VISUALIZATION OF FUZZY SCENES AND PROBABILITY FIELDS

Yee Leung
Department of Geography, and
Center for Environmental Studies
Chinese University of Hong Kong
Shatin, NT, Hong Kong

Michael F. Goodchild and Chih-Chang Lin
National Center for Geographic Information and Analysis, and
Department of Geography
University of California, Santa Barbara, CA 93106, USA
good@ncgia.ucsb.edu

Abstract

Fuzzy classification has the potential to yield richer information from remotely sensed images, but there have been few efforts to deal with the issues involved in working with fuzzy classifications in GIS. Analogous data are also obtained when the multinomial classification given to land is treated as mixed, fuzzy or probabilistic. The paper reports on a series of efforts to develop visualization techniques for such data. To support visualization of the inherent variability in such data, and to propagate uncertainty effectively through GIS operations, it is necessary to introduce the concept of an error model as a stochastic process, and to define a method for creating individual realizations of that process.

Introduction

Increasing emphasis on analysis, modeling and decision support within the GIS applications community in recent years has led to a general concern for issues of data quality. If the purpose of spatial data handling is to make maps, then perhaps it is sufficient to require merely that the output map product be as accurate as the input. But the detailed analytic and modeling applications that underlie much of the recent literature of GIS (Tomlin, 1991; Laurini and Thompson, 1992) demand much more stringent and robust approaches. If the input is known to be inaccurate, uncertain or error-prone, then it is important that the effects of such inaccuracies on the output also be known. Without such knowledge, the apparent value of GIS in supporting spatial decision-making may be illusory.

In this paper we take the position that all geographic information is inaccurate to some degree, because it is impossible to represent the continuous variation of the Earth's surface perfectly in the finite, discrete space of a digital store. We use the term 'accuracy' generically, and assume that it subsumes error from a variety of sources: uncertainty of definition, imperfect replication between observers making subjective judgments, the consequences of mixed pixels in remote sensing, digitizing error *etc*. A spatial database is a representation of geographical reality in digital form, and the output of a GIS process is an estimate of the results of making an equivalent measurement on the ground. In that sense, accuracy in spatial data handling is a measure of the difference between the digital estimate and ground truth. In cases where ground truth is poorly defined, we include variation between observers in this definition of accuracy. Thus a failure of different observers to agree on the class of land cover at a point contributes to the inaccuracy of land cover data.

Although inaccuracy is pervasive in spatial data, some types are clearly less accurate than others. A GPS survey provides known levels of positional accuracy, down to millimeter levels. We focus in this paper on a class of data known to be subject to relatively high levels of uncertainty, and for which there are no such straightforward measures of accuracy. In this class, every point on the plane is characterized by a single value measured on a nominal scale; examples include soil class, land cover class, and land use. We refer to this as a multinomial field. Two data models are commonly used to build digital representations of such fields. The first, the raster model, is used when the field is obtained by remote sensing, by making use of one of a number of standard procedures for classification. In this model, all information on within-pixel variability is lost. The second, or polygon model, partitions the plane into a number of polygons of homogeneous class, thus losing all variability within polygons. The polygon model is also commonly used in mapping multinomial fields, although the boundary lines on such maps are drawn as continuous curves and need not be discretized to polygons.

Both models are clearly approximations, and although both are in common use, it is rare for the degree of approximation to be made explicit in either case, or for uncertainty to be propagated through GIS processes. In the raster case, fuzzy classifiers provide one way of describing uncertainty, by associating each pixel not with a single class, but with a vector of class memberships, each one interpreted as a measure of belonging. Thus pixel x's degree of belonging in class i might be denoted by $\pi_i(x)$, and the vector of class memberships might be written:

$$\{\pi_1(x),\pi_2(x),\ldots,\pi_n(x)\}$$

where n is the number of classes.

In the polygon case, inaccuracy occurs in the form of variation within polygons, perhaps at the edges where boundaries are merely approximations to zones of transition (Mark and Csillag, 1989), or perhaps centrally where small inclusions and islands of different classes have not been mapped. Neither of these issues is dealt with effectively by giving the polygon a fuzzy class membership. Instead, it is necessary to abandon the polygon model because it is fundamentally unable to serve as an adequate basis for representing within-polygon variation. Instead, we see the geometry of the polygon model as an artifact of the mapping process, having little value in an effective approach to data quality, and transform to the raster model. Thus both heterogeneity of polygon class and transition near the boundary are represented through the use of pixel class memberships.

While the concept of fuzzy pixel classification is a familiar feature of the remote sensing literature, there has been very little research on the processing of such data within GIS. In part this may be because of concerns over data volume, since n memberships must be stored for each pixel, rather than one integer between 1 and n. In practice, however, it is rare for more than two class memberships to be significantly greater than zero in any one pixel. Fuzzy-classified scenes are difficult to visualize for similar reasons, and it is not clear how measurements such as class area can be made from such data. Thus despite the availability of fuzzy classifiers, and the greater information content of fuzzy-classified scenes, it is tempting to convert such data to a simple maximum likelihood classification on the grounds that the latter are much easier to handle.

The purpose of this paper is to discuss methods of visualization and processing for fuzzy-classified scenes within GIS. We include with this term not only the results of fuzzy classification in remote sensing, but also derivatives of the polygon model where each pixel is associated with a mixture of classes, or with probabilities of class membership. The next section discusses the meaning of such data from a statistical perspective, and introduces the concept of an error model. The third section discusses a rule-based fuzzy classifier for use in interactive visualization of scenes. This is followed by a description of the environment for visualization of fuzzy-classified scenes developed by the authors. The final summary discusses directions for future research.

Probabilistic Perspective

Consider a raster in which each pixel is associated with a vector of class memberships. The various possible sources and interpretations of this data were discussed in the previous section. To provide a probabilistic interpretation, we assume that the memberships are normalized by pixel:

$$p_i(x) = \pi_i(x) / \sum_k \pi_k(x)$$

Thus $p_i(x)$ is interpreted as the probability that pixel x belongs to class i out of the n classes. This might be interpreted in a mixed pixel context as the proportion of pixel x's area that is of class i; or the proportion of interpreters who would have assigned the pixel's area to class i; or the proportion of pixels with the same spectral response as x that are truly i; and numerous other interpretations are possible also.

We define the term multinomial probability field (MPF) as a vector field whose value at any point is a normalized vector of class membership probabilities of length n. A raster provides a suitable way of creating an acceptable approximation of such a field in a digital database.

Although a display of pixels showing the membership in each class is informative, it nevertheless fails to convey an impression of uncertainty, suggesting that memberships are expressions of deterministic knowledge, rather than of lack of knowledge, or of fuzziness. A similar situation in geostatistics has recently been the focus of a paper by Englund (1992). When the technique of Kriging is used to create an interpolated surface between sample points of known value, the result is both a surface and a map of uncertainty. In fact the surface is the estimated mean, and the map of uncertainty shows estimated variance around the mean. Englund deviates from common practice by showing not the map of estimated means, but sample maps from the distribution of possibilities defined by the means and variances. These, rather than the estimated mean, are then used in GIS processing. As a result, Englund is able to provide visually dramatic illustrations of the uncertainty expressed by the estimated variances, but normally ignored in analyses based on estimated means.

Englund's Kriging means and variances provide an error model, or a stochastic process whose outcomes or realizations represent the uncertainty inherent in the data. Goodchild, Sun and Yang (1992) define an error model in the context of spatial databases as "a stochastic process capable of generating distorted versions of the same reality". The best known error model is the Gaussian, used to describe uncertainty in measurements of a simple scalar quantity like the elevation at a point. Each of the outcomes of such an error model provides one possible version of the truth, as it might be interpreted by one soil scientist, or as it might be digitized by one operator.

In the context of an MPF, the probabilities are the equivalent of Kriging means, and a map of them similarly fails to convey an impression of uncertainty. Goodchild, Sun and Yang (1992) describe an error model for an MPF. Each realization is a map in which each pixel is assigned to a single class. Its two essential properties are:

1. between realizations, the proportion of times pixel x is assigned to class i approaches $p_i(x)$ as the number of realizations becomes large; and
2. within realizations, the outcomes in neighboring pixels are correlated, the degree of correlation being controlled by a spatial dependence parameter.

When the spatial dependence parameter is zero, outcomes are independent in each pixel (the case illustrated by Fisher, 1991). However, this is almost certainly unrealistic since few if any real processes are likely to create such independent outcomes. As the parameter increases, outcomes are correlated over longer and longer distances; one suitable interpretation of this is that larger and larger inclusions within polygons are ignored, or fall below the the minimum mapping unit area.

Many commonly used descriptions of map error fail to meet the requirements of an error model, since they fall short of the complete specification of a stochastic process. Such descriptions include the width of an epsilon band, the measures mandated by many map accuracy standards, the statistics of the misclassification matrix used in remote sensing, and the reliability diagram found on many topographic maps. All of these are useful error descriptors, but fall short of being useful error models. Neither is there a useful connection between many such descriptors and the necessary parameters of error models. For example, it is not possible to connect the parameters in the model described above with such measures as positional accuracy of polygon boundaries, or per-polygon misclassification of attributes.

A Rule-Based Fuzzy Classifier

Uncertainty is endemic to land classification or regionalization, because with few exceptions the Earth's surface is not naturally divided into regions of uniform attributes divided by clear boundary lines. In practice, while some boundaries between classes may follow well-defined lines such as roads, rivers or ridges, other boundaries must be drawn through zones of transition, ecotones, or similarly fuzzy areas. As a consequence, maps of land cover made by different observers may show different boundary positions, and also different numbers of regions and different boundary network topologies. Such uncertainty may be further complicated by imprecision in our language for classification (Leung, 1984, 1985, 1987). Therefore it is essential to have a built-in mechanism for analyzing and displaying uncertainty within a spatial data handling environment.

Conventionally, classification of remotely sensed scenes is performed algorithmically. In supervised classification, techniques such as maximum likelihood (see for example Nilsson, 1965; Duda and Hart, 1973) and the minimum distance method (Wacker and Landgrebe, 1972; Borden *et al.*, 1977; Phillips, 1973) are all procedural. In unsupervised classification, the most common method is cluster analysis (see for example Duda and Hart, 1973; Coleman and Andrews, 1979) which is again algorithmic in structure.

A common drawback of all of these methods is that they cannot handle uncertainty. Fuzzy cluster analysis (see for example Ruspini, 1970, 1973; Bezdek, 1981) and fuzzy graphs (see for example Leung, 1984) can analyze and depict uncertainty in classification in general and image analysis in particular. Nevertheless these methods are mechanical, and cannot communicate to users any knowledge behind the classification.

To make fuzzy classification more flexible, informative and intelligent, a rule-based classifier has been developed within an expert system environment (Leung and Leung, 1992a,b). In place of an algorithm, the classification scheme is represented by a set of rules indicating how spatial classes are conceptualized and spatial data are classified. A rule in the rule set is generally expressed as:

```
(rule <rule-name>
if <object 1> <operator 1> <value 1> and/or
  <object 2> <operator 2> <value 2> and/or
           .
           .
           .
           then <object n> is <value n>
           ) certainty is <certainty factor>.
```

The operators can be ordinary inequalities (>, <, =, >=, <=) or fuzzy inequalities ($\underline{>}$, $\underline{<}$, $\underline{=}$, $\underline{>=}$, $\underline{<=}$, where means approximately). The certainty factor can be a precise value in some fixed range, a fuzzy number, or a linguistic probability (Gopal and Woodcock, 1992).

In the identification of water from MSS data, a typical rule might read:

> If the spectral value in Band 3 (X_3) is *approximately less than 8* and the spectral value in Band 4 (X_4) is *approximately less than 5* then the pixel is a water body, certainty is 1.

Based on evaluations, ground truthing, experts' experience and knowledge gained, rules can be modified, deleted or added according to the rule set without having to rewrite any part of the program, in expert system enviroments such as those provided by Leung and Leung (1992a,b). The knowledge-based approach is thus more versatile than algorithmic approaches.

Regardless of which approach is used (algorithmic or rule-based), fuzzy spatial classification differs from the non-fuzzy scheme in that it can depict the intrinsic uncertainty of spatial data. Intermediate areas, fuzzy boundaries and fuzzy regions can be identified by gradation, while precise boundaries can be handled within the same framework by coupling high levels of certainty with spatially sharp changes in class memberships. However, to communicate uncertainty to the user, we need to devise an effective scheme for visual display.

Tools for Visualization

In this section we describe the tools we have developed for rule-based fuzzy classification, and visualization of MPFs. As we argued in the first section, classification procedures are important for remotely sensed imagery, but it is also desirable to be able to visualize MPFs from sources such as land cover maps, in which classification is performed by other means. For this reason, the system is modular in design, and includes a classification module, display module, and modules for data manipulation. It is interactive and uses a graphic user interface, all instructions and operations being triggered by selecting appropriate screen buttons. Windows are opened and closed as appropriate. The system has been developed in C and X Windows for the IBM RS/6000 under the AIX operating system.

Within the classifier module, fuzzy rules are managed by a built-in mechanism with fuzzy logic connectives. To facilitate rule editing, fuzzy concepts can be modified on-screen by changing critical points in the domain over which the associated membership functions are defined.

In the display module, images can be displayed directly by associating colors with spectral bands without classification, in order to support direct visualization of the preclassified scene. However the most important component of the module supports the display of classified images. In general, techniques of dithering and bit-mapping can be used to display uncertainty (Leung and Leung, 1990) in terms of levels of class membership, to expose the spatial variation in membership within regions or across region boundaries. In addition the system provides several other measures and methods for conveying information about an MPF to the user. The following sections briefly describe the principal tools.

1. Unclassified image

Colors can be assigned to spectral bands to create conventional false-color representations of the unclassified scene. This allows the user to see the raw data before classification.

2. Classified image

The RGB color model is used to display the results generated by the fuzzy classifier, or input from some other source. Each class is associated with a point in RGB space, and each vector of class memberships is mapped to an intermediate point in the color space by linear interpolation. This method is successful for two classes (n=2) provided the pure-class colors are chosen carefully, but it is difficult for the eye to decode the results for n=3, and for n>3 the mapping from class membership vector to color space is no longer unique. Moreover mapping is non-unique for n=3 if the class memberships have not been normalized to sum to 1 (see above).

It is possible to display each pixel's degree of belonging to each class as a numerical value, or graphically as a bar chart. The corresponding location in color space can also be displayed.

To deal with the difficulty of visualizing membership in many classes, it is possible to display each class's memberships separately using a grey scale. By using multiple windows one can display the general distribution of each class for up to four or even six classes simultaneously.

Sometimes it is desirable to have a non-fuzzy image of a fuzzy scene. A simple defuzzing mechanism is maximum likelihood, where the displayed class $f(x) = i$ if $\pi_i(x)>\pi_j(x)$ for all i,j, i not equal to j; that is, a pixel is assigned to class i (and displayed with class i's color) if its degree of membership in class i is highest. The user has control over the colors assigned to each class. Frequency distributions of the entire image can be displayed, and the user can zoom into a selected area, or display the contents of any pixel.

3. Area

Calculation of the area occupied by each class is a common GIS function. For conventionally classified scenes or other forms of raster data it is calculated by counting the pixels assigned to each class and multiplying by pixel area. However the solution is less clear in the case of fuzzy-classified scenes. If $p_i(x)$ is interpreted as the proportion of pixel x that is truly class i, as in a mixed pixel interpretation of fuzziness, then the area of class i will be the sum of such fractions added over the scene. On the other hand if $p_i(x)$ is interpreted probabilistically, the same estimate must be interpreted as the expected area of class i. Similar approaches are appropriate if $p_i(x)$ is given other probabilistic interpretations. Thus the calculation of area on a fuzzy-classified scene seems adequately addressed by calculating:

$$A_i = b \sum_x p_i(x)$$

where b is the area of each raster cell.

More difficult is the estimation of error variance, standard error, or the uncertainty associated with such estimates. In the mixed pixel interpretation A_i is deterministic, with zero uncertainty. In a probabilistic interpretation, and assuming that outcomes in each pixel are independent of outcomes in neighboring pixels (zero spatial dependence) then the uncertainty associated with area estimates can be determined from the statistics of the binomial distribution in the form of a standard error:

$$s_{ei} = b \{ \sum_i p_i(x) [1-p_i(x)] \}^{1/2}$$

where s_{ei} is the root mean square uncertainty in estimate A_i (Fisher, 1991, used Monte Carlo simulation to estimate standard error). But when spatial dependence is present, as it almost always is, and outcomes in neighboring pixels are correlated, it is necessary to resort to the methods described by Goodchild, Sun and Yang (1992).

4. Entropy

The degree of certainty in a pixel's classification can be measured in various ways, but one that expresses the degree to which membership is concentrated in a particular class, rather than spread over a number of classes, is the information statistic or entropy measure:

$$H(x) = - (1/\ln n) \sum_i p_i(x) \ln p_i(x)$$

where H(x) is the entropy associated with pixel x. H(x) varies from 0 (one class has probability 1, all others have probability 0) to 1 (all classes have probability equal to 1/n). The system allows a map of H to be displayed using a grey scale; light areas have high certainty (probability concentrated in one class) while dark areas have low certainty.

The degree of fuzziness associated with membership in each class can be assessed by another form of the entropy measure:

$$H_i = - (1/N \ln 2) \sum_x \{p_i(x) \ln p_i(x) + [1-p_i(x)] \ln[1-p_i(x)]\}$$

where the sum is now over the pixels and N is the number of pixels. H_i is zero if the probability of membership in class i is 0 or 1 in all pixels, and 1 if probability is 0.5 in all pixels. The overall entropy H of the entire fuzzy scene can be obtained by adding these measures over all classes.

5. Realizations

As noted earlier, an important aspect of visualizing uncertainty is the ability to view individual realizations of an error model, rather than its parameters. All of the previously noted methods display some aspect of the probability vectors, which are the parameters of the error model's stochastic process, rather than its outcomes. Viewing a display of probability vectors necessarily diverts attention from the variation between realizations, and focuses more on the average or expected case.

The system includes the ability to display realizations of the error model, using user-determined levels of spatial dependence. Goodchild, Sun and Yang (1992) discuss possible methods for determining appropriate levels, as attributes of the entire map, or of individual classes, or of geographic regions. A display of four or six different realizations in different windows on the screen provides graphic illustration of the implications of uncertainty in spatial data, and draws attention to its influence on analysis, modeling and decision-making.

Summary and Future Directions

It is often argued in the GIS community that while uncertainty is endemic to spatial data and undoubtedly affects the outcomes of spatial data processing, it is best not to draw attention to it because of its complexity and potentially damaging effects on decision-making. The user "does not want to know". Analogous software systems, such as the statistical packages and database management systems, do not include techniques for capturing, storing and manipulating explicit information on uncertainty, so why should GIS? We believe that this argument is both intellectually unsound and disastrously shortsighted. Most spatial decisions, particularly important ones, are made in an environment of conflict and controversy. As GIS matures and becomes available to more and more parties to a debate, the naive view that the party with the GIS somehow carries greater weight will become less and less realistic, and easier and easier to attack. Pressures for better quality assurance and control are already emerging from instances of GIS-related litigation.

Spatial statistics is a complex and difficult field, and few GIS practitioners have more than an elementary understanding of its techniques and concepts. Moreover visual techniques are inherently convincing and communicative. Thus it seems that visualization will have to be a fundamental part of any concerted effort to handle uncertainty within GIS. Goodchild, Sun and Yang (1992) have argued that visualization is the key to user participation in the determination of the key spatial dependence parameters in spatial statistical models of uncertainty.

An MPF is inherently multidimensional, and this paper has presented a number of techniques for improving the user's ability to understand this particular form of spatial variation. However any communication system must satisfy the requirements of the user as much as it exploits the capabilities of the system, and it seems clear to us that an ideal design can only come from the experience of working with these tools in a real analytic environment.

Acknowledgment

The National Center for Geographic Information and Analysis is supported by the National Science Foundation, grant SES-10917.

References

Bezdek, J.C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms* New York: Plenum.

Borden, F.Y., Applegate, D.N., Turner, B.J., Merembeck, B.F., Crenshaw, E.G., Lachowski, H.M. and Thompson, D.N., 1977. "Satellite and Aircraft Multispectral Scanner Digital Data Users Manual" *Technical Report ORSER-SSEL 1-77*, University Park, PA: Pennsylvania State University.

Coleman, G.R. and Andrews, H.C., 1979. "Image Segmentation by Clustering" *Proceedings, IEEE* 67: 773-785.

Duda, R.O. and Hart, P.E., 1973. *Pattern Classification and Scene Analysis*, New York: Wiley.

Englund, E., 1992. "Spatial Simulation: Environmental Applications" in Goodchild, M.F., Parks, B.O. and Steyaert, L. (eds.) *GIS with Environmental Modeling*, New York: Oxford University Press (forthcoming).

Fisher, P.F., 1991. "Modeling Soil Map-Unit Inclusions by Monte Carlo Simulation" *International Journal of Geographical Information Systems* 5(2): 193-208.

Goodchild, M.F., Sun, G. and Yang, S., 1992. "Development and Test of an Error Model for Categorical Data" *International Journal of Geographical Information Systems* (forthcoming).

Gopal, S. and Woodcock, C., 1992. *Accuracy Assessment of Thematic Maps Using Fuzzy Sets 1: Theory and Methods* (unpublished paper).

Laurini, R. and Thompson, D., 1992. *Fundamentals of Spatial Information Systems,* San Diego, CA: Academic Press, 680pp.

Leung, Y., 1984. "Towards a Flexible Framework for Regionalization" *Environment and Planning A* 16: 203-215.

Leung, Y., 1985. "A Linguistically-Based Regional Classification System" in Nijkamp, P., Leitner, H. and Wrigley, N. (eds.) *Measuring the Unmeasurable,* Dordrecht: Martinus Nijhoff, pp. 451-486.

Leung, Y., 1987. "On the Imprecision of Boundaries" *Geographical Analysis* 19: 125-151.

Leung, Y. and Leung, K.S. (1990) *Analysis and Display of Imprecision in Raster-Based Information Systems* (unpublished paper).

Leung, Y. and Leung, K.S. (1992a) An intelligent expert system shell for knowledge-based geographic information systems: 1, the tools. *International Journal of Geographical Information Systems* (forthcoming).

Leung, Y. and Leung, K.S. (1992b) An intelligent expert system shell for knowledge-based geographic information systems: 2, some applications. *International Journal of Geographical Information Systems* (forthcoming).

Mark, D.M. and Csillag, F., 1989. "The Nature of Boundaries in `Area-Class' Maps" *Cartographica* 26(1): 65-78.

Nilsson, N.J., 1965. *Learning Machines: Foundations of Trainable Pattern-Classifying Systems,* New York: McGraw Hill.

Phillips, T.L. (ed.), 1973. *Larsys Version 3 Users Manual* West Lafayette, IN: Laboratory for Applications of Remote Sensing, Purdue University.

Ruspini, E.H., 1970. "Numerical Methods for Fuzzy Clustering" *Information Sciences* 319-350.

Ruspini, E.H., 1973. "New Experimental Results in Fuzzy Clustering" *Information Sciences* 273-284.

Tomlin, C.D, 1991. *Geographic Information Systems and Cartographic Modeling,* Englewood Cliffs, NJ: Prentice-Hall.

Wacker, A.G. and Landgrebe, D.A., 1972. "Minimum Distance Classification in Remote Sensing" *First Canadian Symposium on Remote Sensing, Ottawa.*