Outside-In, Inside-Out: Two Methods of Generating Spatial Certainty Maps

Kim E. Lowell
Centre de recherche en géomatique
Pav. Casault, Université Laval
Ste-Foy, Québec G1K 7P4
CANADA
(418) 656-7998
FAX: (418) 656-7411
Internet: Kim. Lowell@scg.ulaval.ca

Presented at the second annual conference of GeoComputation '97 & SIRC '97, University of Otago, New Zealand, 26-29 August 1997

Abstract

Problems with standard Boolean maps produced through subjective interpretation of a phenomenon (i.e., forest or soils maps) are discussed and two alternatives based on spatial certainty are presented and discussed. One requires multiple interpretations of the phenomenon in order to construct a library of spatial uncertainties; such an Uncertainty Library can be used subsequently to estimate error across cartographic boundaries. The other requires interpreters/cartographers to identify only those map elements which are "100% certain." A spatial interpolation algorithm is subsequently applied to this information to "fill in the gaps" with certainty information for each map type. Both methods have advantages and disadvantages relative to standard Boolean maps and also to each other. These are discussed in general terms and also through the presentation of specific examples. It is concluded that though uncertainty-based cartographic representations provide more flexibility than do conventional Boolean maps, the construction of the former is not without its problems either.

1. Introduction

In recent years, a number of researchers and practitioners have become interested in maps showing spatial certainty/ uncertainty. These maps are often conceptualized as showing fuzzy membership values or probabilities for a given map class (e.g., Hall *et al.,* 1992, Lowell, 1994). One way to produce such maps is to obtain a standard thematic map of the variable of interest and use available information and/or make assumptions about the magnitude and nature of the error inherent in the variable mapped. The information on the Boolean map may then be "perturbed" stochastically to produce a fuzzy map of the variable under study (e.g., Fisher, 1992, Goodchild *et al.,* 1992). In this process, it is often the case that one considers only the cartographic type/value that was originally mapped at a given location rather than also considering additional spatial information that cartographers or others familiar with the variable mapped possess mentally. For example, one might perturb the map type "Forest" using various assumptions about the certainty of this type, but one might not consider whether or not a given "Forest" polygon is surrounded by "Lake" on one side or "Forest Scrub" on another. Yet these two types as neighbors of a Forest polygon imply very different things about the certainty within a polygon labeled "Forest." Hence it would seem to be useful to consider the characteristics of a map type at various places within a given polygon — e.g., close to/far from a boundary — and/or to consider this relative to an adjacent polygon of a given type — e.g., a Forest/Lake boundary should be treated differently than a Forest/Forest Scrub boundary. This also suggests two valid, yet seemingly contradictory approaches to the development of certainty or fuzzy maps.

In one, a single map and a "library of uncertainties" showing the certainty associated with a boundary of a given type are assumed to be available (Fig. 1). Edwards and Lowell (1996) have demonstrated how such an Uncertainty Library may be developed from multiple cartographic interpretations of a given phenomenon. However, such a process has some important limits for the work described; these will be discussed subsequently. For the moment, assume that an Uncertainty Library is available that is known to be applicable to a recently constructed Boolean — i.e., conventional thematic — map for a given phenomenon. The Uncertainty Library consists of the standard deviation for the true location of a boundary line separating any two types that may appear on the map. Given this information, it would seem to be a relatively simple matter to construct a certainty map from the available Boolean map: one identifies the cartographic types on either side of the boundary and applies the appropriate standard deviation. However, it will be demonstrated that the process is more complex than it appears here.

In the other approach, what is considered "a map" is radically different from a conventional Boolean map. Generally, one considers a map to be a complete coverage of a surface — one in which every location has a "value" relative to the variable being mapped. (Note that even a collection of points may be viewed this way ultimately since interpolation is usually used to assign a value to all gaps between points before the map is employed in any analysis.) However, in some disciplines such as natural resources, a standard Boolean thematic map represents a considerable loss of information concerning spatial uncertainty. For example, in the production of forest type maps for Quebec (Ministry of Natural Resources, 1995), aerial photographs are interpreted subjectively by trained human photo-interpreters. The author's experience has shown that in interpreting a photograph, an interpreter works from a "definite" object or area — e.g., a lake, a clearcut — and proceeds to less certain features. At various times the photo-interpreter places a boundary line because of 1)the actual recognition of a definite boundary or dividing line (with types on either side not necessarily being known), 2)the necessity to separate two regions of clearly different types for which the boundary location is not known exactly, or 3)due to the recognition of an actual closed polygon of a given cartographic type (Fig. 2). Note that in the traditional method a photo-interpreter does not always see closed polygons, but is forced to produce them nonetheless. The result is a Boolean map for which one must try to infer certainty from a subjective knowledge of the phenomenon being mapped. If photo-interpreters were permitted to produce an interpretation showing only those features having "100% certainty," it would be possible to

Figure 1. Example of map and associated Uncertainty Library. Library shows one standard deviation for location. (Representation inspired by R. Aspinall, pers. comm.)

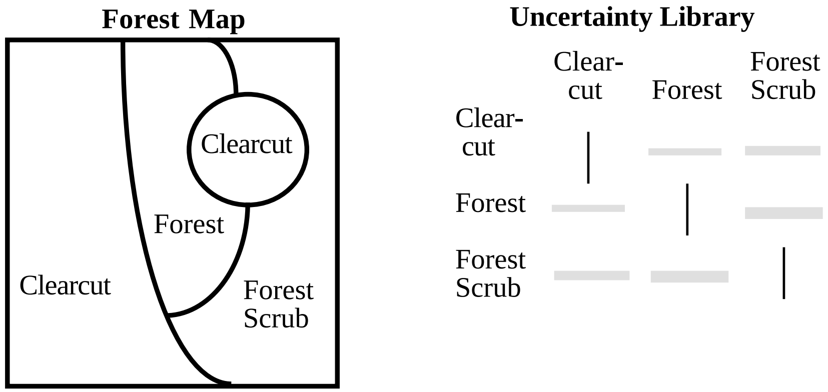
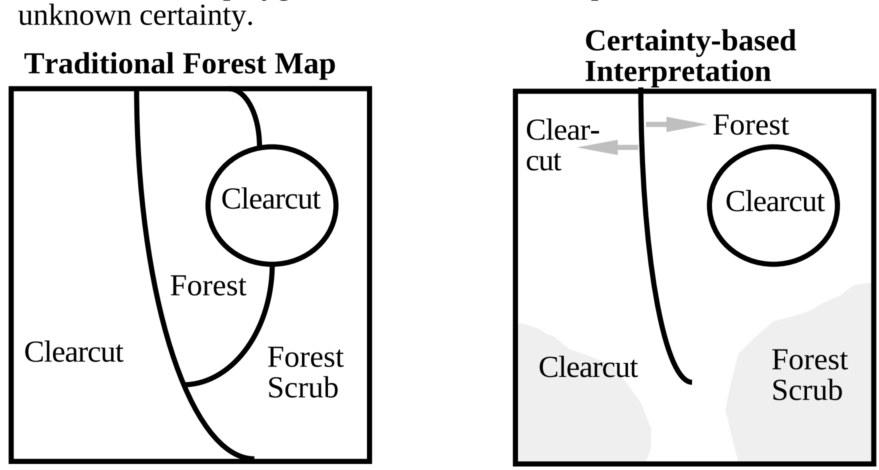


Figure 2. Traditional polygon map and alternative certainty-based interpretation. Regions and boundaries on certainty-based interpretation are "100% certain" whereas polygons on the traditional map have variable and unknown certainty.



derive a certainty map using spatial interpolation. In doing so, instead of inferring uncertainty, one would have explicit information and there would presumably be more consistence among interpretations. However, this process is also not as straightforward as it would seem.

At this juncture, the primary point being made is that, for the production of certainty maps, we have two possible alternatives to map perturbations and its accompanying assumptions. However, these alternatives also are subject to certain difficulties and assumptions. Moreover, the two would seem to be somewhat contradictory. One employs the boundary as its basic unit and works towards the center of polygons. That is, the high certainty implicitly associated with polygon cores is derived from observations at polygon boundaries. In the second, it is conceivable that only polygon cores will be identified by a photo-interpreter as being "100% certain." Thus the low certainty at boundaries is derived from the high certainty at polygon cores. Are the two methods compatible? What are the problems associated with each? Are there any particular benefits of one over the other?

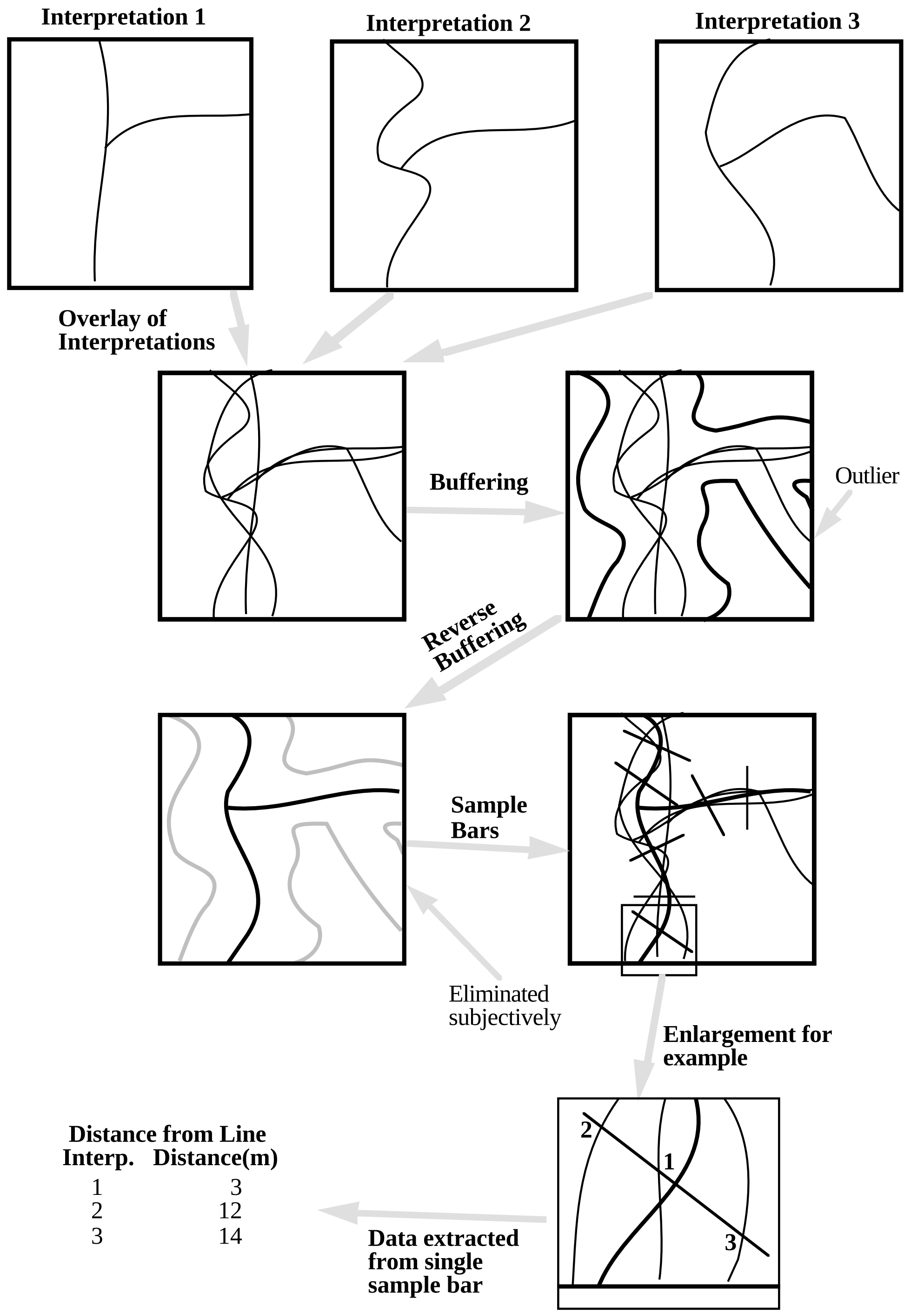
The purpose of this paper is to respond to these and similar questions and also to explore the two methods in greater detail. This includes not just computational aspects regarding the two, but also user considerations including data collection and organization.

2. Method 1: Boundary-based Certainty Maps: Outside-In

2.1 Data collection

If one is to use a single, conventional thematic map to generate a certainty-based map, knowledge about the associated error must be available *a priori*. In this paper, it is assumed that the form of this knowledge is a standard deviation on the location of a boundary of a given type (Fig. 1). In effect, this means that one must fill a k-by-k matrix (in which k is the number of map classes) with a standard deviation or other measure of boundary uncertainty. Put another way, we need to know the locational uncertainty of Forest/Clearcut boundaries, Forest/Forest Scrub boundaries, and Clearcut/Forest Scrub boundaries (assuming k=3 in this case). One of the most straightforward ways to obtain this information is through multiple interpretations of the same phenomenon. The method to be described was developed and described by Aubert (1995) and is presented schematically in Figure 3.

Figure 3. Method for obtaining Uncertainty Library from multiple interpretations.



Distance from Line

Interp.	Distance(m)
1	3
2	12
3	14

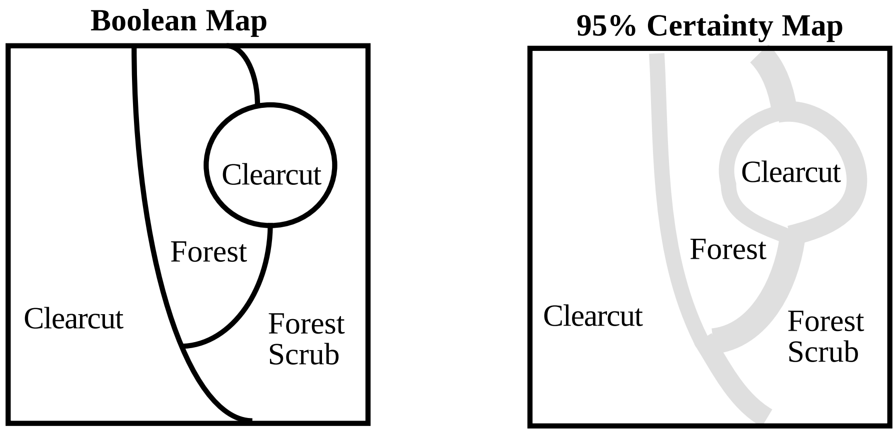
To develop a library of spatial uncertainties, one starts with multiple interpretations of the same phenomenon. As an overall goal, it is desired to use these interpretations to develop a map of "truth" by overlaying the boundaries, identifying the mean location for each boundary, and subsequently quantifying the error around each of these mean boundaries. Suppose for the purposes of explanation, three such interpretations are available (Fig. 3). The boundaries of these are overlaid and a buffering operation around all boundaries is performed. In doing so, the size of the buffer selected will strongly affect the results. This is because all lines within the buffering distance of each other will be bundled together and considered to represent the same "true" line. "Too large" a buffer will cause more than three lines to be bundled into one boundary — something that is clearly impossible if one has only three interpretations; a buffer that is "too small" will cause relatively few boundaries to be bundled together. Once the buffer size has been selected and the buffering operation performed, the outer limits of each buffer are retained and outliers removed manually. Note that this requires a subjective judgment on which lines are outliers (middle of Fig. 3) and sometimes causes less than *n* interpretations to define a line. A reverse buffering operation is then performed to identify the mean location — assumed to be the "true" line location — for each boundary. This mean/true boundary location is then overlaid on the original interpretations and a series of sample bars placed along the mean line. The bars are the size of the original buffer and are placed at selected intervals along the mean location; these sample bars are placed perpendicular to the mean/true line and are spaced far enough apart to avoid the effects of spatial autocorrelation. The distance each line on an original interpretation is from the mean/true line for each sample bar is determined and the mean and standard deviation of these calculated. These are then summarized by boundary type — i.e., all the Forest/Lake boundaries summarized together regardless of their location on the map.

2.2 Treatment and use of information

The method described provides an estimate of the error associated with each type of map boundary line and provides for the development of an Uncertainty Library. Moreover, it provides a means to test if there is a systematic bias in interpretation. For example, it might be the case that in looking for a Forest/Lake boundary on an aerial photograph, there is a consistent tendency to place the separating line towards/away from the Forest. This might be caused by shadows and/or the interpreter's eye consistently being drawn toward/away from water, and/or other factors. This also highlights another use of the methodology developed — one can test the nature of the distribution of the error across the mean location. This was done in the original study (Aubert, 1995); there was no evidence to reject the null hypothesis of the error across the mean line location being distributed according to a Gaussian distribution.

Knowing that the information sought — i.e., a fully populated Uncertainty Library — can be obtained, attention turns to its use. It is assumed here that the Uncertainty Library has been compiled in such a manner as to be robust enough to be used for an area for which an uncertainty map is to be developed. A single Boolean map is produced from a single photo-interpretation of the area, and one may now ask questions such as "Show all the areas which have a probability *p* of being Type A." For example, asking for the 50% confidence interval for all map types will produce the Boolean map itself since, effectively, each boundary represents the point at which there is a 50-50 chance of being the type on either side of the boundary. Similarly, one may want the 95% confidence interval on Clearcuts. A Gaussian distribution can be generated around all Clearcut boundaries using the Uncertainty Library and the point at which 95% of the region is outside this identified (Fig. 4). Note, however, that doing so will produce discontinuities at places where more than two boundaries meet. Furthermore, there is no guarantee that the confidence intervals for all types for a given location will sum to 1.0. That is, if I have a place that is located exactly at the 95% confidence level for Type A, this means that there is a 5% probability that this location is actually some other map type. Yet if I ask for the 95% confidence interval for Type B, it is possible that the same location will be located

Figure 4. Boolean map and 95% certainty map. Gray zones are less than 95% certain of being any given type.



within the area for Type B (Fig. 5a). Moreover, in the case of a Softwood/Hardwood boundary for a given Softwood polygon, the uncertainty would (presumably) be small meaning that the 95% confidence interval would be relatively close to the boundary on the thematic map. However, a Softwood/Mixed boundary for the same polygon would have much less certainty (presumably) meaning that there is a discontinuity even for the same polygon boundary (Fig. 5b). It is even possible that, if uncertainties are large enough for certain types, the very existence of a polygon in a given place is questionable (Fig. 5c).

There are other problems with this approach which are inherent in the way that the Uncertainty Library is developed. Of critical importance in this construction process is the size of the buffer zone selected for use. Not only does this affect which lines from the original interpretations will be bundled as representing the same "true" line, but it also affects the maximum uncertainty that will be found in the Uncertainty Library. For example, if the buffer zone selected is 20 m with three interpretations, then the maximum uncertainty will be 40 m — i.e., three lines spaced equidistant at 20 m which will be bundled together. Moreover, because the 20 m may not be applicable over the entire length of a line, one must decide subjectively when something is to be considered an outlier (Fig. 3). By definition, certain lines will be assessed as outliers even though this may be because they cause problems for the methodology and not because they are truly statistical outliers. The net result of all of these factors is that the values in the Uncertainty Library are likely to be underestimates of the actual spatial uncertainties associated with a given boundary line. Finally, this method must make the assumption of a Gaussian distribution across a line. While there is evidence to support this assumption for the synthetic images on which the method was originally developed, there has not been exhaustive testing of this under a wide variety of cartographic conditions.

3. Method 2: "Polygon core"-based Certainty Maps: Inside-Out

3.1 Data collection

Traditionally, photo-interpretation is conducted by identifying boundaries of homogeneous areas with the constraint that the boundaries form closed polygons over the entire map. Each polygon is then labeled with its appropriate cartographic appellation. In the proposed certainty-based photo-interpretation, it is only required that interpreters identify those features which are "100% sure." It is not necessary that these form closed polygons Theoretically, these features can be one of three elements (Fig. 6). First, one may have an actual, definite polygon. In forestry, such

Figure 5a. A situation in which the 95% certainty of one boundary is completely within the 95% certainty of another.

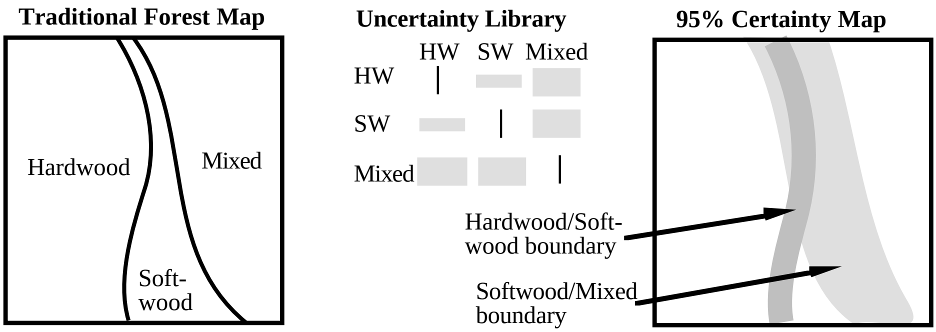


Figure 5b. A situation in which the boundary of a single feature has discontinuities.

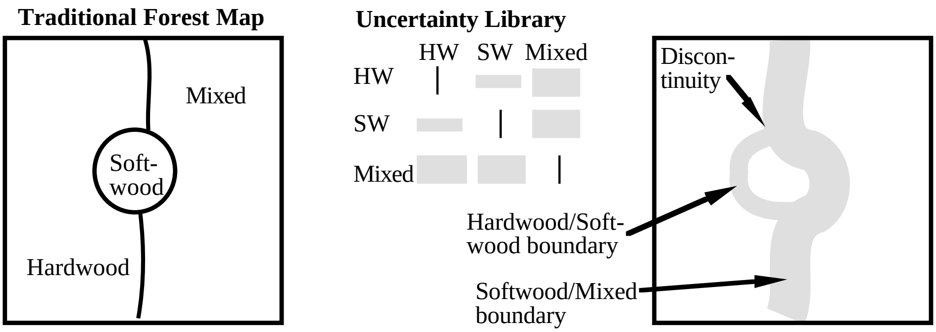


Figure 5c. A situation in which the magnitude of the uncertainty calls into question the existence of a polygon.

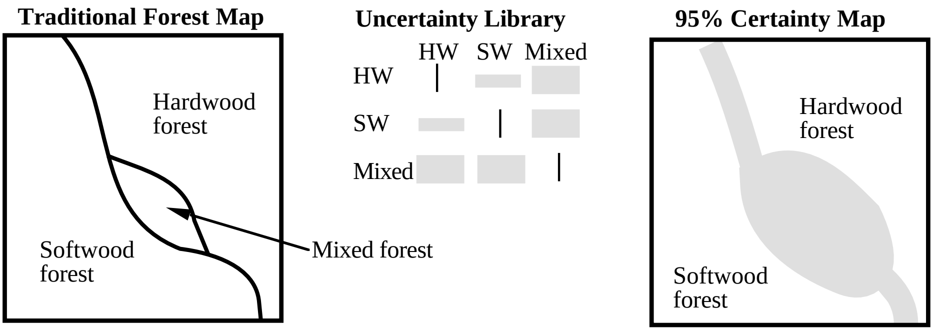
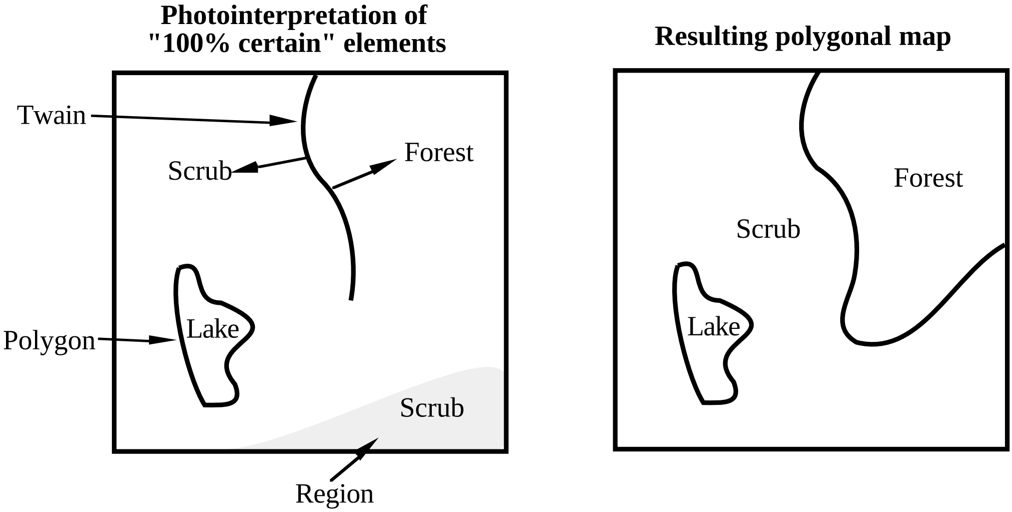


Figure 6. Certainty-based photo-interpretation and polygonal map resulting from spatial interpolation and maximum likelihood classification.



elements are most likely to be lakes, or clearcuts, or power line right-of-ways, etc. — — i.e., distinct elements with definite boundaries that truly are polygons. Note that with such objects, only the interior of the polygon is known to be a given type; what is on the outside of the line delimiting the polygon is not necessarily identified. The second type of feature possible is a line for which the cartographic type on both sides of the line is labeled, but the interpreter is not obliged to form a closed polygon with the line. This type of feature is referred to as a "twain" herein. The third and final element possible is a region or point of a known cartographic type whose boundaries are not exact; instead it is the core of the polygon that is recognizable and identifiable with "100% certainty." Note that the resulting "map" has virtually no use for human interpretation because of humans being accustomed to closed polygon maps. It is critical that this information be treated or post-processed subsequently in order to render it useful for human interpretation.

3.2 Treatment and use of information

Treating this information requires a thorough understanding of the nature of the data. Effectively, a certainty-based map has a series of points labeled "100" (% certain of being a given type). In the case of a closed polygon on such a map, the interior and the bounding line are known to be, "without doubt," the cartographic type labeled. A twain is two sets of points side-by-side which is known to have one type "on the left" and a different type "on the right." Thus one has a set of "100% certain" points for one map type abutting a set of "100% certain" points for another map type arranged linearly. In effect, therefore, a twain acts as an impermeable membrane that prevents one type from "bleeding" into the other. Finally, in the case of a region — i.e., a polygon core — the boundaries of the region remain to be defined, but that core is a set of "100% certain" points for the map type labeled. Note that this way of looking at the data as a series of points labeled 100 also implies an equally valid inversion of the data. That is, if a set of points are "100% certain" to be Forest, then they are also "0% certain" to be Lake and/or Clearcut, and/or any other map type. Thus if one wants to produce a certainty map for Forest, one need only label all Forest points "100" and all others "0" and conduct a spatial interpolation.

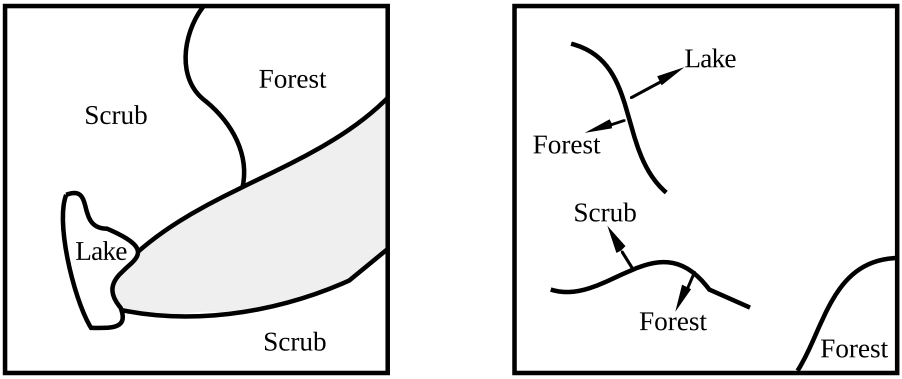
If this is done for all map types individually, one obtains a certainty surface for each type which may be treated very similarly as the certainty map produced from multiple comparisons. For example, to produce a polygon map that identifies map type boundaries, one may do a maximum likelihood classification: assign each point to the class for which its certainty is the largest. Note that the boundaries so identified will effectively be the "50% line" between two classes, or the "33% line" among three classes, etc. One may also ask more specific questions of the certainty surfaces than just polygon boundaries. For example, one may ask — as before — for the map showing 95% certainty for all types (Fig. 7). This request highlights a potential problem, however.

In performing the interpolation of uncertainty in the manner described, it must be assumed that the form of the distribution of certainty from one 100% certain element to another is linear — something that may not be true. The reason for this imposed linearity involves the interpolation method that is required. Normally when one interpolates spatially, one has a single variable or set of values — e.g., points of known elevation. However, in the present case one has a set of values for Lake, a set of values for Forest, etc. When the interpolation is conducted, not only is a certainty value needed for each type at every location, but the certainty values for a given location must sum to 100. Thus we interpolate in a seemingly typical fashion, but with an added constraint. The only method for doing this known to the author is area-stealing interpolation (Gold 1989) — a variant of natural neighbor interpolation. With this method, one effectively determines the certainty value for a given type by assessing geometrically the influence of all neighboring "100% certainty" points and their associated types. (For more detail see Lowell (1994).) Because one is not literally interpolating across a boundary, one cannot change the form of the distribution across the boundary.

Another more subtle problem is in the nature of the data. In the example presented (Fig. 6), the data "made sense" and could be understood without the use of a computer. However, it is easy to imagine a situation in which the data do not "make sense" (Fig. 7). Nonetheless, because a non-intelligent algorithm will be applied to these data, a result/surface will be produced even though it may be nonsensical. Note that the problem is not with the algorithm employed; no interpolation algorithm is capable of understanding that certain data will not produce "meaningful" polygons. The problem is simply that the data make no sense relative to the way in which human beings interpret the world — something that is related to the desire to have homogeneous polygons — whereas the interpolation algorithm is certainly capable of using the data. In fact, ensuring a surface that "makes sense" is one of the reasons

Figure 7. Map showing areas which are "95% certain" (gray area is less than this for all types) and a certainty-based interpretation that would produce a nonsense polygonal map.

95% Certainty Map



that conventional methods have been employed — there is always an internal check on the consistency of polygons. This is not the case with certainty-based interpretations which may cause problems for unwary users.

4. Synthesis and Conclusions

Two methods have been presented for producing certainty surfaces from data derived from subjective human interpretation of a particular phenomenon. These are applicable to phenomena for which the production process is subject to considerable subjectivity.

One method relies on the boundaries identified by an individual and a library of uncertainties. It also involves some subjectivity in that in producing the Uncertainty Library, the size of a buffer employed during one operation, as well as the determination of which interpretations or portions of interpretations are outliers, are subjective decisions that will vary from one individual to another. It also has the drawback that it makes an assumption about the form of the distribution of the error across a boundary once the Uncertainty Library is available. Furthermore, there are discontinuities in the nature of error at places at which three or more polygons join. And finally, it requires multiple interpretations of a phenomenon and/or the availability of a pre-existing Uncertainty Library.

The other method relies on data showing only those features which are 100% certain on a single interpretation and an interpolation algorithm. This method, though less subjective, also has inherent limitations and drawbacks. Simply the manner in which photographs must be interpreted is a drawback. Photo-interpreters are currently trained to identify closed polygons over an entire surface. Suddenly asking them to change their method of interpreting from "Identify closed polygons" to "Identify only those map elements that are 100% certain" is sure to cause a certain amount of discomfort and misunderstanding initially. This method also suffers from the impossibility of defining a particular frequency distribution for error as one moves from one "100% certain" element to another. Although relatively little work has been done on determining the true form of error distributions across map boundaries, it is certainly conceivable that this is not linear as must be supposed herein. Finally, this method can and will produce polygon surfaces from data even if the basic data are essentially nonsensical.

Despite these drawbacks for either method, the concept of developing certainty-based maps for interpreted phenomenon is sound. Regardless of the method of construction, such maps clearly provide more flexibility to a user than existing Boolean maps. However, it remains that, just as with Boolean maps, the certainty-based maps will only be as good as the assumptions and data used to construct them.

Bibliography

Aubert, E. (1995) Quantification of the spatial uncertainty inherent in photo-interpretations using a GIS for temporal monitoring of forest stands. Unpublished M.Sc. thesis, Université Laval (in French). 92 pp.

Edwards, G., Lowell, K. (1996) Modeling uncertainty in photo-interpreted boundaries. *Photogrammetric Engineering and Remote Sensing* 62:337-391.

Fisher, P. (1992) First experiments in viewshed uncertainty: simulating fuzzy viewsheds. *Photogrammetric Engineering and Remote Sensing* 58:345-352.

Gold, C. (1989) Surface interpolation, spatial adjacency, and GIS. In *Three-dimensional Applications in Geographical Information System*. Edited by J.R. Raper. Taylor and Francis, London. Pp. 21-35.

Goodchild, M., Guoqing, S., Shiren, Y. (1992) Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6:87-104.

Hall, G.B., Wang, F., Subaryono, G. (1992) Comparison of Boolean and fuzzy classification methods in land suitability analysis by using geographical information systems. *Environmental Planning A* 24:497-516.

Lowell, K. (1994) A fuzzy surface cartographic representation for forestry based on Voronoi diagram area stealing. *Canadian Journal of Forest Research* 24:1970-1980.

Ministry of Natural Resources (1995) *Norme de stratification Écoforestière: troisième programme de connaissance de la ressource forestiere*. Ministère de Ressources naturelles du Québec, Service des inventaires forestiers. 116 pp.