Measuring Uncertainty in Class Assignment for Natural Resource Maps under Fuzzy Logic

A-Xing Zhu

Abstract

There are two kinds of uncertainty associated with assigning a geographic entity to a class in the classification process. The first is related to the fuzzy belonging of the entity to the prescribed set of classes and the second is associated with the deviation of the entity from the prototype of the class to which the entity is assigned. This paper argues that these two kinds of uncertainty can be estimated if a similarity model is employed in spatial data representation. Under this similarity model, the uncertainty of fuzzy belonging can be approximated by an entropy measure of membership distribution or by a measure of membership residual. The uncertainty associated with the deviation from the prototype definitions can be estimated using a membership exaggeration measure. A case study using a soil map shows that high entropy values occur in areas where soils seem to be transitional and that areas which are mis-classified have higher entropy values. The membership exaggeration is high for areas where soil experts have low confidence in identifying soil types and predicting their spatial distribution. These measures helped in identifying that the high elevation areas were mapped with high accuracy and that error reduction efforts are needed in mapping the soil resource in the low elevation areas.

Introduction

With the increasing popularity of geographic information systems (GIS), geographic data in GIS are often being used to support policy decisions under the assumption that they are free of errors. However, this error-free assumption about geographic data is often not warranted due to a variety of reasons (Goodchild and Gopal, 1989, pp. xii-xiii; Burrough, 1986, pp. 103-135). Errors in geographic data would have a profound impact on the reliability of the resulting policy decision based on GIS analyses because the quality of data affects the quality of decisions and the evaluation of decision alternatives (Barraba, 1989; Anderson and Stewart, 1994).

One must assess the fitness of geographic data being used when deriving policy decisions based on GIS (Agumya and Hunter, 1996; Stanek and Frank, 1993). The first step towards assessing the fitness of geographic data for a specific application is the derivation of data quality information (Hunter and Goodchild, 1993). There are many potential sources of error in spatial data (Hunter and Beard, 1992), and the quality of spatial data can be described by various accuracy elements (Guptill and Morrison, 1995). Goodchild (1995), in his discussion of attribute accuracy, raises the important issue of the spatial structure of error associated with categorical data although his emphasis was on spatial dependence rather than spatial variation in accuracy. Knowledge of spatial variation of data quality can be very useful to users in revealing the areas where the quality meet the needs of the application and in identifying areas where special error reduction efforts must be carried out.

Categorical data sets derived from either remote sensing classification or field surveys are the most widely used GIS data in reaching management decisions and in other GIS-based analyses. Information on the spatial variation of data quality is rarely available for these data sets although global accuracy statistics such as PCC (percent correctly classified), Kappa (see Campbell (1996), pp. 389-392), and RMSE are often provided with the data sets but these global indices give no information on the spatial nature of the classification accuracy (Goodchild, 1995).

This paper examines the types of uncertainty in categorical maps and provides means to measure the spatial distribution of these types of uncertainty. In the next section, two types of errors (omission and commission errors) associated with class assignment in classification are discussed. It is then followed by the discussion of a similarity model based on which uncertainty measures are devised. Three uncertainty measures for estimating the commission and omission errors are discussed. A case study conducted in the Lubretcht Experimental Forest, Montana will be presented to show the usefulness of these measures for depicting the spatial variation of uncertainty in a soil series map of the area. The paper ends with conclusions and summaries.

Class Assignment and Uncertainty

Humans are particularly skilled at distilling structure or patterns from complex reality (Burrough and Frank, 1995). One way of extracting structure from complex data sets is classification through which categorical maps depicting the distribution of spatial phenomena are produced. There are two phases in conventional classification: class definition and class assignment. During class definition, the parameter space of a spatial phenomenon is discretized into regions (classes) with each region assigned a class name and represented by the centroid of that region (Figure 1). The centroid is the central concept of that class. This central concept is often the typical case for this class. It must be noted here that during class definition the multi-dimensional parameter space is being divided into distinct and discrete regions (Figure 1a) and each of these regions is then condensed into a point. This reduction of the parameter space provides the basis for errors to occur in the class assignment phase.

During class assignment, which is often performed under crisp logic, an entity is assigned to one and only one class based on a comparison between the observed attributes of the entity and the typical attributes of prescribed classes.

Department of Geography, University of Wisconsin, 550 North Park Street, Madison, WI 53706 (axing@geography.wisc.edu).

Photogrammetric Engineering & Remote Sensing, Vol. 63, No. 10, October 1997, pp. 1195–1202.

0099-1112/97/6310-1195$3.00/0

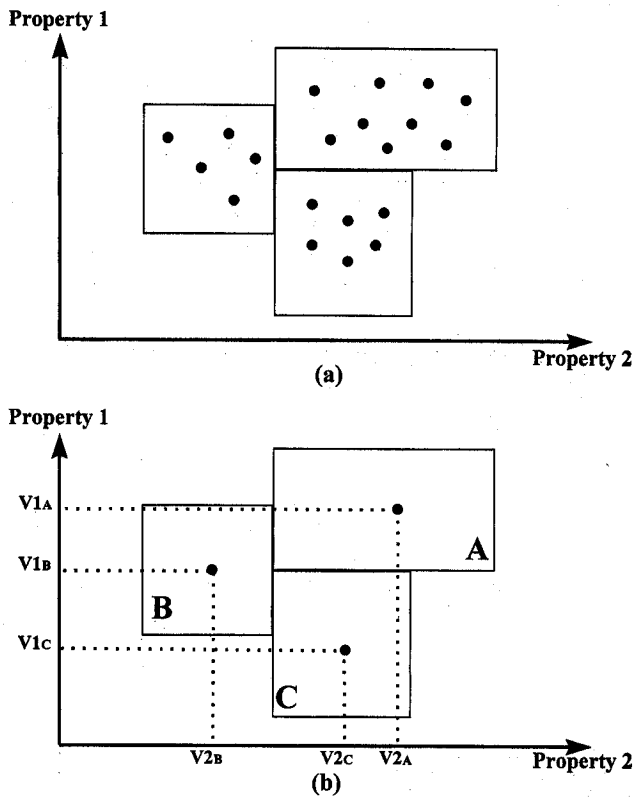

Figure 1. Discretization of parameter space into classes. (a) Division of parameter space. (b) Labeling each region as a class by giving it a class label and class centroid.

The comparison may be performed in terms of distance in parameter space, as is done in minimum-distance-to-means and parallelepiped classifiers, or may be carried out statistically, as in Gaussian maximum-likelihood classification. Once an entity is assigned to a class, it is said to carry the properties of that class, which (the properties) are often represented by the central concept (the centroid) of the class, and the "individuality" of this entity is then lost in this class assignment process. The loss of this individuality is the error introduced into the final product.

The errors introduced in class assignment can be perceived from two perspectives. Let's assume that there are two entities represented as E_1 and E_2, and their positions in relation to the classes in the parameter space are shown in Figure 2. During class assignment, both of these two entities would be assigned to Class A according to Figure 2 (assuming that distances to the class centroids are the basis for classification), and both E_1 and E_2 will carry the properties of Class A in this case. It is apparent that neither E_1 nor E_2 are located in the center of Class A. By assigning these entities to Class A and having them bear the properties of Class A, we ignore the differences between the properties of these entities and the typical conditions of Class A. In this case, we committed a *commission error*, that is, we assigned a label to an entity which does not *fully* "qualify" for it. During class assignment, we also ignore the fact that E_1 also bears resemblance to Classes B and C, and so does E_2 but at a different degree. By ignoring the similarities between an entity and other classes, we committed an *omission error*.

It must be pointed out that the concepts of commission error and omission error outlined here pertain to a single classified entity. It is argued here that any classified entity contains some degrees of commission error and omission error even if the entity is correctly classified unless when the entity is a typical case of a class. Under this notation, entities E_1 and E_2 in our example would both contain commission and omission errors but at different degrees.

The occurrence of these two kinds of error in class assignment is very common in the creation of categorical resource maps because geographic entities rarely conform to the definitions of some discreet and distinctive classes. The properties of geographic entities often vary in some continuous fashion in parameter space. Therefore, it would be difficult or impossible not to commit commission error and omission error in assigning a geographic entity to a class under crisp logic. Furthermore, the degrees of these errors also vary over space because properties of geographic entities often vary continuously over geographic space (Mark and Csillag, 1990). This variation of errors over geographic space makes global accuracy statistics (such as PCC and kappa) inadequate for users to assess the fitness of the data and to allocate error reduction efforts.

As discussed above, the creation of the two types of error in conventional class assignment is due to the direct employment of crisp logic which exaggerates the similarity (membership) between an object and the class to which the object is assigned, and ignores the similarities between the object and the other classes. It would be desirable to provide measures on this membership exaggeration and ignorance so that the magnitude of exaggeration and ignorance can be reported to users. In order to estimate the degree of exaggeration and ignorance, we must know the location of an object in parameter space in relation to the class centers before the object is assigned to any class. This means that we need to obtain the membership value for the object in each of the prescribed classes. Once this set of membership values is obtained, the degrees of exaggeration and ignorance associated with the assignment of the object to a class can then be estimated. In other words, categorical resource map should now be created in two steps: (1) class membership derivation and (2) class labeling and uncertainty estimation.

People may question the necessity of this two-step generation of categorical maps. One may argue that if the memberships in classes are known, then the memberships should

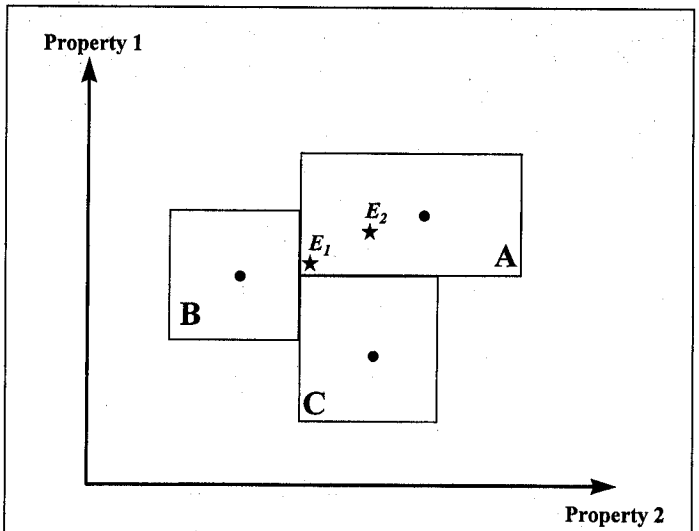

Figure 2. Assigning entities E_1 and E_2 into Class A introduces two kinds of errors: commission error (exaggerating these entities to be the prototype of Class A) and omission error (ignoring the similarity between these entities and other classes, such as Class B and C).

be used instead of class labels, which would prevent loss of information. This is certainly true, but users are not yet used to reading membership maps, particularly when using many of these maps to identify the spatial distribution of certain natural resources. The conventional categorical resource map would be more effective for portraying spatial distributions of natural resources. To this end, one may also argue that if classification is necessary, then the users should just accept loss of information. This has been the case for a long time, particularly when the categorical maps are produced manually, but it would be useful to know how much and where the information loss has occurred. With the use of GIS, one should be able to estimate this loss of information and report it along with the categorical map. Spatial variation of information loss is certainly useful to the end users for assessing the fitness of the categorical map for their applications, particularly when the end users are not involved in the production of the categorical map. For these reasons, it is argued that the two-step process should be employed in generating categorical maps.

The class membership derivation step calls for a spatial model which allows the representation of partial memberships to a set of prescribed classes for every location over the area to be mapped. Once this a model is populated, the variation of exaggeration and ignorance over space can then be derived. The next section presents and discusses such a model.

Similarity Representation of Spatial Phenomena

Many researchers (for example, Burrough (1989), Leung *et al.* (1992), Lowell (1994), McBratney and De Gruijter (1992), Odeh *et al.* (1992), Robinson (1988), and Wang (1990)) have examined the effects and usefulness of fuzzy classification (continuous classification) and found that allowing varying degrees of membership in multiple classes prevents information loss in the classification process. While applying fuzzy logic to soil inference using environmental data, Zhu (1997a) developed a model, called a soil similarity model, to allow the representation of spatial gradation of soils. The model consists of two parts: the similarity representation of soils in parameter space and the raster representation of soils in geographic space.

The similarity representation of soils is based on fuzzy logic under which a soil can be assigned to more than one class with varying degrees of assignment. Under this similarity representation, a soil at point (i,j) can be represented as an n-element vector, called a soil similarity vector, $\mathbf{S}_{ij}$ $(S_{ij}^1, S_{ij}^2, \ldots, S_{ij}^k, \ldots, S_{ij}^n)$, where S_{ij}^k is a measure of the similarity between the local soil at (i,j) and soil category k, and n is the total number of prescribed soil categories. This notation is very similar to the notation of the probability vector described by Goodchild *et al.* (1992). The difference is that S_{ij}^k is a value measuring the similarity between the local soil and the prescribed soil category k, and it is not a probability measuring the chance for the prescribed soil category k to occur at location (i,j).

The elements in a similarity vector do not have to sum up to unity because they are similarity measures (Zhu, 1997a). It is possible that an object has high similarity values to many similar classes and the sum of these similarities can exceed unity while another object may be very unique and it may not bear much similarity to any of the prescribed classes, and the sum of its similarity values can then be less than a unity. It must be emphasized that it is the whole similarity vector which is important, not just the highest membership value in the vector.

Under a raster representation scheme, soils over an area can be represented as an array of soil similarity vectors with each vector corresponding to a pixel in the raster database. Because the whole vector is important in representing a soil in the parameter space, subtle differences in soil between two neighboring pixels can be accommodated by the differences in their respective similarity vectors. By combining this representation power of the similarity vector in the parameter space and the ability of representing high spatial details of a raster database, spatial gradation of soil information can be preserved under this similarity model.

Apparently, this model can be easily used for representing spatial variation of other natural resources. The key to apply this model is the derivation of membership values in these vectors. There are many techniques which can be used to populate this model. A detailed discussion of these techniques is beyond the scope of this paper. Some of examples are cited as follows. Zhu *et al.* (1996) developed an approach to populate the similarity model using expert system development techniques and fuzzy mathematics. Neural network classification (for example, Civco (1993), Foody (1996), and Gong *et al.* (1996)) and fuzzy classification (for example, Bezdek *et al.* (1984), Gopal and Woodcock (1994), McBratney and De Gruijter (1992), Odeh *et al.* (1992), and Wang (1990)) can also be used to compute the similarity vectors. Many classification techniques used in remote sensing applications often compute "membership" of an object in classes. These techniques can be used to populate the similarity model with little modifications.

Estimating Uncertainty under the Similarity Model

Once the similarity model for a given resource is populated, a categorical map of the resource can be derived and spatial variation of class assignment errors in that categorical resource map can be estimated. The generation of a categorical map is done by converting the similarity vector for each pixel into a class label. The meaningful way is to use the class which has the highest membership value in the vector to represent the local object. This process is referred to as the hardening process, which is similar to the crisp class assignment discussed earlier. The result from this hardening process is a raster map with each pixel being labeled as the class to which the object is assigned.

As discussed earlier, there are two types of errors involved in the hardening process for each pixel: commission error and omission error. The manifestation of these two types of error and their associated uncertainty can be understood through the following example. Let us say there are two objects at locations P_1 and P_2. The vectors describing the objects in relation to some classes at these two points are V_1(0.2, 0.25, 0.3, 0.25) for P_1 and V_2(0.1, 0.0, 0.85, 0.05) for P_2. The elements in the vectors are the membership values to class A, B, C, and D, respectively. If both these vectors are hardened, the objects at these two points will be labeled as Class C. The omission error can be perceived as the degree of ignoring membership values in the vector other than these corresponding to Class C and the commission error would be the exaggeration of the objects to have full membership in Class C.

The term "error" often implies the difference between an obtained value and the corresponding truth. Because the membership values in the vector may very well be an approximation to the real similarity values between an object and the prescribed classes and we may never know the real similarity values, the term "uncertainty" is used here instead of "error." By the definition of the two types of error, the uncertainty associated with omission error is referred to as *ignorance uncertainty* and the uncertainty due to commission error is referred to as *exaggeration uncertainty.*

Ignorance Uncertainty

Ignorance uncertainty is related to the fuzziness of an object compared with the definitions of classes. In other words, it is related to membership diffusion. The fuzzier the object in

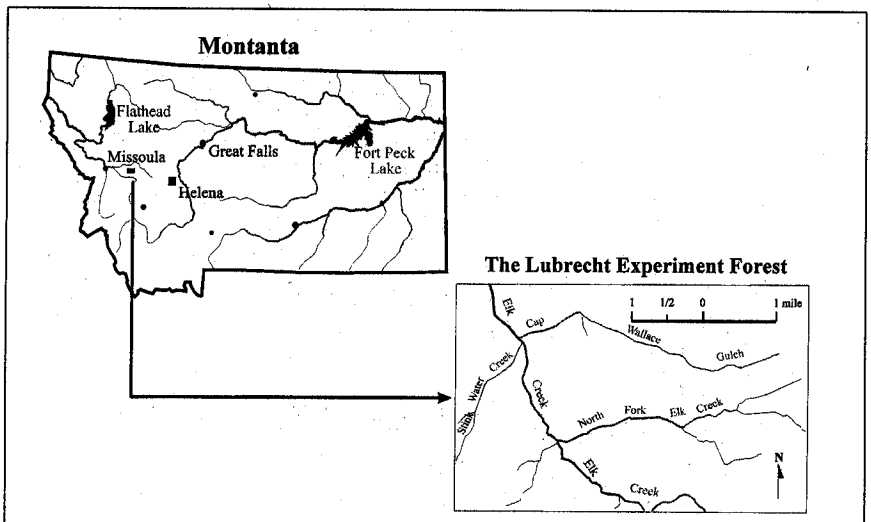

Figure 3. Location of the Lubrecht Experiment Forest, Montana.

relation to the classes, the more evenly distributed the membership in the vector, and the greater is the ignorance uncertainty. Because the membership values are more evenly distributed in V_1 than in V_2, the ignorance error associated with the hardening of V_1 is greater. Ignorance uncertainty can be estimated using an entropy measure (Goodchild *et al.*, 1994; Zhu, 1996), which expresses the degree to which membership is concentrated in a particular class, rather than spread over a number of classes. The entropy measure can be calculated as follow:

$$H_{ij} = \frac{1}{\log_e n} \sum_{k=1}^{n} [(S'^{k}_{ij}) \log_e (S'^{k}_{ij})] \quad (1)$$

where $S'_{ij}{}^{k}$ is the normalized similarity value (the sum of the normalized values in a vector is 1.0). H_{ij} is the entropy associated with the similarity vector for point (i,j) and has a range of 0 to 1. An entropy value of 0 means that the object at a given pixel has full membership in one of the prescribed classes and 0 membership in all other classes. In other words, the local object is the prototype of a class. Because the membership values in all other classes are 0 ($H_{ij} = 0$), there is no membership ignorance involved in the hardening process and the ignorance uncertainty for this pixel is then 0. When the entropy value of a similarity vector reaches 1, it means the object is similar to all classes at the same degree and none of the classes would be a good representative for this object. Assigning the object to anyone of these classes would induce the highest degree of membership ignorance and the ignorance uncertainty for this pixel would then be 1.

It should be realized that the entropy measure is insensitive to the class assignment. The entropy about a similarity vector remains the same whether or not the object is assigned to the class with the highest similarity value in the vector. Because the hardening process always assigns an object to the class with the highest membership value in the vector, the entropy statistics can be used to measure the dispersion of membership from this class and it therefore can be considered as a good index for measuring ignorance uncertainty in the hardening process. The magnitude of an entropy value also depends on number of classes, whether or not these classes are relevant to the object. The more the classes, the smaller the value is. It is recommended that the relative magnitude of an entropy value is more informative than the absolute one in interpreting entropy as the ignorance uncertainty.

Another index for estimating membership ignorance is the membership residual after the highest. This measure can be expressed as

$$\Delta M_{ij} = (1 - S_{ij}{}'^{g}) \quad (2)$$

where ΔM_{ij} is the membership residual and $S_{ij}{}'^{g}$ is the normalized similarity value for class g to which the object is assigned. The difference between ΔM_{ij} and H_{ij} is that ΔM_{ij} is sensitive to class assignment but insensitive to the distribution of membership residual and number of classes in the vector.

Exaggeration Uncertainty

Exaggeration uncertainty is inversely related to the membership saturation in the class to which an object is assigned. Obviously, the higher the membership in the assigned class, the less the exaggeration. For example, membership exaggeration in assigning the object at P_2 to Class C is much less that for the object at P_1. A simple measure of exaggeration uncertainty is

$$E_{ij} = (1 - S^{g}_{ij}) \quad (3)$$

where E_{ij} is the exaggeration uncertainty measure and $S_{ij}{}^{g}$ is the similarity measure between the object at (i,j) and class g to which the object is assigned. Equation 3 is different from Equation 2 although they look alike. $S_{ij}{}^{g}$ is a similarity measure expressing the object's membership saturation in Class g and is not related to other similarity values in the vector while $S'_{ij}{}^{g}$ is a normalized value expressing the relative importance to other values in the vector. Equation 3 should be applied to a vector which is not normalized. Otherwise, information about exaggeration is lost through the normalization process.

Case Study Results and Discussions

Study Area

The study area is the Lubrecht Experiment Forest located about 50 km northeast of Missoula, Montana (Figure 3). The Forest was established in 1937 to foster research on natural resources. The elevation in the area ranges from about 1,200 m to about 2,000 m, with high elevation in the northeast and southwest and low elevation in the northwest (Figure 4). The area is considered as a semi-humid to semi-arid region with strong moisture contrasts between low elevation regions and

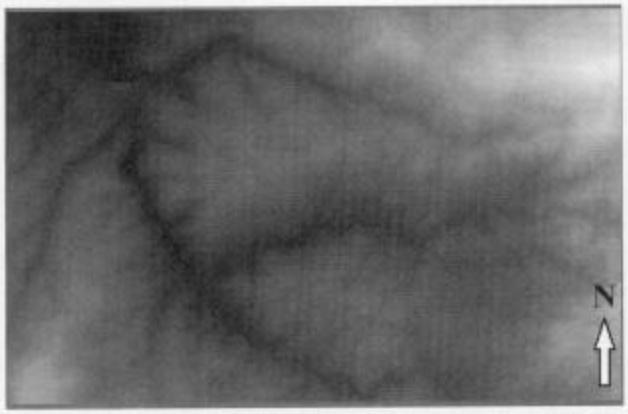

Figure 4. Digital elevation model of the study area (light tones mean high elevations).

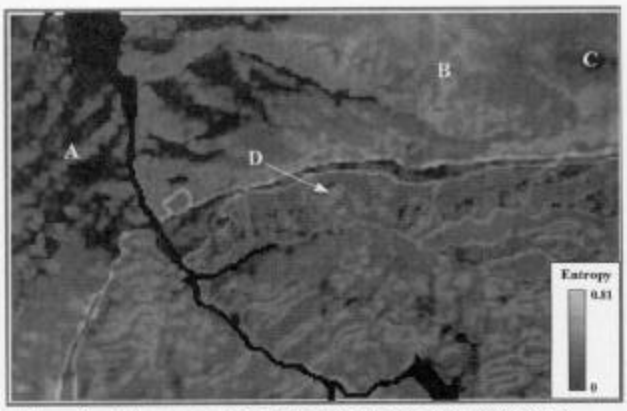

Figure 6. Entropy map for the Lubrecht area with light tones indicating high entropy values (the black areas along Elk Creek were not included in this study).

high elevation areas, and between south-facing slopes and north-facing slopes (Nimlos, 1986).

Derivation of Membership Vectors and Generation of Soil Series Map

Zhu *et al.* (1996) developed a knowledge-based inference approach for populating the soil similarity model. The approach was based on the Jenny's classic concept that soil is a result of the interaction among soil forming factors (Jenny, 1980). The details of this approach are beyond the scope of this paper. In general, they employed a set of knowledge elicitation techniques to extract soil scientists' knowledge on soil-environment relationships and used a set of GIS techniques to characterize the soil formative environment. The extracted knowledge was then combined with the spatial information on a soil formative environment under a set of fuzzy inference techniques to derive soil similarity vectors over the study area. Zhu *et al.* (1997) explored the use of the similarity vectors for deriving continuous soil property maps and found that the resultant soil information had a higher quality at both the spatial and attribute levels than that in the conventional soil map.

The hardening process outlined earlier was used to convert the similarity vectors over the Lubrecht area to a soil series map (Figure 5). For each pixel, the proposed uncertainty indices were computed from the membership vector of that pixel. Three images of uncertainty were produced, with one for each of the three indices.

Ignorance Uncertainty

Figures 6 and 7 show the spatial variation of ignorance uncertainty estimated by the entropy measure and the membership residual measure. One can conclude that these two images are essentially the same. The only difference is that the uncertainty values at the boundaries of soil bodies on the membership residual image are higher than corresponding values on the entropy image. This is due to the fact that only the total residual is used in calculating the membership residual index, and the distribution of this residual membership in the vector was not considered. Because the two images are essentially the same, discussion of ignorance uncertainty is focused on the entropy measure. The validity of the entropy measure on ignorance uncertainty is evaluated in two aspects: the point accuracy and the spatial patterns of ignorance uncertainty revealed by this measure.

Point Accuracy of the Entropy Measure

The point accuracy of the entropy measure can be assessed using hypothesis testing. If the entropy calculated from a similarity vector is a useful measure of ignorance uncertainty

Figure 5. Soil series map from hardening the similarity vectors (the black areas along Elk Creek were not included in this study).

Figure 7. Distribution of membership residuals. This image depicts the spatial pattern of uncertainty similar to that in Figure 6.

Figure 4. Digital elevation model of the study area (light tones mean high elevations).

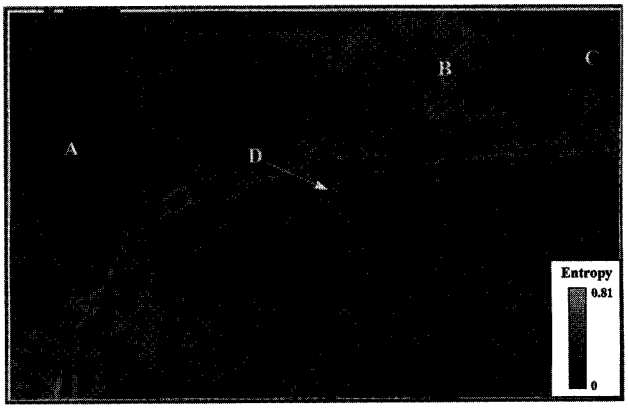

Figure 6. Entropy map for the Lubrecht area with light tones indicating high entropy values (the black areas along Elk Creek were not included in this study).

high elevation areas, and between south-facing slopes and north-facing slopes (Nimlos, 1986).

Derivation of Membership Vectors and Generation of Soil Series Map

Zhu *et al.* (1996) developed a knowledge-based inference approach for populating the soil similarity model. The approach was based on the Jenny's classic concept that soil is a result of the interaction among soil forming factors (Jenny, 1980). The details of this approach are beyond the scope of this paper. In general, they employed a set of knowledge elicitation techniques to extract soil scientists' knowledge on soil-environment relationships and used a set of GIS techniques to characterize the soil formative environment. The extracted knowledge was then combined with the spatial information on a soil formative environment under a set of fuzzy inference techniques to derive soil similarity vectors over the study area. Zhu *et al.* (1997) explored the use of the similarity vectors for deriving continuous soil property maps and found that the resultant soil information had a higher quality at both the spatial and attribute levels than that in the conventional soil map.

The hardening process outlined earlier was used to convert the similarity vectors over the Lubrecht area to a soil series map (Figure 5). For each pixel, the proposed uncertainty indices were computed from the membership vector of that pixel. Three images of uncertainty were produced, with one for each of the three indices.

Ignorance Uncertainty

Figures 6 and 7 show the spatial variation of ignorance uncertainty estimated by the entropy measure and the membership residual measure. One can conclude that these two images are essentially the same. The only difference is that the uncertainty values at the boundaries of soil bodies on the membership residual image are higher than corresponding values on the entropy image. This is due to the fact that only the total residual is used in calculating the membership residual index, and the distribution of this residual membership in the vector was not considered. Because the two images are essentially the same, discussion of ignorance uncertainty is focused on the entropy measure. The validity of the entropy measure on ignorance uncertainty is evaluated in two aspects: the point accuracy and the spatial patterns of ignorance uncertainty revealed by this measure.

Point Accuracy of the Entropy Measure

The point accuracy of the entropy measure can be assessed using hypothesis testing. If the entropy calculated from a similarity vector is a useful measure of ignorance uncertainty

Figure 5. Soil series map from hardening the similarity vectors (the black areas along Elk Creek were not included in this study).

Figure 7. Distribution of membership residuals. This image depicts the spatial pattern of uncertainty similar to that in Figure 6.

TABLE 1. STATISTICS OF ENTROPY FOR THE MATCHED AND MISMATCHED SITES

Statistics	Matched Sites	Mismatched Sites
Mean	0.54	0.64
Variance	0.047	0.049
No. of Sites	38	26

in the hardening process, it can then be assumed that a local soil with a similarity vector whose entropy is high would have a high probability to be mis-classified. This is because a local soil highly similar to two or more soil series would have a high probability of not being classified, in the field, as the soil series for which the computed similarity value is the highest. If the above assumption is true, then we should be able to observe that the mean entropy value for all of the mis-classified pixels is higher than the mean entropy value for the correctly classified pixels.

To test the above hypothesis, the soil series at 64 field sites was determined over the summers of 1991, 1992, and 1993. The sites were distributed over the entire study area. The coordinates of these sites were determined using a Global Positioning System (GPS) receiver and USGS 1:24,000-scale topographic maps of the area. The inferred soil series for these sites was determined from the hardened soil series map. These inferred soil series were then compared to the field observed soil series. The 64 sites were divided into two groups. The first group (the matched group) contains all of the sites whose inferred soil series match the observed soil series and the second group (the mismatched group) contains the rest. The entropy values for all 64 sites were obtained from the entropy map. The mean and standard deviation of entropy for each of the two groups were calculated and are reported in Table 1.

The null hypothesis (H_o) is that the mean entropy value for the mismatched sites is equal to or smaller than the mean value for the matched sites and the alternate hypothesis (H_A) is that the mean value for the mismatched sites is larger than that for the matched sites. A student-t test with the assumption that the population variances of the two groups of sites are unequal was used to test the hypothesis (Burt and Barber, 1996, p. 314). The calculated t value is 1.727 with a degree of freedom of 53. The critical t value greater than that for which the null hypothesis can be rejected with 95 percent of confidence at the degree of freedom of 40 is 1.684. Because the calculated t with a degree of freedom of 53 is greater than the critical t value, the above H_0 can then be rejected at 95 percent of confidence. This indicates that, statistically speaking, the mean entropy value for the mismatch sites is greater than that for the matched sites. It can then be concluded that the entropy is a useful measure of ignorance uncertainty in this case.

Spatial Patterns of Ignorance Uncertainty

Comparing Figures 5 and 6, one can observe that entropy values are high in the mid-elevation areas (Area B in Figure 6) where soils are transitional to the soil series prescribed for the low elevations (Area A) and those for the high elevations (Area C). Assigning these transitional soils to any prescribed soil series would imply that these soils can be treated as the same as the prototypes of the prescribed soil series. This implication could result in misuses of these soil resources because managerial practices on these transitional soils may have to be substantially different from those on the prescribed soil series. The high entropy values in these areas indicate that the above implication is incorrect and other managerial measures may have to be applied in using the soil resource over these transitional areas.

Another observation which can be made from Figures 5 and 6 is the high entropy values at the fringe of a soil body. This is not difficult to understand because the soils at the boundary areas often bear high similarities to many different soil types. Assigning these transitional soils to any single soil category would introduce high degrees of uncertainty, too.

The third observation that can be made about Figure 6 is that at low elevations (Areas around A) the entropy values seem to be higher on the south-facing slopes than those on the north-facing slopes. This can be explained by two factors. First, because the study area is in a semi-arid to semi-humid region, moisture condition is the dominant factor during the soil forming process. At low elevation, the moisture condition on the north-facing slopes is more spatially homogeneous than that on the south-facing slopes where subtle changes in slope aspect would result in a significant difference in moisture conditions. Therefore, the soils on the north-facing slopes are spatially more contiguous than those on the south-facing slopes where different soils are often intermittently distributed (D in Figure 6). Second, because the soils on the north-facing slopes are more spatially contiguous, it is much easier for soil experts to relate these soils to environmental conditions during the knowledge acquisition process (Zhu, 1997b). On the other hand, soil experts would have a low confidence in relating the soils on the south-facing slopes to environmental conditions because the different soil types are so intermittently distributed on these slopes. Therefore, it is not surprising to find that the soils on the north-facing slopes are mapped with a higher confidence than those on the south-facing slopes.

Exaggeration Uncertainty

The image of membership exaggeration (Figure 8) shows a spatial pattern very different from that of ignorance uncertainty. Membership exaggeration is very high for areas at low elevations, particularly for areas with south-facing slopes. This observation means that the local soils at low elevation bear low similarities to the assigned soil series, although that similarity values are the largest in their respective vectors. This situation can be created under two scenarios. First, the local soil is really different from any of the prescribed soil series; therefore, it bears low similarity values to all of the prescribed soil series. The second scenario is that the soil similarity vector is not an accurate representation of the local soil, and the low similarity values are the result of low confidence of soil experts in mapping the soils in these areas. In this study, the latter would be a more proper expla-

Figure 8. Spatial distribution of membership exaggeration in class assignment (light tones indicate high degrees of membership exaggeration).

TABLE 1. STATISTICS OF ENTROPY FOR THE MATCHED AND MISMATCHED SITES

Statistics	Matched Sites	Mismatched Sites
Mean	0.54	0.64
Variance	0.047	0.049
No. of Sites	38	26

in the hardening process, it can then be assumed that a local soil with a similarity vector whose entropy is high would have a high probability to be mis-classified. This is because a local soil highly similar to two or more soil series would have a high probability of not being classified, in the field, as the soil series for which the computed similarity value is the highest. If the above assumption is true, then we should be able to observe that the mean entropy value for all of the mis-classified pixels is higher than the mean entropy value for the correctly classified pixels.

To test the above hypothesis, the soil series at 64 field sites was determined over the summers of 1991, 1992, and 1993. The sites were distributed over the entire study area. The coordinates of these sites were determined using a Global Positioning System (GPS) receiver and USGS 1:24,000-scale topographic maps of the area. The inferred soil series for these sites was determined from the hardened soil series map. These inferred soil series were then compared to the field observed soil series. The 64 sites were divided into two groups. The first group (the matched group) contains all of the sites whose inferred soil series match the observed soil series and the second group (the mismatched group) contains the rest. The entropy values for all 64 sites were obtained from the entropy map. The mean and standard deviation of entropy for each of the two groups were calculated and are reported in Table 1.

The null hypothesis (H_0) is that the mean entropy value for the mismatched sites is equal to or smaller than the mean value for the matched sites and the alternate hypothesis (H_A) is that the mean value for the mismatched sites is larger than that for the matched sites. A student-t test with the assumption that the population variances of the two groups of sites are unequal was used to test the hypothesis (Burt and Barber, 1996, p. 314). The calculated t value is 1.727 with a degree of freedom of 53. The critical t value greater than that for which the null hypothesis can be rejected with 95 percent of confidence at the degree of freedom of 40 is 1.684. Because the calculated t with a degree of freedom of 53 is greater than the critical t value, the above H_0 can then be rejected at 95 percent of confidence. This indicates that, statistically speaking, the mean entropy value for the mismatch sites is greater than that for the matched sites. It can then be concluded that the entropy is a useful measure of ignorance uncertainty in this case.

Spatial Patterns of Ignorance Uncertainty

Comparing Figures 5 and 6, one can observe that entropy values are high in the mid-elevation areas (Area B in Figure 6) where soils are transitional to the soil series prescribed for the low elevations (Area A) and those for the high elevations (Area C). Assigning these transitional soils to any prescribed soil series would imply that these soils can be treated as the same as the prototypes of the prescribed soil series. This implication could result in misuses of these soil resources because managerial practices on these transitional soils may have to be substantially different from those on the prescribed soil series. The high entropy values in these areas indicate that the above implication is incorrect and other managerial measures may have to be applied in using the soil resource over these transitional areas.

Another observation which can be made from Figures 5 and 6 is the high entropy values at the fringe of a soil body. This is not difficult to understand because the soils at the boundary areas often bear high similarities to many different soil types. Assigning these transitional soils to any single soil category would introduce high degrees of uncertainty, too.

The third observation that can be made about Figure 6 is that at low elevations (Areas around A) the entropy values seem to be higher on the south-facing slopes than those on the north-facing slopes. This can be explained by two factors. First, because the study area is in a semi-arid to semi-humid region, moisture condition is the dominant factor during the soil forming process. At low elevation, the moisture condition on the north-facing slopes is more spatially homogeneous than that on the south-facing slopes where subtle changes in slope aspect would result in a significant difference in moisture conditions. Therefore, the soils on the north-facing slopes are spatially more contiguous than those on the south-facing slopes where different soils are often intermittently distributed (D in Figure 6). Second, because the soils on the north-facing slopes are more spatially contiguous, it is much easier for soil experts to relate these soils to environmental conditions during the knowledge acquisition process (Zhu, 1997b). On the other hand, soil experts would have a low confidence in relating the soils on the south-facing slopes to environmental conditions because the different soil types are so intermittently distributed on these slopes. Therefore, it is not surprising to find that the soils on the north-facing slopes are mapped with a higher confidence than those on the south-facing slopes.

Exaggeration Uncertainty

The image of membership exaggeration (Figure 8) shows a spatial pattern very different from that of ignorance uncertainty. Membership exaggeration is very high for areas at low elevations, particularly for areas with south-facing slopes. This observation means that the local soils at low elevation bear low similarities to the assigned soil series, although that similarity values are the largest in their respective vectors. This situation can be created under two scenarios. First, the local soil is really different from any of the prescribed soil series; therefore, it bears low similarity values to all of the prescribed soil series. The second scenario is that the soil similarity vector is not an accurate representation of the local soil, and the low similarity values are the result of low confidence of soil experts in mapping the soils in these areas. In this study, the latter would be a more proper expla-

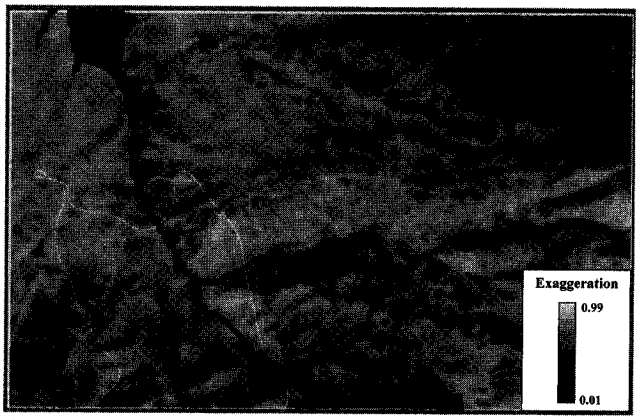

Figure 8. Spatial distribution of membership exaggeration in class assignment (light tones indicate high degrees of membership exaggeration).

nation. In this semi-arid to semi-humid region, the low elevation areas, particularly these with south-facing slopes, are more susceptible to moisture stress than the high elevation areas due to higher temperatures and less precipitation at low elevations. The soils on these slopes are not as well developed as those at higher elevations and they are highly variable spatially. While using environmental indices to derive soil similarity vectors (Zhu *et al.*, 1996), soil experts would be very conservative in giving high similarity values to the soils in these areas. On the other hand, the soils at high elevations are better developed and more spatially contiguous, and soil experts understand the relationships between these soils and their environment better (Zhu, 1997b). Therefore, the experts would be more confident in mapping the soils in these high elevation regions.

Quality of the Resultant Soil Series Map

Through the examination of and discussion on the spatial distribution of the uncertainty associated with the soil series map, the spatial patterns of quality of this soil series map can be understood. A user can now clearly see that the soils at high elevations are mapped with higher accuracy because both ignorance and exaggeration uncertainty are low over these areas. The middle elevation areas are the ecotone between the soil series prescribed for high elevations and those for low elevations. Managerial practice for the soils over these areas may have to be different from that for any of the prescribed soil series. The problem areas are the low elevations, particularly those with south-facing slopes where both uncertainties are very high. Error reduction efforts are recommended in mapping the soil resources in these areas.

Summary

This paper outlined and discussed the uncertainty due to membership diffusion and membership exaggeration during class assignment in generating categorical resource maps. It is recommended that categorical resource maps should be created in two steps: (1) generating a similarity representation of the natural resource to be mapped and (2) hardening this similarity representation to create the categorical resource map and deriving images depicting the spatial variation of uncertainty in the so-derived categorical map. Three uncertainty measures computed from the similarity representation are devised: entropy, membership residual, and membership exaggeration. Entropy and membership residual measures are for estimating uncertainty due to membership ignorance (omission error), and the spatial variation portrayed by these two indices are often very similar for categorical maps derived through the hardening process. Membership exaggeration measures the degree of exaggerating partial membership of an object to full membership in the assigned class (commission error).

In a case study of mapping soil resource using a similarity model, uncertainty images derived using these measures helped to identify areas of high accuracy and areas of potential problems on the resultant soil map. It was also found that, in this case study, the mean entropy value for the misclassified soil sites was larger than that for the correctly classified sites. It is concluded here that the two-step resource map generation based on the similarity representation is advantageous because it allows information regarding spatial patterns of uncertainty associated with the categorical map to be derived. It is also concluded that the measures for estimating the uncertainty related to membership ignorance and exaggeration are meaningful.

Although these uncertainty measures were meaningful in depicting the spatial variation of uncertainty in this case study, the usefulness of these measures depends on the quality of the membership values in the similarity vectors. If the membership values are not good approximations to reality, then these measures would not be providing users any useful information. Clearly, any potential advantage of using the similarity model depends on the quality of the membership values populating it. The quality of the membership values in turn depends on the methodologies and the information used to generate these membership values.

Acknowledgments

This research was supported by a startup grant from the Graduate School at the University of Wisconsin-Madison.

References

Agumya, A., and G.J. Hunter, 1996. Assessing fitness for use of spatial information: Information usage and decision uncertainty, *Proceedings of GIS/LIS'96*, Denver, Colorado, pp. 349–359.

Anderson, J.A., and A.R. Stewart, 1994. Local government liability for erroneous data: Law and policy in a changing environment, *Proceedings of the Conference on Law and Information Policy for Spatial Databases*, Tempe, Arizona, pp. 267–79.

Barraba, V., 1989. Keynote address, *Proceedings of Specialist Meeting on NCGIA Research Initiative 4: Use and Value of Geographic Information*, Technical Report 89-7, National Center for Geographical Information and Analysis, University of California, Santa Barbara, pp. 59–74.

Bezdek, J.C., R. Ehrlich, and W. Full, 1984. FCM: the fuzzy c-means clustering algorithm, *Computers & Geosciences*, 10:191–203.

Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment*, Oxford University Press, Oxford, 194 p.

———, 1989. Fuzzy mathematical methods for soil survey and land evaluation, *Journal of Soil Science*, 40:477–492.

Burrough, P.A., and A.U. Frank, 1995. Concepts and paradigms in spatial information: Are current geographical information systems truly generic? *International Journal of Geographical Information Systems*, 9(2):101–116.

Burt, J., and G. Barber, 1996. *Elementary Statistics for Geographers*, The Guilford Press, New York, 640 p.

Campbell, J.B., 1996. *Introduction to Remote Sensing*, The Guilford Press, New York, 622 p.

Civco, D.L., 1993. Artificial neural networks for land-cover classification and mapping, *International Journal of Geographical Information Systems*, 7(2):173–186.

Foody, G.M., 1996. Relating the land-cover composition of mixed pixels to artificial neural network classification output, *Photogrammetric Engineering & Remote Sensing*, 62(5):491–499.

Gong, P., R. Pu, and J. Chen, 1996. Mapping ecological land systems and classification uncertainties from digital elevation and forest-cover data using neural networks, *Photogrammetric Engineering & Remote Sensing*, 62(11):1249–1260.

Goodchild, M.F., 1995. Attribute accuracy, *Elements of Spatial Data Quality* (S.C. Guptill and J.L. Morrison, editors), Pergamon, Oxford, pp. 59–79.

Goodchild, M.F., and S. Gopal (editors), 1989. *Accuracy of Spatial Databases*, Taylor and Francis, New York, 290 p.

Goodchild, M.F., G. Sun, and S. Yang, 1992. Development and test of an error model for categorical data, *International Journal of Geographical Information Systems*, 6:87–104.

Goodchild, M.F., L. Chin-Chang, and Y. Leung, 1994. Visualizing fuzzy maps, *Visualization in Geographical Information Systems* (H.M. Hearnshaw and D.J. Unwin, editors), John Wiley & Sons, New York, pp. 158–167.

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

Guptill, S., and J. Morrison (editors), 1995. *Elements of Spatial Data Quality*, Pergamon, London, 250 p.

Hunter, G.J., and K. Beard, 1992. Understanding error in spatial databases, *The Australian Surveyor*, 37(2):108–119.

Hunter, G.J., and M.F. Goodchild, 1993. Managing uncertainty in

spatial databases: Putting theory into practice, *Journal of the Urban and Regional Information Systems Association*, 5(2):55–62.

Jenny, H., 1980. *The Soil Resource: Origin and Behaviour*, Springer-Verlag, New York, 377 p.

Leung, Y., M.F. Goodchild, and C.C. Lin, 1992. Visualization of fuzzy scenes and probability fields, *Proceedings of the 5th International Symposium on Spatial Data Handling*, Charleston, South Carolina, 1:480–490.

Lowell, K., 1994. An uncertainty-based spatial representation for natural resources phenomena, *Advances in GIS Research: Proceedings of the Sixth International Symposium on Spatial Data Handling* (T.C. Waugh and R.G. Healey, editors), Taylor & Francis, London, pp. 933–944.

Mark, D.M., and F. Csillag, 1990. The nature of boundaries on 'area-class' maps, *Cartographica*, 27:65–78.

McBratney, A.B., and J.J. De Gruijter, 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades, *Journal of Soil Science*, 43:159–175.

Nimlos, J.T., 1986. *Soils of Lubrecht Experimental Forest*, Miscellaneous Publication No. 44, Montana Forest and Conservation Experiment Station, Missoula, Montana, 36 p.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough, 1992. Soil pattern recognition with fuzzy-c-means: Application to classification and soil landform interrelationships, *Soil Science Society of America Journal*, 56:505–516.

Robinson, V.B., 1988. Some implications of fuzzy set theory applied to geographic databases, *Comput. Environ. and Urban Systems*, 22:010-1–010-9.

Stanek, H., and A.U. Frank, 1993. GIS-based decision making must consider data quality, *Proceedings of the Fourth European Conference and Exhibition on Geographical Information Systems (EGIS'93)*, Genoa, Italy, pp. 685–692.

Wang, F., 1990. Improving remote sensing image analysis through fuzzy information representation, *Photogrammetric Engineering & Remote Sensing*, 56(8):1163–1169.

Zhu, A.X., 1996. Generating an uncertainty map of soil spatial information, *Proceedings of GIS/LIS'96*, Denver, Colorado, pp. 895–905.

———, 1997a. A similarity model for representing soil spatial information, *Geoderma*, in press.

———, 1997b. A personal construct-based knowledge acquisition process for natural resource mapping under fuzzy logic, submitted to *International Journal of Geographical Information Systems*.

Zhu, A.X., L.E. Band, B. Dutton, and T. Nimlos, 1996. Automated soil inference under fuzzy logic, *Ecological Modeling*, 90:123–145.

Zhu, A.X., L.E. Band, R. Vertessy, and B. Dutton, 1997. Derivation of soil properties using a soil land inference model (SoLIM), *Soil Science Society of America Journal*, 61:523–533.