Fuzzy Spatial Queries in Digital Spatial Data Libraries

Michael F. Goodchild, Daniel R. Montello, Peter Fohl, Jon Gottsegen
National Center for Geographic Information and Analysis, and Department of Geography
University of California, Santa Barbara, CA, USA

Abstract

Three forms of uncertainty have been discussed in the context of geographic information systems: in positions of features, in attributes, and due to generalization. A fourth type occurs with uncertain footprints in digital spatial data libraries, when location is used as the primary key to drive search and retrieval. The concept of a geolibrary is defined. Fuzzy footprints are encountered frequently when the area covered by the query or the information object is defined by a vernacular place-name. Three models are proposed for representing fuzzy footprints. Cartographic display imposes limitations on the visualization of fuzzy footprints. A measure of goodness of fit is proposed and generalized to fuzzy footprints.

Introduction

Concepts of uncertainty are now firmly established as part of the apparatus of geographic information systems (GIS). Maps are of limited positional accuracy, due to the mechanics of map production and the limitations of Earth measurement systems, and it is unusual for paper maps to have positional accuracies better than about 1 part in 10^4 or 10^5 [12]. Yet much higher precisions are standard in computational systems. Many GIS designers have chosen to represent positions using double precision, or 1 part in 10^{14}. At this precision, a map of the entire Earth would be capable of recording accurately the positions of large protein molecules.

In order to formalize uncertain position, consider a point with true location U. The uncertainty over the point's location is often modeled by a bivariate Gaussian distribution, using a probability density function $q(X)$, the probability density of the point being observed at X instead of U. Because q is probability density it can be integrated over an area to obtain the probability that the observed point lies in that area, rather than elsewhere. Difficulties arise in the cases of uncertainly-located lines and areas, however, where the object is defined by a set of points, rather than a single point. In the case of an area, q is no longer meaningful as a probability density; rather, some new variable or *probability field* $p(X)$ might represent the probability that X lies in the true area. But now p is a probability, rather than a probability density, and it is meaningless to integrate it over an area. In the case of a line there is no available interpretation of either a probability density or a probability field.

More recently, concepts of uncertainty have been applied to the classification of land according to soil type, vegetation cover, and land use (e.g., [3]). Uncertainty has been conceptualized in both probabilistic [13] and fuzzy [7] frameworks [4], and Fisher [8], [9] has attempted to establish the fundamental differences between them in this context. The type present at some point X is represented by a vector $\{p_1,p_2,\ldots,p_n\}$, where p is either the probability of membership in a class, or the fuzzy membership, respectively; and the sum of p over all classes is 1.

Finally, concepts of uncertainty have also been applied to the formalization of scale effects. Maps are always generalized, to a degree that depends largely on their *representative fraction* or *scale*, or the ratio of distance on the map to distance on the ground. Thus a map at scale 1:24,000 is less generalized than one at 1:100,000. Frequently, data are not available at the scale required by an application, but only at a coarser scale. For example, accurate analysis of the suitability of a given area for development may require data at the level of generalization corresponding to 1:24,000. But only data at 1:100,000 are available in digital form from the supplier. In such circumstances it is desirable to know the uncertainty introduced into the results of analysis when too-coarse data are used. Ehlschlaeger et al. [6] and others have developed methods for simulating the information that is not available due to excessive generalization, based on geostatistical models calibrated in areas where both data are available at both scales.

In summary, several different conceptualizations have appeared in the literature on uncertainty in geographical data: the probability density of the observed location of a point feature; the probability of a point lying in an area; and the probability of a point belonging to a given class.

Although models of the uncertainty in line locations have been described [14], [15], there is no available interpretation of either a probability field or a probability density.

In this paper we focus on a new application of fuzzy and probability concepts to geographic data—the concept of a digital spatial data library. The Alexandria Digital Library (ADL) is one of six projects funded by the joint Digital Library Initiative of the U.S. National Science Foundation, National Aeronautical and Space Administration, and Advanced Research Projects Agency. Its purpose is to research and build a distributed digital library for geographically-referenced information—in effect, to provide the services of a map and imagery library over the Internet [22] (see *http://www.alexandria.ucsb.edu*). Users of map and imagery libraries frequently pose queries based on ill-defined regions, and a complex interaction between user and librarian is often needed before the query can be satisfied. Moreover, information in the library may relate to ill-defined regions on the ground; for example, the library may contain a guidebook to 'downtown Santa Barbara'. Thus both footprints—that of the query, and that of the information-bearing object (IBO)—may be ill-defined. The central problem of this paper is to provide a service in the digital environment that supports such ill-defined geographic references. The next section describes the background to the paper in more detail.

Geographic footprints and the geolibrary

Consider a region A (e.g., Illinois, Midwest, Riviera neighborhood). The region is *well-defined* if for every geographic location X, either X is in A or X is not in A. In other words, region A has a precise *footprint* on the ground. In most cases region A will be *connected*, that is, between any pair of points in A there will exist a path that lies entirely within A. Region A is *ill-defined* if one or more additional responses to the query 'is X in A' exist, such as 'X may be in A', 'X is sometimes in A', 'X is probably in A', 'some people think X is in A', etc.

In practice, geographic regions are more likely to be defined by their limits than by enumerating their contents. Because the geographic surface is continuous, a finite region will contain an infinite number of locations; thus definition by enumeration is possible only if a region is defined as an aggregate of a finite number of smaller regions. Moreover, the simple connectivity of geographic regions allows their limits to be defined simply. Thus specification of a region is often reduced to specification of its boundary. The basis of many point-in-polygon routines in geographic information systems (e.g., [25], p.216) is also therefore an adequate definition of a region:

A location X is in region A if a line drawn in any direction from X to infinity crosses the boundary of X an odd number of times.

Users of map libraries, and by extension their digital equivalents such as ADL, use footprints in several ways in support of requests for information. A user might be looking for information on Illinois, or the Midwest, or any of a vast number of named places. The library will also have cataloged its holdings by footprint. In this case, however, there is the option of identifying the footprint by name ('Illinois'), or identifying its defining coordinates, in some universal coordinate system such as latitude/longitude. While same choice exists in principle for the user—to ask for information either by place-name or by latitude/longitude—the latter is likely only in areas like the Sahara or the Pacific, where the absence of named features makes coordinates the only workable option.

Footprints are a powerful basis for cataloging the holdings of map libraries. Although several other keys to information might be used, such as date of creation, name of author, or title, in our experience the vast majority of users of map libraries rank location as the primary key. In this way map library search is very different from search through the catalogs of conventional libraries, where books are cataloged by author, title, and subject but not by the footprint of the book's information, if one exists (although place-names may appear in titles, and subject indices often include place-names as a type of subject).

Imagine a library in which the primary search mechanism consisted of a globe; and the user requested information about an area by pointing to it on the globe ("what have you got about here?"). We term such a library a *geolibrary* [11], since it provides access to information by geographic footprint, rather than by subject, author, or title. Such a library would retrieve information by location, and would be of great value in finding information on specific places, regardless of their size.

However, footprints pose a major difficulty for map libraries, and are one basis for the observation that the contents of map libraries are difficult to catalog and search effectively (and for the high level of human assistance required by users of map libraries). Footprints are multidimensional keys (normally two, but conceivably more). Moreover the space of these keys is continuous, unlike the space of subject, author, or title keys. Thus it is possible for footprints to divide geographic space in an infinite number of ways. Finally, while the set of place-names is finite and discrete, place-names have complex hierarchical relationships of containment. Thus needed information on Santa Barbara may be found on a map of California, or on a map dominated by neighboring Goleta, and cataloged under that name. Traditional subject catalogs do not support these complex relationships.

But in the digital world it is entirely possible to support retrieval by such complex, multidimensional keys. A user of ADL first identifies an area of interest by zooming from a map of the world displayed on the screen. Additional keys can be identified, such as date of the data, or theme, and then the database is searched for data sets fitting the conditions of the search. The contents of each data set can be browsed, or returned to the user, or the user can determine the conditions of access to the data in restricted cases.

Fuzzy regions and queries

A *gazetteer* is a list of named places, together with geographic coordinates identifying locations. In most cases of large places these are representative, typically central, points. The gazetteer of the *Times Atlas of the World* [23] contains over 210,000 items. But gazetteers list only well-defined places, mostly places recognized as having some kind of legal status. For example, 'Central Valley', a common informal term for the combined region of the Sacramento and San Joaquin Rivers in California, points in the *Times Atlas* to a small town in Northern California, 1990 population 4340; there is no entry for the U.S. 'Midwest'. The Geographic Names Information System (GNIS) of the U.S. Geological Survey (*http://mapping.usgs.gov/www/gnis/*), a list of place-names derived from the USGS's topographic maps and arguably a digital equivalent of an atlas gazetteer, similarly reflects a preference for places with some form of official recognition, and omits less formal terms such as 'Riviera', 'Central Coast', or 'Midwest', all terms in common use and likely subjects of library searches.

Human discourse assigns complex priorities to the status of geographic terms. Reference to 'the population of Los Angeles' can be interpreted in principle using any recognized footprint associated with the name, including the boundary of the City of Los Angeles (1996 population 3,638,100), or the boundary of the County of Los Angeles (1996 population 9,488,200). The interpretation may also depend on the location of the interpreter—'the population of Los Angeles' to a resident of London may be closer to the population of the entire Southern California conurbation, in excess of 15 million, than to that of either the County or the City. The statement 'I live in Santa Barbara' may be interpreted as indicating that the person lives inside the footprint of the City of Santa Barbara (possibly the narrowest definition), has a mailing address in Santa Barbara (broader, practical, and a factor in determining property values), or lives in Goleta (an unincorporated populated area adjacent to the City of Santa Barbara), depending again on the context of the query. In general, the scale at which the query is interpreted is likely to depend on the physical separation of the user and the feature; a query from the most distant user will be interpreted at the coarsest scale, while a local user's query will be interpreted at the most detailed scale.

In summary, the ability of a digital library to respond effectively to a query about a named place, or to catalog a data set about a named place, will depend several factors. If the place has been formalized, by being incorporated into some level of an administrative hierarchy, then it will likely appear in published gazetteers and their digital descendants. Feature names that have received recognition by being used on topographic maps produced by national mapping agencies are also readily found. But regions that remain ill-defined will likely not appear in such lists, even though they may be used frequently in everyday communication, and may convey strong images and associations to their users. Small regions such as city neighborhoods are also unlikely to be listed, however strong their significance to local residents. We believe that effective digital libraries will need to decouple these two issues of official recognition and ability to search, by making it possible for users to construct queries for both ill-defined and well-defined regions, and for librarians to build catalog entries for data sets about ill-defined regions as well.

Modeling Fuzzy Regions

Numerous techniques for digital representation of well-defined regions have been developed over the past three decades, in GIS and other technologies. In a raster, a region is denoted by enumerating all of the discrete elements contained in the region, and information is lost because of the finite size of the elements. In a vector representation, the region's boundary is represented as a series of mathematical functions (usually straight lines) connecting points; information is lost if the precision of the coordinates is less than their accuracy, or if the mathematical function is not a perfect representation of the boundary between adjacent points. Further complications arise when the boundary is defined as a straight line on the curved surface of the Earth.

We conceive of an ill-defined object as a field $z(X)$, giving a measure of the object's presence at any point in the plane (or on the surface of the Earth). A well-defined object is a binary field, $z = \{0,1\}$. But for ill-defined objects various probabilistic and fuzzy interpretations of z are possible.

Blakemore [1] and others have suggested that z be three-valued, with an intermediate value denoting 'X may be in A'. In the egg-yolk model of Cohn and Gotts [5] the 'yolk' is 'in A', the 'white' is 'may be in A', and the two sets together form the 'egg'. Since z has only three values in this model, its spatial variation can be treated as a simple

variant of the two-valued case, using standard vector or raster representations.

In the more general case, however, the scale of z is continuous. It could be interpreted as a probability, $0 \le z \le 1$, either strictly frequentist as the probability that a randomly chosen person or observer would assign the point to A, or subjectively as a measure of the willingness of an observer to assign the point to A. It could also be interpreted as a measure of the membership of X in the fuzzy set A [26]. The distinctions between 'probable' and 'fuzzy' are profound, and are not explored here (see, for example, [8], [9]). Instead, we assume here that z is always confined to the interval [0,1], and interpreted in one of two ways: as the proportion of times something occurred (the *frequentist* view), or as a subjective assessment (the *subjectivist* view).

In principle, representation of a field $z(X)$ in a digital store can require an infinite amount of space, if the value of the field is independent at every location. In practice, however, geographic fields tend to be strongly autocorrelated, so that effective representations can be constructed in limited space, and it is often sufficient to store only a generalized version of the field, ignoring more detailed variation. Six standard field representations are commonly supported in GIS databases [10]:

- point samples over a rectangular grid;
- averages over a raster of rectangular cells;
- samples at irregularly-spaced points;
- averages over irregularly-shaped polygons;
- digitized isolines; and
- values at the nodes of a triangular mesh.

Others, such as the triangular and quadrilateral meshes commonly used for solving partial differential equations with finite element methods [21], are not supported by GIS. But these GIS methods have been devised for the representation of general surfaces with no assumed characteristics, other than strong spatial autocorrelation. Much more efficient methods of storage could be devised if $z(X)$ could be shown to follow certain general rules.

Models implemented in this research

We have focused in this research on three models. For the general case, in which no assumptions are made about the form of $z(X)$, we use a uniform raster, assigning a value of z to each cell. We term this *Model A*, and implement it in a frequentist interpretation in which $z(X)$ is the proportion of a sample of respondents reporting that X lay inside their conceptualization of the region of interest. *Model B* is a version of the Mark and Csillag [19] monotonic model, in which a crisp boundary is blurred by a mathematical function. For computational reasons we have implemented a simple linear function; thus $z(X) = 0.5+d/2\varepsilon$ for X inside the region and $d<\varepsilon$; $z(X)=1$ for X inside the region and $d \ge \varepsilon$; $z(X) = 0.5-d/2\varepsilon$ for X outside the region and $d<\varepsilon$; and $z(X)=0$ for X outside the region and $d \ge \varepsilon$; and ε is a parameter of the region [17]. *Model C* is the simplest model, and identical to Model B except that the region is circular. Define a central point Y, and the distance from X to Y as c. Then $z(X) = 0$ for $c \ge r+\varepsilon$; $z(X) = 1-(c+\varepsilon-r)/2\varepsilon$ for $r-\varepsilon<c<r+\varepsilon$; and $z(X)=0$ for $c \le r-\varepsilon$, where ε is a parameter of the region and r is the radius of the circle at which $z(X)=0.5$.

Visual display

In this section we discuss alternative methods for displaying $z(X)$, the representation of a fuzzy region. A common cartographic technique for indicating uncertainty is to replace a solid line with dashes. A dashed stream conventionally indicates uncertainty about its flow rather than about its position, while a dashed political boundary may indicate uncertainty of position because of dispute between the bordering jurisdictions. A dashed region boundary would be easy to draw, but it would not communicate the amount of positional uncertainty or anything about the form of the $z(X)$ surface.

Any of the cartographic techniques for representing scalar fields would be appropriate, including all of the methods used to represent the surface of ground elevation [2], [16], [18], [20], although some of these may be infeasible at the comparatively coarse densities available on today's display units (dot-density methods, for example, are much easier to implement for printed documents with much higher dot density). The choice will depend to some extent on whether other information must be presented at the same time, since there are strict limits on the ability of the eye to interpret visual information, and on the ability of displays to portray multiple items of information simultaneously. For example, suppose that two facts A and B must be portrayed at a single location X in a display. If B dominates, then the presence of A must be inferred; for example, roads can be displayed over colors indicating geologic conditions because roads are comparatively thin objects, and the eye will infer geologic conditions from the colors shown on either side of the road. Attempts to use color to display multiple conditions (e.g., use green to display the existence of both blue and yellow) are rarely successful because of the eye's limited ability to decompose color into its primitive components [24].

In the digital library case, it will often be desirable to

display a fuzzy region simultaneously with a base map of the area that includes such features as major roads, coastline, and rivers. It may also be desirable to display a fuzzy region used as a query at the same time as the fuzzy footprint of an IBO stored in the library; consider, for example, a search for information on 'downtown Santa Barbara' that finds a guide to 'lower State Street', an ill-defined part of downtown.

In general, two fields can be displayed simultaneously provided at least one is shown using an isoline (contour) representation, or some other representation that makes use of thin lines. Thus the query region could be shown using isolines, and the IBO's footprint using continuously-shaded grey fill. But it would not be feasible to display both fields using continuously-shaded fill. It would be feasible to overlay a base map of thin objects provided their colors were different from those of the fill.

Goodness of fit

Assume now that an IBO has been defined by a fuzzy region, such that $p(X)$ defines the probability that X lies in the region. Assume further that this IBO has been identified by a digital search mechanism as of possible interest to a user looking for information about a fuzzy query region defined by a probability field r, where $r(X)$ is the probability that X lies in the region of the user's interest. In other words, the IBO is a possible 'hit' for the query. Needed is a measure of the 'goodness of fit', or an estimate of the degree to which the IBO fits the query.

In comparing well-defined IBO footprints with well-defined query regions a suitable measure of goodness of fit is the ratio of the intersection to the square root of the product of the footprint and region areas. This will be 1 when the footprint and region coincide; small if the footprint occupies only a small part of the query region; and small if the query region occupies only a small part of the footprint. For footprints and regions defined by polygons it is also readily computed using standard vector overlay techniques.

When either footprint or query region or both are fuzzy, we propose the measure:

$$\int prdA \Big/ \left[\int pdA \int rdA\right]^{1/2}$$

which generalizes the measure for well-defined regions to the ill-defined case. The three integrals will be estimated from fuzzy region representations using various methods, depending on the digital representation being used. For example, for a pixel-based representation the integral will be approximated by the sum; for other cases it may be efficient to rasterize first and then sum, or to compute the integral analytically.

Conclusion

We have proposed methods for searching digital spatial data libraries that extend to the case of ill-defined geographic footprints. Such cases are common when information is cataloged or searched using vernacular place-names, rather than the officially-recognized place-names that tend to populate gazetteers. We have implemented these methods successfully within the framework of the Alexandria Digital Library, a project to implement a digital spatial data library for geo-referenced materials. We have proposed three approaches to the representation of ill-defined regions. The methods include tools for visualization of fuzzy footprints, and for measuring the goodness of fit of IBOs to query regions when either or both are ill-defined.

Not discussed in this paper, but also an important part of our research, are methods for eliciting definitions of fuzzy regions from users of the library, and from experts.

Acknowledgment

The National Center for Geographic Information and Analysis and the Alexandria Digital Library project are supported by the National Science Foundation, NASA, and the Advanced Research Projects Agency.

References

[1] M.J. Blakemore. Generalization and error in spatial databases, *Cartographica*, 21:131–139, 1984.

[2] D. Brandes. Sources for relief representation techniques, *The Cartographic Journal*, 20:87–94, 1983

[3] P.A. Burrough. Fuzzy mathematical models for soil survey and land evaluation, *Journal of Soil Science*, 40:477–492, 1989.

[4] P.A. Burrough and A.U. Frank, editors. *Geographic Objects with Indeterminate Boundaries*, Taylor and Francis, London, 1996.

[5] A.G. Cohn and N.M. Gotts. The 'egg-yolk' representation of regions with indeterminate boundaries, in P.A. Burrough and A.U. Frank, editors, *Geographic Objects with Indeterminate Boundaries*. Taylor and Francis, London, pp. 171–188, 1996

[6] C.R. Ehlschlaeger, A.M. Shortridge, and M.F. Goodchild. Visualizing spatial data uncertainty using animation, *Computers and Geosciences*, 23(4):387–395, 1997.

[7] P.F. Fisher. First experiments in viewshed uncertainty: simulating the fuzzy viewshed. *Photogrammetric Engineering and Remote Sensing*, 58:345–352, 1992.

[8] P.F. Fisher. Probable and fuzzy models of the viewshed operation, in M.F. Worboys, editor, *Innovations in GIS 1*, Taylor and Francis, London, pp. 161–175, 1994

[9] P.F. Fisher. Boolean and fuzzy regions, in P.A. Burrough and A.U. Frank, editors, *Geographic Objects with Indeterminate Boundaries*, Taylor and Francis, London, pp. 87–94, 1996

[10] M.F. Goodchild. The state of GIS for environmental problem-solving, in M.F. Goodchild, B.O. Parks, and L.T. Steyaert, editors, *Environmental Modeling with GIS*, Oxrod University Press, New York, pp. 8–15, 1993.

[11] M.F. Goodchild. The geolibrary, in S. Carver, editor, *Innovations in GIS V*. Taylor and Francis, London, 1998.

[12] M.F. Goodchild and S. Gopal, editors. *Accuracy of Spatial Databases*. London: Taylor and Francis.

[13] M.F. Goodchild, G. Sun, and S. Yang. Development and test of an error model for categorical data, *International Journal of Geographical Information Systems*, 6(2):87–104, 1992.

[14] G.J. Hunter and M.F. Goodchild. A new model for handling vector data uncertainty in geographic information systems. Journal of the Urban and Regional Information Systems Association, 8(1):51-57, 1996.

[15] H.T. Kiiveri. Assessing, representing, and transmitting positional uncertainty in maps, *International Journal of Geographical Information Science*, 11(1):33–52, 1997.

[16] M.P. Kumler and R.E. Groop. Continuous-tone mapping of smooth surfaces, *Cartography and Geographic Information Systems*, 17:279–289, 1990.

[17] P. Lagacherie, P. Andrieux, and R. Bouzigues. Fuzziness and uncertainty of soil boundaries: from reality to coding in GIS, in P.A. Burrough and A.U. Frank, editors, *Geographic Objects with Indeterminate Boundaries*, Taylor and Francis, London, pp. 275–286, 1996.

[18] S. Lavin. Mapping continuous geographical distributions using dot-density shading, *The American Cartographer*, 13:140–150, 1986.

[19] D.M. Mark and F. Csillag. The nature of boundaries on 'area-class' maps, *Cartographica*, 26:65–78, 1989.

[20] R.K. Phillips. Experimental method in cartographic communication: research on relief maps, *Cartographica*, 21:120–128, 1984.

[21] L.J. Segerlind. *Applied Finite Element Analysis*, Wiley, New York, 1976.

[22] T.R. Smith, D. Andresen, L. Carver, R. Dolin, and others. A digital library for geographically referenced materials, *Computer*, 29(7):14, 1996.

[23] Times Books and Bartholomew. *The Times Atlas of the World*, Eighth Comprehensive Edition, Times Books, New York, 1990.

[24] E.R. Tufte. *The Visual Display Of Quantitative Information*, Graphics Press, Cheshire, Conn., 1983.

[25] M.F. Worboys. *GIS: A Computing Perspective*, Taylor and Francis, London, 1995.

[26] Zadeh, L.A. (1965) Fuzzy sets. *Information and Control* 8: 338-353.