48].

Kappa is explained by Congalton [15, 16], Congalton and Green [17] and Skidmore [59], among others. A formula for its variance of was developed by Fleiss et al. [27]. Correct formulas for *kappa* and its variance are presented by Hudson and Ramm [37]. A clear mathematical derivation is given by Bishop et al. [5§11.4.2].

The basic idea behind *kappa* is that some of the apparent classification accuracy could be due to chance. This is especially relevant when some classes are more likely to be encountered during field sampling than others. As a simple illustration, consider a soil map where soil type A is mapped over 90% of the survey area, and soil type B over the other 10%. This can be used as an *a priori* estimate of the actual coverage, in which case a randomly-selected field sampling point would have a 90% chance of being in fact soil type A. The joint probability is then 0.81 of a this point being correctly mapped as soil type A, strictly by chance. A similar argument for class B gives a joint probability of 0.01, so the map would be expected to have an overall accuracy of 82% simply by a chance assignment of ground truth points to mapped classes, irrespective of the actual pattern shown on the map, so long as the mapped proportions are as given. This makes it hard to distinguish a good map from one that is simply lucky.

Use of *kappa* instead of naïve accuracy measures becomes less important as the number of classes increases, as is typically the case for detailed soil maps. For example, for a map with twenty classes, each with $p_{i+} = p_{+i} = 0.05$, chance agreement is only $20 \cdot (0.05)^2 = 0.05$. It also becomes less important as the balance between classes becomes more even. For two equally-likely soil types which are each mapped over half of the area, the chance overall accuracy is 50%, compared to 82% for the 90% – 10% split calculated above.

Kappa as presented here is based on the assumption that the marginal proportions are known to the classifier before assigning observations to classes [9, 50]. This has a clear interpretation in psychiatric diagnosis. In soil survey, an estimate of the proportions of each soil class is established by the survey itself. The resulting row (map) proportions p_{i+} can be reasonably considered as prior probabilities for the column (reference) proportions p_{+j}, if we assume that the mapper is at least somewhat competent. That is, before the accuracy assessment takes place, the evaluator already knows the proportions that 'should' be found. In §4.3 we discuss the case where prior knowledge or ignorance of marginal proportions is explictly considered.

4.2.1 Unweighted *kappa*

The usual form of *kappa* is where all errors are equally important. We define four intermediate coefficients $\theta_1 \ldots \theta_4$ as follows:

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{r} x_{ii} = \sum_{i=1}^{r} p_{ii} \tag{13}$$

$$\theta_2 = \frac{1}{n^2}\sum_{i=1}^{r} x_{i+} \cdot x_{+i} = \sum_{i=1}^{r} p_{i+} \cdot p_{+i} \tag{14}$$

$$\theta_3 = \frac{1}{n^2}\sum_{i=1}^{r} x_{ii} \cdot (x_{i+} + x_{+i}) = \sum_{i=1}^{r} p_{ii} \cdot (p_{i+} + p_{+i}) \tag{15}$$

$$\theta_4 = \frac{1}{n^3}\sum_{i=1}^{r}\sum_{j=1}^{r} x_{ij} \cdot (x_{+i} + x_{j+})^2 = \sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij} \cdot (p_{+i} + p_{j+})^2 \tag{16}$$

The first coefficient, θ_1, is the overall accuracy A_o. The second, θ_2, is the expected overall accuracy if there were chance agreement between reference and mapped data. The other two coefficients are used in the computation of variance. Note that in the formula for θ_4, the *reference* total x_{+i} for mapped class i, is added to the *mapped* total x_{j+} for reference class j. That is, these are not the marginal totals for the cell, but rather those from the transpose $\mathbf{X}^{\mathrm{T}}$.

These coefficients are then combined to compute *kappa*, and its variance as follows:

$$\hat{k} = \frac{\theta_1 - \theta_2}{1 - \theta_2} \tag{17}$$

$$\sigma^2[\hat{k}] = \frac{1}{n}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\right] \tag{18}$$

In Equation 17, the denominator is the proportion of disagreement that is predicted by the marginal totals. The numerator is the actual proportion of disagreement. So $\hat{k}$ is the proportion of disagreements expected by chance that did *not* occur. If it is 1, this means that there were no disagreements. If it is 0, this means that the disagreements that occur are exactly those expected by chance. If it is less than 0, this means that the classifier actually has more disagreement with the reference data than would be expected from a pure random assignment of classified pixels to classes.

The *kappa* statistic may be difficult to explain to map users. The zero point is easy enough to understand, but the increasing accuracy of the map as *kappa* increases is not a direct relation as with overall accuracy. In particular, it depends on the marginal proportions.

It would be quite disturbing for *kappa* to be not significantly different from zero, as this would indicate that the mapper was incompetent; fortunately, in practice this is rarely the case. More interesting is the confidence interval for *kappa*, which allows us to compare two classifications to see if one is significantly better than the other [e.g

49], in the same way as was explained for the binomial test (§3.1), substituting $\hat{k}$ and $\sigma^2[\hat{k}]$ for $\hat{p}$ and $\hat{s}^2$, respectively, in Equation 4.

A numerical example for unweighted *kappa*

This example continues the analysis of the data in Table 3. The coefficients and statistics computed from this confusion matrix are shown in Table 10.

Statistic	Value
θ_1	0.5276
θ_2	0.3054
θ_3	0.3575
θ_4	0.4037
$\hat{k}$	0.3199
$\sigma^2[\hat{k}]$	0.00274
$\sigma[\hat{k}]$	0.05234
C.V.	16.4%
95% C.I.	0.2173 . . . 0.4225

Table 10: *Kappa* statistics for the example confusion matrix

In this example, $\hat{k}$ ($\approx 32\%$) is substantially lower than the naïve measure of agreement, namely overall accuracy $A_o = \theta_1$ ($\approx 53\%$), indicating that a large portion of the apparent classification accuracy could be due to chance agreement. Because of the small sample size, the confidence interval for $\hat{k}$ is wide. This is also shown by the high coefficient of variability. These are decreased as the square root of the increase in sample size. For example to double the precision, the sample size must be quadrupled.

4.2.2 Conditional *kappa*

Kappa may also be calculated for each class, either by row (user's accuracy) or column (producer's reliability) [50, 5]. This allows us to determine which classes are well-mapped. The calculation follows the same logic as overall *kappa*, but is restricted to one row (Equation 19) or column (Equation 20):

$$\begin{aligned}
\hat{k}_{i+} &= \frac{(p_{ii}/p_{i+}) - p_{+i}}{1 - p_{+i}} = \frac{C_i - p_{+i}}{1 - p_{+i}} && (19)\\
&= \frac{p_{ii} - (p_{i+} \cdot p_{+i})}{p_{i+} - (p_{i+} \cdot p_{+i})}\\
\hat{k}_{+j} &= \frac{(p_{jj}/p_{+j}) - p_{j+}}{1 - p_{j+}} = \frac{O_j - p_{j+}}{1 - p_{j+}} && (20)\\
&= \frac{p_{jj} - (p_{+j} \cdot p_{j+})}{p_{+j} - (p_{+j} \cdot p_{j+})}
\end{aligned}$$

Conditional *kappa* is in the interval $[0 \ldots p_{+i}]$ (user's) or $[0 \ldots p_{j+}]$ (producers's). It is thus a downward adjustment of the naïve measure of accuracy to take into account chance agreement. In general, $\hat{k}_{i+} = C_i$ iff $C_i = 1$, i.e. perfect user's accuracy for the class. For a given agreement C_i, $\hat{k}_{i+}$ is lower to the extent that the marginal proportion p_{+i} of the reference class increases. That is, as the reference class becomes more likely, the apparent agreement represented by C_i must be adjusted more.

The variance of conditional *kappa* for user's class i is calculated as [5, Eqn. 11.4-10]:

$$\begin{aligned} \sigma^2[\hat{k}_{i+}] &= \frac{1}{n} \cdot \left\{ \frac{p_{i+} - p_{ii}}{{p_{i+}}^3 \cdot (1 - p_{+i})^3} \right. \\ &\quad \left. \cdot [(p_{i+} - p_{ii})(p_{i+} \cdot p_{+i} - p_{ii}) + p_{ii} \cdot (1 - p_{i+} - p_{+i} + p_{ii})] \right\} \end{aligned} \tag{21}$$

For producer's class i, Equation 21 is permuted by interchanging the row and column summaries, i.e. p_{i+} and p_{+i}.

A numerical example for conditional *kappa*

Table 11 shows conditional *kappa* and their variances for the four classes of the sample confusion matrix of §4.2.1 (Table 11).

	User's accuracy				Producer's reliability			
Class	C_i	$\hat{k}_{i+}$	$\sigma[\hat{k}_{i+}]$	C.V.,%	O_j	$\hat{k}_{+j}$	$\sigma[\hat{k}_{+j}]$	C.V.,%
A	0.5738	0.3684	0.0763	20.7	0.6604	0.4573	0.0899	19.6
B	0.6111	0.4888	0.1440	29.5	0.2821	0.1929	0.0673	34.9
C	0.6032	0.3466	0.0824	23.8	0.5938	0.3378	0.0806	23.9
D	0.0952	0.0546	0.0603	110.3	0.2857	0.1801	0.1906	105.8

Table 11: Conditional *kappa* statistics for the example confusion matrix

As expected, both user's and producer's conditional *kappa* are smaller than the naïve measures of accuracy. Class D is clearly very poorly mapped. We can see an important discrepancy between user's and producer's accuracy both for this class and class B. The coefficients of variation are all higher than for overall $\hat{k}$, due to the small sample sizes.

4.2.3 Weighted *kappa*

Kappa may also be computed for the weighted confusion matrix presented in §4.1.1, which gives partial credit for incorect classifications. This was first developed by Cohen [14] and then refined by Fleiss et al. [27]. Its practical application in remote sensing accuracy assessment was explored by Næsset [51]. The formulas presented here are equivalent to, but computationally simpler than, those from these sources.

From the confusion matrix and the notation of Table 5, we define several intermediate coefficients:

$$\theta_{w1} = \sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij} \cdot p_{ij} \tag{22}$$

$$\theta_{w2} = \sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij} \cdot p_{i+} \cdot p_{+j} \tag{23}$$

$$\theta_{w4} = \sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij} \cdot [w_{ij} \cdot (1-\theta_{w2}) - (\overline{w}_{i+} + \overline{w}_{+j}) \cdot (1-\theta_{w1})]^2 \tag{24}$$

The coefficients θ_{w1} and θ_{w2} are the weighted equivalents of θ_1 and θ_2 from Equations 13 and 14, and have the same meaning, namely actual and chance agreement, but here with 'agreement' including partial agreements. These are then combined to produce weighted *kappa* and its variance, as follows:

$$\hat{k}_w = \frac{\theta_{w1} - \theta_{w2}}{1 - \theta_{w2}} \tag{25}$$

$$\sigma^2[\hat{k}_w] = \frac{\theta_{w4} - (\theta_{w1} \cdot \theta_{w2} - 2\theta_{w2} + \theta_{w1})^2}{n \cdot (1-\theta_{w2})^4} \tag{26}$$

If $\mathbf{W} = \mathbf{I}$, i.e. the diagonals are 1 and the off-diagonals 0, these formulas give the same results as the formulas for unweighted *kappa* (Equations 17 and 18).

A numerical example for weighted *kappa*

This example continues that of §4.1.1. We calculate the coefficients and statistics shown in Table 12.

Statistic	Value
θ_{w1}	0.7332
θ_{w2}	0.6312
θ_{w4}	0.0187
$\hat{k}_w$	0.2776
$\sigma^2[\hat{k}_w]$	0.004741
$\sigma[\hat{k}_w]$	0.06886
C.V.	24.1%
95% C.I.	0.1417 . . . 0.4116

Table 12: *Kappa* statistics for the example confusion and weight matrices

As in the unweighted case, $\hat{k}_w$ is substantially lower than the naïve measure of agreement, here the *weighted* overall accuracy A_{ow} = θ_{w1}. In this example, there is even a greater discrepancy than for the unweighted example, because, first, the apparent agreement is much higher due to the partial credit given to incorrect answers, and second, the weighted chance agreement is higher. Weighted *kappa* may be higher or lower than unweighted *kappa*, depending on the confusion and weight matrices. In the present case, the large proportions of classes A and C in both the reference and mapped population, along with the high partial credit given to the misclassification between these two, leads to a high chance agreement.

Weighted *kappa* gives an insight into the value of the map as compared to a random assignment of areas to map units with the specified proportions. In this example it is quite low, about 28%. Thus the map is quite good for the user (as shown by $A_{ow} \approx 73\%$ from Table 6), but the producer didn't improve the pre-survey knowledge very much.

4.3 The *tau* index

The various *kappa* indices all assume that both the user's and producer's marginal probabilities for each class are known before the classification. In remote sensing applications, this is not the case for the user's marginals (rows of the confusion matrix), which are not known until the automatic classifier completes its work. The producer's marginals (columns), by contrast, are considered to be known before the classification, since they are independently-determined by field sampling. Ma and Redmond [44] proposed the *tau* index of agreement for this situation. Their work was extended by Næsset [50] to assess per-class producer's reliability; the per-class user's accuracy is the same as conditional *kappa*. *Tau* may be especially appropriate for assessing soil maps made by automatic classifiers [e.g. 73].

To assess overall accuracy, Ma and Redmond [44] define (using the notation of Table 2 and §4.2.1):

$$\tau = \frac{\theta_1 - \theta_2'}{1 - \theta_2'}, \text{ where} \tag{27}$$

$$\theta_1 = \sum_{i=1}^{r} p_{ii} \tag{28}$$

$$\theta_2' = \sum_{i=1}^{r} p_i \cdot p_{+i} \tag{29}$$

$$\theta_3' = \sum_{i=1}^{r} p_{ii} \cdot (p_i + p_{+i}) \tag{30}$$

$$\theta_4' = \sum_{i=1}^{r} \sum_{j=1}^{r} p_{ij} \cdot (p_{+i} + p_j)^2 \tag{31}$$

Equation 27 has the same form as the definition of $\hat{k}$ in Equation 17, except that θ_2 from Equation 14, representing chance agreement, is replaced by θ_2' in Equation 27. This equation in turn uses p_i instead of p_{i+} to represent the row proportions. These p_i are the *prior* probabilities of class membership. In the absence of specific information, they may all be set to $1/r$, i.e. equal probability for each of the r classes. In this situation, τ is the same as the modified $\hat{k}$ proposed by Foody [28]. *Tau* can also be applied with any other prior assignment of classes, for example, an estimate of the distribution of soil classes based on a detailed map of a sample area.

The coefficients θ_3' and θ_4' are modifications of Equations 15 and 16, with prior row proportions substituted for actual. They can then be used to calculate the variance of *tau*, $\sigma^2[\tau]$, with Equation 18, and this variance can be used to compute confidence intervals as for *kappa*.

A numerical example for *tau*

Table 13 shows the computed coefficients and statistics using the data of Table 3, for three cases: (1) assuming that each class is equally likely *a priori*, i.e. $p_i = 0.25, \ \forall i$ (left column); (2) with unequal priors, in this case $\{0.1, \ 0.4, \ 0.1, \ 0.4\}$, i.e. with the highest expectation for the classes that are not mapped over the largest areas (center column); and (3) with unequal priors, in this case $\{0.4, \ 0.1, \ 0.4, \ 0.1\}$, i.e. with the highest expectations closely matching the largest mapped areas (right column).

Statistic	**Value**†	**Value**‡	**Value***
θ_1	0.5276	0.5276	0.5276
θ'_2	0.2500	0.1847	0.3153
θ'_3	0.3099	0.2547	0.3651
θ'_4	0.3209	0.2667	0.4202
τ	0.3701	0.4206	0.3100
$\sigma^2[\tau]$	0.00239	0.002064	0.002186
$\sigma[\tau]$	0.0489	0.04543	0.05307
C.V.	13.2%	10.8%	17.1%
95% C.I.	0.2743 . . . 0.4660	0.3316 . . . 0.5097	0.2060 . . . 0.4141

†Equal priors, all 0.25

‡Priors $\{0.1, 0.4, 0.1, 0.4\}$

*Priors $\{0.4, 0.1, 0.4, 0.1\}$

Actual proportions $\{0.3742, 0.1104, 0.3865, 0.1288\}$

Table 13: Tau statistics for the example confusion matrix

With equal priors (left column), θ'_2 is substantially lower than the chance agreement θ_2 computed for $\hat{k}$. This leads to a value of τ that is significantly higher than $\hat{k}$. This is because, in this example, the marginal proportions for the rows are far from equal, and furthermore the two largest match the corresponding row proportions fairly closely. The standard deviation for *tau* is somewhat lower than for *kappa*.

If the prior probabilities are quite different from the actual proportions (center column), *tau* is even closer to the overall accuracy; this is because the map is not what was expected, so agreement is a pleasant surprise. In this case, as shown in the rightmost column of Table 13, there is a smaller chance agreement, the index of agreement is closer to the overall accuracy, and the variability of the estimate is lower. The map is providing a large increase in information over the prior state of knowledge.

If the prior probabilities are equal to the actual row proportions, *tau* is the same as *kappa*. This situation is closely approximated in the rightmost column of Table 13, where the statistics are quite close to those shown in the example for unweighted *kappa* of Table 10.

5 Multinomial tests applied to compound map units

A variant of the confusion matrix may be used to simultaneously verify the overall accuracy and the agreement with stated proportions of a compound map unit. Here the rows are legend categories, typically compound map units (with simple map units being a special case), and the columns are the possible categories for validation observations, i.e. simple classes, some of which may be named constituents. The matrix is not symmetrical, and only the row-wise statistics, corresponding to map legend categories, are important. To avoid confusion, we call this a *validation* matrix Y, with a corresponding row-wise proportions matrix Q.

In a typical application, the simple classes include all the possible taxonomic classes that may be encountered in the study area, even if they are not named in any legend category. Observations that do not fall in any class may be placed either with the class which they most closely resemble (if sufficiently similar) or in a separate 'other' class that will always be treated as a mapping failure.

We define another matrix R, termed the *predicted proportions* matrix, with the same structure as Q, and containing the predicted proportions of each validation class according to the legend. Most of the cells in this matrix will be zero; the only non-zero proportions will be for classes mentioned in the map unit name, and perhaps in the descriptive legend. Each row must sum to unity; column sums are meaningless. In the ideal situation $Q = R$.

Symbol	**Meaning**	**Computation**
$\mathbf{Y}$	validation matrix $\rightarrow$ rows $[1 \ldots r]$ are legend categories $\rightarrow$ columns $[1 \ldots c]$ are validation classes	
y_{ij}	number of observations in row i, column j, i.e. mapped in legend category i and observed in validation class j	
y_{i+}	total number of observations for legend category i	$\sum_{j=1}^{c} y_{ij}$
$n = y_{++}$	total number of observations	$\sum_{i=1}^{r} y_{i+}$
$\mathbf{Q}$	row-wise proportions matrix	
q_{ij}	proportion of the observations in legend category i that are in validation class j	y_{ij}/y_{i+}
q_{i+}	sum of proportions for legend category i	$\sum_{j=1}^{c} q_{ij} = 1$
$\mathbf{R}$	predicted proportions matrix	
r_{ij}	proportion of legend category i that should be in validation class j	
r_{i+}	sum of predicted proportions for legend category i	$\sum_{j=1}^{c} r_{ij} = 1$
δ_{ij}	indicator of a class with non-zero expectation	$\begin{cases} 1 & : & r_{ij} > 0 \\ 0 & : & r_{ij} = 0 \end{cases}$
d_i	count of classes with non-zero expectation for legend category i	$\sum_{j=1}^{c} \delta_{ij}$

Table 14: Notation for the validation matrix for compound map units

Table 14 shows the notation of the validation and predicted proportion matrices. We may consider the validation matrix Y to be a realisation of a multinomial distribution whose underlying probabilities are the declared proportions of each constituent, multiplied by the number of observations for the legend category. That is, we expect

$$y_{ij} \quad \rightarrow \quad r_{ij} \cdot y_{i+} \tag{32}$$

where the left-hand side is the observed, and the right hand side the expected, count. This implies that for cells with $r_{ij} = 0$, we expect no observations; however, we know that in any map unit it is possible to encounter any class (perhaps because of poor mapping, perhaps because of inclusions). So the table is not incomplete in the statistical sense [5, p. 58], i.e. containing cells in which it is impossible to encounter an observation. The difficulty is in assigning a maximum-likelihood estimate of the expected count in the cells which the map predicts should be empty. The usual method of multiplying the marginal proportions is not valid here, because the expected cell proportions to test are already set in the legend.

The appropriate measure of fit of the actual data to this model is Pearson's χ^2, calculated over the cells with non-zero expectation:

$$\chi^2 \quad = \quad \sum_{i,j} \frac{(y_{ij} - r_{ij} \cdot y_{i+})^2}{r_{ij} \cdot y_{i+}}, r_{ij} > 0 \tag{33}$$

For each legend category i there are as many degrees of freedom as named constituents d_i. Observations found in non-named classes are grouped into an 'other' class for purposes of computing the degrees of freedom. The degrees of freedom are summed over all legend categories (rows) to assess the entire map: $df = \sum_i d_i$. Then the accuracy of the map is measured by the probability that the observed χ^2 calculated by Equation 33 could have arisen by chance from the X^2 distribution:

$$\Pr\{X^2_{df} \leq \chi^2\} \tag{34}$$

The probability of the observed χ^2 may be used directly as a measure of map accuracy; the lower the value, the better. Two maps may be compared by their respective probabilities, or, if the legend is the same, directly by the value of χ^2.

Table 15 shows an example of four map units with varying declared proportions of three classes. Observations not falling in any of these have been recorded in a separate 'Other' class. The table shows the observed counts with their proportions, the expected counts, the contribution to χ^2 for each cell, the combined χ^2 (here ≈ 7.4) and its probability. In this example, there is a fairly good fit between the prediction and validation. There is still $\approx 10\%$ probability that the observed discrepancy could be due to chance.

From the row summaries we can see that Map Unit 1 is by far the most poorly mapped, because class B, which was supposed to occur in 40% of the map unit in fact was found in only $\approx 30\%$ of the validation observations; the rest were mostly class C, which was not named in the legend. It is also possible to compute the probability that the observed χ^2 could be due to chance separately for each legend category, with d_i degrees of freedom; in this example they are 0.05, 0.38, 0.24, and 0.25, respectively. This implies that Map Unit 1 may be considered to be incorrectly characterised at the $\alpha = 0.05$ level, but the others must be accepted as correctly characterised.

To emphasize the importance of the class proportions in this assessment, we take the predicted proportions matrix $\mathbf{R}$ of Table 15 and replace it with one where the expected proportions are reversed, e.g., for Map Unit 1, $40\% - 60\%$ instead of $60\% - 40\%$. With the same validation observations (Y), the computed χ^2 is ≈ 145, which is effectively impossible by chance.

		Reference Class				Total
		A	B	C	O	
	Map Unit 1	35	14	11	1	61
	A & B	*0.5738*	*0.2295*	*0.1803*	*0.0164*	*1*
	60%–40%	36.6	24.4	0	0	61
		0.0699	4.4328	–	–	4.5027
	Map Unit 2	4	11	3	0	18
	B & A	*0.2222*	*0.6111*	*0.1667*	*0.0000*	*1*
	70%–30%	5.4	12.6	0	0	18
Legend		0.3630	0.2032	–	–	0.5661
Category	Map Unit 3	12	9	38	4	63
	C & A & B	*0.1905*	*0.1429*	*0.6032*	*0.0635*	*1*
	60%–25%–15%	15.75	9.45	37.8	0	63
		0.8929	0.0214	0.0011	–	0.9153
	Map Unit 4	2	5	12	2	21
	C & B	*0.0952*	*0.2381*	*0.5714*	*0.0952*	*1*
	60%–40%	0	8.4	12.6	0	21
		–	1.3762	0.0286	–	1.4048
						163
Cell contents: y_{ij}, q_{ij}, $(r_{ij} \cdot y_{i+})$, contribution to χ^2						–
						163
				$\Pr(X^2{}_{df=9} \leq 7.389) = 0.8964$		7.389

Table 15: Example validation matrix for compound map units

6 Sampling schemes for accuracy assessment

The accuracy of a thematic map with respect to ground truth is assessed by means of a sample, where the observed class, here considered the true or 'reference' class, is compared with the class as mapped. There are three types of reference observations:

1. point observations in the field, of a minimum representative volume for classification, usually a pedon [64, p. 18];
2. point observations interpreted from a more detailed map or photo, presumed to be correct for the purposes of accuracy assessment; and
3. an entire reference map, presumed to be correct for the purposes of accuracy assessment. This is typically a more detailed survey.

For the first two types of observations, two issues must be resolved: (1) the sample design, that is, where the sample points should be located and what should be their relation, if any, to each other (§6.1), and (2) how many samples must be taken to ensure acceptable estimates of accuracy (§6.3). For last type of observation, the entire map is used, and the only question is how to discretize the area into samples (§6.4).

6.1 Sample design

For point observations, sampling may be simple random or stratified random by map unit, using transects across, or small blocks within, each map unit [32].

Simple random sampling [12, Ch. 1] is the most statistically-defensible, especially for binomial and multinomial tests discussed later; however it suffers from several defects. First, small classes may be under-sampled by chance. Second, it is expected to have the highest cost of field access, and in fact may be impractical if some areas of the map are very difficult to visit.

The problem of accessibility may be solved by discarding inaccessible sample points, as long as we can ensure this does not lead to bias; unfortunately this is rarely the case in soil survey, where inaccessible points are likely to be hillier, wetter, of lower productivity, etc. than accessible points. Another approach is to select a compact *representative area* where all classes are expected to be present in the same proportion to the entire mapped area, but where accessibility is satisfactory.

Stratified random sampling has several advantages [12, Ch. 5]. First, it ensures that all classes are sampled in proportion to their occurrence. Second, it provides estimates for each class separately. Third, if the classes differ in their internal variability, it may lead to a better estimates of overall accuracy.

Forbes et al. [29, p. 28] suggest stratifying the accuracy assessment by major soil landscape, since these almost always have soil sets with different patterns, soil-landscape relations, and heterogeneity, present different mapping difficulties, and are therefore expected to have different accuracies. These strata can be major physiographic units, soil systems [19], major land resource areas [62], or high levels in a soil-geomorphic classification [e.g. 80].

Samples may be selected independently or grouped by transects. Transects sacrifice some theoretical randomness for major practical advantages. First, the samples are nearby and can be efficiently located. Second, they are usually set up to run perpendicular to the major variations in the landscape, thereby covering a wide range of map units.

Spacing in transects may be regular or random. It is a common mis-perception that spatial dependence makes close-by samples invalid. Brus and Gruijter [10] show that in a design-based approach, any spatial dependence may result in less precise estimates, but does not invalidate the sampling scheme.

6.2 Sample distribution for multinomial tests

The various multinomial tests of §4 and §5 do not test whether the sample represents the population. For example, if only well-mapped units were sampled, the accuracy statistics may be quite favourable, with no indication that the more difficult units were not included in the test. A stratified random sample by definition represents its strata, but for other designs, purposive sampling, or a design not stratified by map unit, we need a test of representativeness.

A sample is considered to be well-distributed if its distribution over some stratification is 'close enough' to the true proportions of the strata. Since we are evaluating a area-class maps, the obvious stratification is by map unit (legend category), and the appropriate test is Pearson's χ^2 on a one-dimensional array of class proportions:

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - r_i \cdot n)^2}{r_i \cdot n} \tag{35}$$

where y_i is the number of samples in legend category $i \in [1 \ldots k]$, r_i is the proportion $(0 \ldots 1]$ of the map covered by category i, and n is the sample size. The distribution of the sample over the legend is measured by the probability that the observed χ^2 calculated by Equation 35 could have arisen by chance from the X^2 distribution.

$$\Pr\{X^2_{df=(k-1)} \leq \chi^2\} \tag{36}$$

6.3 Sample size

An important issue is the number of samples that would be necessary to achieve certain statistical criteria. This is especially difficult in soil survey because of the expense of soil sampling.

6.3.1 Sample size for binomial tests

The binomial test is used to estimate the overall proportion of agreement between a map and a validation sample (§3). Cochran [12, §4.4] presents a derivation of the required sample size, which we follow here, with the difference that, since the population of possible sample sites (pedons) is very large, we can ignore finite-population corrections. This topic is also examined by Webster and Oliver [77, Ch. 3]. The required sample size n is estimated as:

$$n = \frac{{Z_\alpha}^2}{b^2} \cdot \hat{p}\,(1 - \hat{p}) \tag{37}$$

where $\hat{p}$ is a prior estimate of the proportion that will be found to be correct, b is the desired absolute precision as a proportion, and Z_α is the normal score for the desired two-tailed probability α of Type I error. Note that as the proportion expected to be correct $\hat{p} \rightarrow 1$, the required sample size n for a given absolute precision decreases; that is, the more successful the map, the fewer samples are needed to prove this.

The expected proportion $\hat{p}$ may be estimated from previous experience or a preliminary sample. Type I error α is the risk that the true proportion p is outside the confidence limits determined by the standard error of the estimate:

$$\Pr\left\{|\hat{p} - p| \geq Z_\alpha \cdot \sqrt{p \cdot (1-p)/n}\right\} = \alpha \tag{38}$$

That is, by specifying α we specify both the width of the confidence interval and the probability that the true proportion is outside it.

For example, for $\alpha = 0.05$, then $Z_\alpha = 1.96$, that is, 5% of the normal distribution lies in the tails beyond ± 1.96. If the allowable error $b = \pm 5\%$ (± 0.05 as an absolute precision) on a prior estimate $\hat{p} = 0.9$, we calculate:

$$n = \frac{(1.96)^2}{(0.05)^2} \cdot 0.9 \cdot 0.1 \approx 138$$

If the required confidence is relaxed to $\alpha = 0.1$, then $Z_\alpha = 1.64$, and the required sample size is reduced by a factor of $1.64/1.96 = 0.84$ to 116.

The sample size may also be calculated as a simplification of the calculation for multinomial tests (next §), setting $k = 1$.

6.3.2 Sample size for multinomial tests

A multinomial accuracy tests (§4) such as *kappa* estimates per-class accuracies as well as overall accuracy. Tortora [68] developed the following criteria for large populations. The required sample size n_i to obtain an absolute precision of b_i for class $i = [1 \ldots k]$ and for probability of Type I error α is computed from the expected proportion $\hat{p}_i$ as:

$$n_i = \frac{B}{{b_i}^2} \cdot \hat{p}_i \; (1 - \hat{p}_i), \text{where} \tag{39}$$

$$B = \chi^2_{(1,\ \alpha/k)} \tag{40}$$

where χ^2 are the tabulated values of the cumulative *Chi-square* distribution with one degree of freedom such that $\Pr\{\chi^2\} > \alpha/k$. The $\hat{p}_i$ may be estimated from the proportions of each legend class as mapped.

The total number of samples to estimate all classes with the required precision is found as the maximum of the individual estimates:

$$n = \max_i \{n_i\} \tag{41}$$

Here are some useful values of B:

α/k	0.25	0.1	0.05	0.025	0.01	0.005	0.001
B	1.323	2.706	3.841	5.024	6.635	7.879	10.828

Some tables of χ^2 use instead the lower tail, i.e. $\Pr\{\chi^2\} < \alpha/k$, in which case use $1 - (\alpha/k)$ in the formulas for B.

For example, to assess the accuracy of a soil map with five legend categories ($k = 5$), covering 40%, 25%, 20%, 10%, and 5% of the area, respectively, to an absolute precision of 5% for each class, we use $\hat{p}_i = 0.4$, i.e. the class proportion closest to 0.5. If the probability of a Type I error is set to 0.05, we calculate:

$$B = \chi^2_{(1,\ 0.05/5)} = \chi^2_{(1,\ 0.01)} = 6.635$$

$$n = \frac{6.635}{(0.05)^2} \cdot 0.4 \; (1 - 0.4) \approx 637$$

The number of samples required increases as the expected proportion of any class approaches 0.5. So maps with balanced class distributions require fewer samples to validate. This is determined by the landscape, and is beyond the evaluator's control. However, two factors may be set by the evaluator: the required absolute precision of class proportions and the acceptable risk of Type I error. Low precision and high acceptable error lead to decreased sample sizes. For most field sampling programs, the precisions that may be realistically obtained are limited. It also seems realistic accept an $\alpha = 0.1$ probability of Type I error.

Soil maps often have many categories, each with a correspondingly small proportion of the map. This makes the situation even worse if we want to keep the relative precision within reasonable bounds. For example, if $k = 20$ with $\hat{p}_i \approx 0.05$ for all classes, per-class relative errors of 20% correspond to absolute errors of 0.01, i.e. each class should be estimated as $5\% \pm 1\%$. If we allow 10% probability of a Type I error, n is computed as:

$$\begin{aligned} B &= \chi^2_{(1,\ 0.10/20)} = \chi^2_{(1,\ 0.005)} = 7.879 \\ n &= \frac{7.879}{(0.01)^2} \cdot (0.05\ (1 - 0.05)) \approx 3,742 \end{aligned}$$

which is truly prohibitive. Thus it seems unreasonable to expect that a field sampling program can establish tight bounds on the true proportions of a large number of classes. If classes are grouped so that related soils are considered together, the number of classes is reduced and the same relative precision corresponds to a wider absolute precision; then the number of samples may be achievable. For example, with five equally-likely classes to be estimated as $20\% \pm 5\%$ (relative precision 0.25), with $\alpha = 0.05$, the required sample size is $n \approx 505$.

6.4 Sampling within a reference map

The accuracy of a thematic map may also be assessed by comparison with a reference *map* (as opposed to a set of reference *samples*) which is considered to be 'correct' or at least the best that can be achieved. In soil survey, this reference map may be one that was produced by detailed survey, and the map to be evaluated may be a map produced by an experimental method such as automated landform classification [48]. This is an exhaustive sample, over the entire mapped area, so that sample size is no problem. A naïve measure of accuracy expresses the adequacy of the map for the user; however, *kappa* or *tau* should be used to compare several maps, especially if some classes are mapped over a large proportion of the area.

The maps must be discretized for sampling by dividing the area into cells of a size consistent with the scale of the thematic map. A reasonable choice for cell size is one-quarter of the Minimum Legible Area (MLA), which corresponds to the ground area represented by the Minimum Legible Delineation (MLD) of 25 mm² [75] or 40 mm² [29] on the map. The MLA may be computed from the scale number SN (denominator of the scale fraction) as:

$$\mathrm{MLA}_{25} = (\mathrm{SN}/100)^2 \cdot 0.25\ \mathrm{m}^2 \quad (42)$$

$$\mathrm{MLA}_{40} = (\mathrm{SN}/100)^2 \cdot 0.40\ \mathrm{m}^2 \quad (43)$$

and the required cell size as one-quarter of this. For example, on a 1:50 000 map, $\mathrm{MLA}_{25} = 62\,500\ \mathrm{m}^2$ (6.25 ha) and $\mathrm{MLA}_{40} = 100\,000\ \mathrm{m}^2$ (10 ha), so that a reasonable cell size for accuracy assessment would be $15\,625\ \mathrm{m}^2$ to $25\,000\ \mathrm{m}^2$, corresponding to grid cells $125\ \mathrm{m}^2$ to $\approx 158\ \mathrm{m}^2$ on a side.

As an illustration, consider the confusion matrix of Table 16, adapted from a study of the success of guided landform classification compared to conventional aerial photo-interpretation for soil survey [34]. Six landform classes in a hilland were mapped by photo-interpretation (columns) using the geo-pedological approach of Zinck [81]. In this hierarchical legend, there were four 'relief types' (Hi1, Hi2, Hi3 and Hi4), each with one lithology; two of the relief types were further divided into two 'landforms'.

The area was then mapped by supervised classification based on landform parameters, as shown in the rows. This is quite close to the concept of 'independent diagnosis' for which *kappa* was first developed. A total of 20 107 30m² pixels, corresponding to a map scale of $\approx$ 1:12 000 (using the definition for MLA_{25}), were compared; this is an exhaustive sample over the study area of about 18km².

		API Class						Total
		Hi111	Hi211	Hi212	Hi311	Hi312	Hi411	p_{i+}
Mapped Class	Hi111	4401	37	0	294	13	118	4863
		0.2189	*0.0018*	*0*	*0.0146*	*0.0006*	*0.0059*	*0.2419*
	Hi211	34	541	231	1263	51	0	2120
		0.0017	*0.0269*	*0.0115*	*0.0628*	*0.0025*	*0*	*0.1054*
	Hi212	0	284	700	542	143	0	1669
		0	*0.0141*	*0.0348*	*0.0270*	*0.0071*	*0*	*0.0830*
	Hi311	1054	454	182	4409	130	15	6244
		0.0524	*0.0226*	*0.0091*	*0.2193*	*0.0065*	*0.0007*	*0.3105*
	Hi312	7	18	782	851	626	309	2593
		0.0003	*0.0009*	*0.0389*	*0.0423*	*0.0311*	*0.0154*	*0.1290*
	Hi411	0	0	14	66	31	2507	2618
		0	*0*	*0.0007*	*0.0033*	*0.0015*	*0.1247*	*0.1302*
	Total	5496	1334	1909	7425	994	2949	20107
	p_{+j}	*0.2733*	*0.0663*	*0.0949*	*0.3693*	*0.0494*	*0.1467*	*1.0000*

Table 16: Confusion matrix for an exhaustive sample

Inspecting the columns of the table, we see that API class Hi211 was often mis-mapped by the automatic classifier as Hi311 or Hi212; class Hi212 as class Hi312; and class Hi311 as Hi211 or Hi312. Comparing the marginal proportions, we see that the automatic classifier greatly over-mapped Hi211 and Hi312 (the scarp slopes and the bottoms of small vales), and somewhat under-mapped the other classes. Class Hi211 thus causes problems from both the API and classifier's points of view. This is shown clearly in the per-class (conditional) values (Table 17); here also we see the narrow confidence limits.

The naïve overall agreement between automated and manual classification is 0.6557 (C.V. = 0.5%). *Kappa* is significantly lower, 0.5579 (C.V. = 0.7%), indicating the large chance agreement; *tau* with equal prior probabilites has a value somewhat closer to the naïve value, 0.5868 (C.V. = 0.6%). Because of the large sample numbers, the variance is very low and the confidence intervals are quite narrow, so that the success of different automated classification algorithms can easily be distinguished.

If we give partial credit of 0.8 to mis-classifications within the same relief type (i.e. confusion between Hi211 and Hi212 or Hi311 and Hi312), the weighted naïve overall agreement is raised somewhat to 0.7152 and the weighted *kappa* to 0.6023. This shows that most of the errors were between, not within, relief types.

Class	User's accuracy (*Success of automated classifier*)				Producer's reliability (*Reproducibility of API*)			
	C_i	$\hat{k}_{i+}$	$\sigma[\hat{k}_{i+}]$	C.V.,%	O_j	$\hat{k}_{+j}$	$\sigma[\hat{k}_{+j}]$	C.V.,%
Hi111	0.9050	0.8693	0.0056	0.6	0.8008	0.7372	0.0067	0.9
Hi211	0.2552	0.2023	0.0094	4.6	0.4055	0.3355	0.0144	4.3
Hi212	0.4194	0.3585	0.0127	3.5	0.3667	0.3094	0.0113	3.7
Hi311	0.7061	0.5341	0.0081	1.5	0.5938	0.4109	0.0068	1.7
Hi312	0.2414	0.2020	0.0080	4.0	0.6298	0.5750	0.0172	3.0
Hi411	0.9576	0.9503	0.0046	0.5	0.8501	0.8277	0.0074	0.9

Table 17: Per-class statistics for an exhaustive sample

7 Conclusions

To assess the thematic accuracy of an area-class soil map, a series of decisions must be taken, which can greatly affect the assessment. For accuracy assessment from the user's point of view, they are taken with reference to the user's requirement for the map's *adequacy*, i.e. its fitness for use [29]. For quality control from the producer's point of view, they are taken with reference to quality standards of the producing organisation.

1. Decide on the *kind of thematic accuracy* that is to be measured (§1.4). This is one of
   - legend,
   - taxonomic,
   - attribute,
   - interpretive, or
   - interpretive classification.
2. Decide on the *thematic precision* that is required. Either
   - as precise as the legend, or
   - generalised.

   If the second, classes must be grouped or attributes or interpretations must be generalised.
3. Decide on how close an agreement between map and observation should qualify as a 'success' (a match). This may be
   - exact,
   - within a given tolerance (similar classes or a range of an attribute), or
   - weighted (partial credit).
4. Decide on a binomial or multinomial test (§4).
5. If a multinomial test, decide which statistic best expresses the accuracy:
   - naïve (§4.1),

   - *kappa* (§4.2),
   - *tau* (§4.3), or
   - validation of compound map units (§5).

6. If a weighted multinomial assessment has been selected, define a weights matrix (§4.1.2), which may be based on:
   - utility functions, or
   - producer's judgement.

7. If a *tau* test, decide on prior probabilities (§4.3):
   - equal, or
   - based on earlier surveys or estimates.

8. Decide on the level of statistical precision, in particular, the acceptable probability of a Type I error (§3); for soil survey applications the 90% level is often used.

9. Decide on a sampling design (§6.1):
   - random,
   - stratified by map unit,
   - stratified by region, e.g. major landscape, or
   - complete (using a reference map).

10. For complete sampling from a reference map, calculate the discretization (grid cell size) (§6.4).

11. For point sampling, calculate the required sample size (§6.3).

12. Decide on some other issues in point sampling:
    - What to do with inaccessible points?
    - What to do with points near spot or line symbols (§1.3)?
    - What to do with points near boundaries (§1.3)?

For dataset documentation (metadata), both the choices made and the results must be reported. It may also be useful to include explanations, in clear but non-technical language, of concepts such as Type I error, chance agreement, and confidence intervals. For example, an appropriate statement for accuracy assessment might be:

> "The accuracies for this soil geographic database are reported as a most probable value and a 80% 'confidence interval', which gives a minimum and maximum accuracy. This was based on an independent field sample to compare the soil types predicted by the map with the soil types actually found in field at these points. We believe that this sample was unbiased and representative of all the possible samples that could have been chosen. If this is true, the confidence intervals that we report have an 80% probability of actually containing the true accuracy. This means that, if we had been able to take a large number of similar samples, 80% of the

> confidence intervals that we calculated from these would have contained the true accuracy. Obviously, we do not know if our one sample is one of those 80% or one of the 20% that does not contain the true accuracy."

Supplementary material

R [38, 55] and MATLAB [53, 67] functions to compute naïve measures, *tau*, unweighted, weighted, and conditional *kappa*, as well as sample confusion and weight matrices, may be found on the author's web site:

`http://www.itc.nl/~rossiter/pubs/accuracy.html`

Acknowledgements

Thanks to Steve Carlisle of the USDA Natural Resources Conservation Service, New York State, for his unique ability to explain the mental processes of the field soil surveyor, and for stimulating discussions which helped me clarify many of the issues in this paper.

References

[1] Arbeitskreis für Bodensystematik der Deutschen Bodenkundlichen Gesellschaft. Systematik der Böden und der bodenbildenden Substrate Deutschlands. *Mitteilungen der Deutschen Bodenkundlichen Gesellschaft*, 86:1–180, 1998.

[2] B W Avery. *Soil classification for England and Wales: higher categories*. Soil Survey Technical Monograph 14. Cranfield University, Soil Survey & Land Research Centre, Cranfield, England, 1980.

[3] P H T Beckett. Method and scale of land resource survey, in relation to precision and cost. In G A Stewart, editor, *Land evaluation: Papers of a CSIRO Symposium, organized in cooperation with UNESCO, Canberra 26-31 August 1968*, pages 53–63. Macmillan Company of Australia, South Melbourne, 1968.

[4] P H T Beckett and P A Burrough. The relation between cost & utility in soil survey. IV. Comparison of the utilities of soil maps produced by different survey procedures, and to different scales. *Journal of Soil Science*, 22(4):466–480, 1971.

[5] Y Bishop, S Fienberg, and P Holland. *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, MA, 1975.

[6] D B Boorman, J M Hollis, and A Lilly. Hydrology of soil types: a hydrologically-based classification of the soils of the United Kingdom. Technical Report 126, UK Institute of Hydrology, 1995.

[7] J Bouma, H W G Booltink, and P A Finke. Use of soil survey data for modelling solute transport in the vadose zone. *Journal of Environmental Quality*, 25:519–526, 1996.

[8] G R Brannon and B F Hajek. Update and recorrelation of soil surveys using gis and statistical analysis. *Soil Science Society of America Journal*, 64(2):679–680, 2000.

[9] R L Brennen and D J Prediger. Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41:687–699, 1981.

[10] D J Brus and J J de Gruijter. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma*, 80(1–2):1–59, 1997.

[11] P A Burrough and R A McDonnell. *Principles of geographical information systems*. Oxford University Press, Oxford, 1998.

[12] W G Cochran. *Sampling Techniques*. John Wiley, New York, 3rd edition, 1977.

[13] J Cohen. A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, 20:37–46, 1960.

[14] J Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:426–443, 1968.

[15] R G Congalton. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 54:593–600, 1988.

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

[17] R G Congalton and K Green. *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis, Boca Raton, FL, 1999.

[18] R G Congalton, R G Oderwald, and R A Mead. Assessing landsat classification accuracy using discrete multivariate-analysis statistical techniques. *Photogrammetric Engineering & Remote Sensing*, 49(12):1671–1678, 1983.

[19] R B Daniels, H J Kleiss, S W Buol, H J Byrd, and J A Phillips. *Soil systems in North Carolina*. North Carolina Agricultural Research Service Bulletin 467. North Carolina Agricultural Research Service, Raleigh, NC, 1984.

[20] J A Deckers, F O Nachtergaele, and O C Spaargaren, editors. *World reference base for soil resources : introduction*. ACCO, Leuven, 1998.

[21] D Dent and A Young. *Soil survey and land evaluation*. George Allen & Unwin, London, England, 1981.

[22] C A van Diepen, H van Keulen, J Wolf, and J A A Berkhout. Land evaluation: from intuition to quantification. In B A Stewart, editor, *Advances In Soil Science*, volume 15, pages 139–204. Springer, New York, 1991.

[23] J Dumanski and C Onofrei. Techniques of crop yield assessment for agricultural land evaluation. *Soil Use & Management*, 5(1):9–16, 1989.

[24] FAO. *World reference base for soil resources*. World Soil Resources Report 84. Food and Agriculture Organization of the United Nations, Rome, 1998.

[25] Federal Geographic Data Committee. *Content Standard for Digital Geospatial Metadata*. FGDC-STD-001-1998. U.S. Geological Survey, Reston, VA, 1998.

[26] W Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, New York, 3rd edition, 1968.

[27] J L Fleiss, J Cohen, and B S Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[28] G M Foody. On the compensation for chance agreement in image classification and accuracy assessment. *Photogrammetric Engineering & Remote Sensing*, 58: 1459–1460, 1992.

[29] T R Forbes, D Rossiter, and A Van Wambeke. *Guidelines for evaluating the adequacy of soil resource inventories*. SMSS Technical Monograph 4. Cornell University Department of Agronomy, Ithaca, NY, 1987 printing edition, 1982.

[30] S Gopal and C Woodcock. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering & Remote Sensing*, 60(2):181–188, 1994.

[31] B Gorte. *Probabilistic segmentation of remotely sensed images*. ITC Publication 63. International Institute for Aerospace Survey & Earth Sciences (ITC), Enschede, the Netherlands, 1998.

[32] J J de Gruijter and B A Marsman. Transect sampling for reliable information on mapping units. In D R Nielson and J Bouma, editors, *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*, pages 150–163, Las Vegas, 1984. PUDOC.

[33] J J de Gruijter, D J J Walvoort, and P F M van Gaans. Continuous soil maps - a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, 77:169–195, 1997.

[34] T Hengl. *Improving soil survey methodology using advanced mapping techniques and grid-based modelling*. Msc, International Institute for Aerospace Survey & Earth Sciences (ITC), Enschede, the Netherlands, 2000.

[35] G B M Heuvelink. *Error propagation in quantitative spatial modelling: applications in Geographical Information Systems*. Netherlands Geographical Studies 163. Faculteit Ruimtelijke Wetenschappen, Universiteit Utrecht, Utrecht, 1993.

[36] A F van Holst. Bodemkartering en bodemkaarten. In H de Bakker and W P Locher, editors, *Bodemkunde van Nederland*, volume 2 : Bodemgeografie, pages 85–99. Malmberg, Den Bosch, 2de edition, 1990.

[37] W D Hudson and C W Ramm. Correct formulation of the Kappa coefficient of agreement (in remote sensing). *Photogrammetric Engineering & Remote Sensing*, 53(4):421–422, 1987.

[38] R Ihaka and R Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[39] W Kainz. Logical consistency. In Stephen C Guptill and Joel L Morrison, editors, *Elements of spatial data quality*, pages 109–137. Elsevier (on behalf of the International Cartographic Association), Oxford, 1995.

[40] A A Klingebiel and P H Montgomery. *Land capability classification*. USDA Agricultural Handbook 210. US Government Printing Office, Washington, DC, 1961.

[41] P Lagacherie, P Andrieux, and R Bouzigues. Fuzziness and uncertainty of soil boundaries: from reality to coding in gis. In P A Burrough and A U Frank, editors, *Geographic objects with indeterminate boundaries*, GISDATA 2, pages 275–286. Taylor & Francis, London, 1996.

[42] T M Lillesand and R W Kiefer. *Remote sensing and image interpretation*. John Wiley & Sons, New York, 3rd edition, 1994.

[43] A Lilly, D B Boorman, J M Hollis, P A Finke, J Bouma, and M R Hoosbeek. The development of a hydrological classification of UK soils and the inherent scale changes. *Nutrient Cycling in Agroecosystems*, 50(1–3):299–302, 1998.

[44] Z Ma and R L Redmond. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering & Remote Sensing*, 61(4):435–439, 1995.

[45] B A Marsman and J J de Gruijter. *Quality of soil maps : a comparison of soil survey methods in a sandy area*. Soil survey papers 15. Soil Survey Institute, Wageningen, 1986.

[46] A B McBratney, J J de Gruijter, and D J Brus. Spacial prediction and mapping of continuous soil classes. *Geoderma*, 54:39–64, 1992.

[47] A B McBratney, I O A Odeh, T F A Bishop, M S Dunbar, and T M Shatar. An overview of pedometric techniques for use in soil survey. *Geoderma*, 97(3–4): 293–327, 2000.

[48] N J McKenzie, P E Gessler, P J Ryan, and D A O'Connell. The role of terrain analysis in soil mapping. In J P Wilson and J Gallant, editors, *Terrain analysis : principles and applications*, pages 245–265. Wiley & Sons, New York, 2000.

[49] R A Monserud and R Leemans. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*, 62(4):275–293, 1992.

[50] E Næsset. Conditional tau coefficient for assessment of producer's accuracy of classified remotely sensed data. *ISPRS Journal of Photogrammetry & Remote Sensing*, 51(2):91–98, 1996.

[51] E Næsset. Use of the weighted Kappa coefficient in classification error assessment of thematic maps. *International Journal of Geographical Information Systems*, 10(5):591–603, 1996.

[52] D R Nielson and J Bouma, editors. *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA, Las Vegas, USA 30 November - 1 December 1984*. PUDOC, Wageningen, 1985.

[53] E Pärt-Enander and A Sjöberg. *The MATLAB 5 Handbook*. Addison-Wesley, Harlow, Herts, 1999.

[54] R G Pontius. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, 66(8):1011–1016, 2000.

[55] R Core Development Team. The R project for statistical computing, 04-July-2001 2001. URL: `http://www.r-project.org/`.

[56] D G Rossiter. A theoretical framework for land evaluation (with Discussion). *Geoderma*, 72:165–202, 1996.

[57] D G Rossiter. A compendium of on-line soil survey information: Digital Soil Survey, 25-March-2001 2000. URL: `http://www.itc.nl/~rossiter/research/rsrch_ss_digital.html`.

[58] P A Sánchez, W Couto, and S W Buol. The fertility capability soil classification system: interpretation, applicability and modification. *Geoderma*, 27(4):283–309, 1982.

[59] A K Skidmore. Accuracy assessment of spatial information. In A Stein, F van der Meer, and B G F Gorte, editors, *Spatial statistics for remote sensing*, pages 197–209. Kluwer Academic, Dordrecht, 1999.

[60] C W Smith. *The Fertility Capability Classification System (FCC) - 3rd Approximation: A technical soil classification system relating pedon characterization data to inherent fertility characteristics*. Ph.d., North Carolina State University, 1989.

[61] Soil Classification Working Group. *Soil classification: a taxonomic system for South Africa*. Department of Agricultural Development, RSA, Pretoria, 1991.

[62] Soil Conservation Service. *Land resource regions and Major Land Resource Areas of the United States*. Agriculture Handbook 296. US Government Printing Office, Washington, DC, 1981.

[63] Soil Resources Inventory Group. *Soil resource inventories and development planning: Proceedings of workshops at Cornell University 1977-1978*. Soil Management Support Services (SMSS) Technical Monograph 1. Soil Conservation Service, USDA, Washington, DC, 1981.

[64] Soil Survey Division Staff. *Soil survey manual*. United States Department of Agriculture Handbook No. 18. US Department of Agriculture, Washington, DC, 1993.

[65] Soil Survey Staff. *Soil Taxonomy: a basic system of soil classification for making and interpreting soil surveys*. Agricultural Handbook 436. US Department of Agriculture Soil Conservation Service, Washington, DC, 2nd edition, 1999.

[66] R G D Steel and J H Torrie. *Principles and procedures of statistics: a biometrical approach*. McGraw-Hill, New York, 2nd edition, 1980.

[67] The Math Works. *Using MATLAB Version 6*. The Math Works, Natick, MA, 2000.

[68] R Tortora. A note on sample size estimation for multinomial populations. *American Statistician*, 32(3):100–102, 1978.

[69] U S Department of Agriculture Natural Resources Conservation Service. State soil geographic (statsgo) database for north carolina, 10-July-1995 1995. URL: `http://www.ftw.nrcs.usda.gov/stat_data.html`.

[70] U S Department of Agriculture Natural Resources Conservation Service. Soil Survey Geographic (SSURGO) database for Edgecombe County, North Carolina, 01-December-1998 1998. URL: `http://www.ftw.nrcs.usda.gov/ssur_data.html`.

[71] U S Department of Agriculture Soil Conservation Service. *Soil survey of Seneca Nation of Indians, New York*. Government Printing Office, Washington, DC, 1984.

[72] A Van Wambeke and T Forbes, editors. *Guidlines for using Soil Taxonomy in the names of soil map units*. SMSS Technical Monograph No. 10. Soil Management Support Service, Washington, DC, 1986.

[73] S J Ventura and B J Irvin. Automated landform classification methods for soil-landscape studies. In J P Wilson and J Gallant, editors, *Terrain analysis : principles and applications*, pages 267–294. Wiley & Sons, New York, 2000.

[74] K Verheyen, D Adriaens, M Hermy, and S Deckers. High-resolution continuous soil classification using morphological soil profile descriptions. *Geoderma*, 101: 31–48, 2001.

[75] A P A Vink. *Land use in advancing agriculture*. Springer-Verlag, New York, 1975.

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

[77] R Webster and M A Oliver. *Statistical methods in soil and land resource survey*. Oxford University Press, Oxford, 1990.

[78] L P Wilding. Spatial variability: its documentation, accommodation and implication to soil surveys. In D R Nielson and J Bouma, editors, *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*, pages 166–189, Las Vegas, 1984. PUDOC.

[79] F J Young, R D Hammer, and F Williams. Estimation of map unit composition from transect data. *Soil Science Society of America Journal*, 61(3):854–861, 1997.

[80] J A Zinck and C R Valenzuela. Soil geographic database: structure and application examples. *ITC Journal*, 1990(3):270–294, 1990.

[81] J Alfred Zinck. *Physiography and Soils*. ITC Lecture Notes SOL.41. ITC, Enschede, the Netherlands, 1988.