A Sketch-based Language for Representing Uncertainty in the Locations of Origin of Herbarium Specimens

Barry J. Kronenfeld
Department of Geography and Geoinformation Science
George Mason University
Fairfax, Virginia, USA
bkronenf@gmu.edu

Andrea Weeks
Department of Environmental Science and Policy &
Ted R. Bradley Herbarium - George Mason University
Fairfax, Virginia, USA
aweeks3@gmu.edu

***Abstract*—Uncertainty fields have been suggested as an appropriate model for retrospective georeferencing of herbarium specimens. Previous work has focused only on automated data capture methods, but techniques for manual data specification may be able to harness human spatial cognition skills to quickly interpret complex spatial propositions. This paper develops a formal modeling language by which location uncertainty fields can be derived from manually sketched features. The language consists of low-level specification of critical probability isolines from which a surface can be uniquely derived, and high-level specification of features and predicates from which low-level isolines can be derived. In a case study, five specimens of *Kolsteletzkya pentacarpos* housed in the Ted Bradley Herbarium at George Mason University are retrospectively georeferenced, and locational uncertainties of error distance, possibility region and uncertainty field representations are compared.**

Keywords: Herbarium databases; Retrospective georeferencing; Uncertainty fields;

I. INTRODUCTION

Herbaria store archivally-prepared, dried and pressed specimens of plant species and accumulate deep historical records about the flora of particular geographic areas. An important element of herbarium databases is the textual description of the location at which each specimen was collected. The process of translating these descriptions into quantitative data, including best estimates of the geographic coordinates of the collection location as well as associated uncertainty, is referred to as retrospective georeferencing (Murphey et al. 2004, Beaman et al. 2004).

Retrospective georeferencing is important in facilitating biologists in returning to the precise field site of collection and creating distribution models used in biogeographical analyses. For instance, herbarium specimen locality information contributes to the granularity of detail in species distribution maps, forecasting climate-induced range-shifts of vegetation types (Téllez-Valdés et al. 2006), and locating previously unknown populations of rare species (Ferriera de Siqueria et al. 2009). Recent studies (Graham et al. 2008; Fernandez et al. 2009) have quantified substantial effects of locational uncertainty on ecological niche models. To improve such models and to assess data usability and model error, there has been much interest in developing methods to capture and record locational uncertainty.

In this paper, we present a modeling language for sketch-based specification of an uncertainty field indicating the probability distribution of a specimen location. Our approach differs from previous work on uncertainty fields (e.g. Liu et al. 2009) in that its aim is to enable manual data specification rather than automatic parsing of location descriptions. Building on the practice of specifying a set of possible locations with a polygon sketch (Proctor 2004), we use point, line and polygon sketches to specify a full probability surface. In a case study, the approach is demonstrated for five specimens of Virginia saltmarsh mallow (*Kosteletzkya pentacarpos* (L.) Ledeb.; Malvaceae), and the potential benefits of the modeling language are computed.

II. BACKGROUND

At least four models have been proposed to represent uncertainty in georeferenced specimen locations: (a) qualitatively defined confidence values (Murphey et al. 2004), (b) error distances (Chapman 2005), (c) possibility regions (Proctor 2004), and (d) uncertainty fields (Liu et al. 2009). Current best practices recommend that numerical error distances be recorded, but there is considerable interest in uncertainty probability surfaces that would more fully capture the information contained in textual descriptions (Chapman and Wieczorek 2006).

In addition to data model, a choice must be made between using manual or automated methods to perform georeferencing. Efforts to develop automated methods are justified by the sheer number of herbarium specimens in museums around the world (BioGeomancer Working Group 2007). However, despite work to develop operational elements (e.g. Liu et al. 2009), full automation is a long way off, and may ultimately be less accurate and less efficient, than protocols involving semi-automated assistance to human-guided georeferencing (Murphey et al. 2004). On the other hand, at present no language exists to manually specify an uncertainty field. Our aim is to fill this gap.

The base for our proposed sketch-based modeling language is the common cartographic technique of drawing isolines connecting points of equal value to depict

Kronenfeld, BJ and A Weeks. 2010. A sketch-based language for representing uncertainty in the locations of origin of herbarium specimens. In N. Tate and P. Fisher (eds.), *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Leicester, UK: University of Leicester, pp. 237-240.

topographic and other surfaces. The "egg yolk" model of spatial vagueness (Cohn and Gotts 1996) builds on this concept by specifying lines that represent transitions between zones of definite inclusion (the "yolk"), vagueness (the "white), and definite exclusion. Zhan and Lin (2003) utilize the egg yolk model to conceptually represent fuzzy polygons, implicitly defining numerical membership values for the inner and outer boundaries. To interpolate values at locations in between isolines, methods built from the medial axis (Blum 1967) have been developed for elevation contours (Thibault and Gold 2000) and gradation between categorical regions (Kronenfeld 2007).

In adapting these methods, two unique characteristics of location uncertainty fields are apparent. First, unlike a topographic surface, the height of a probability density function cannot be directly observed. Instead, the herbarium specialist is more likely to be more comfortable specifying cumulative probabilities, which are constrained to sum to 100% across all locations.

Second, uncertainty fields are themselves uncertain, leading to recursive logic. Just as any statement about vagueness must itself be vague (Fisher et al. 2007), any representation of the uncertainty of a location must itself be uncertain; we refer to this phenomenon as "higher-order uncertainty". The existence of higher-order uncertainty does not diminish the importance of modeling lower-order uncertainty, but it suggests that more emphasis should be placed on semantic transparency, simplicity and efficiency of implementation.

III. Proposed Sketch-based Modeling Language

Our aim was to create a modeling language based on sketched isolines that would be intuitive and easy to implement, but also robust and flexible enough to enable a wide variety of input. The proposed language uses sketches and simple parameters that can be stored in a relational database to represent an uncertainty field, which can later be translated into a triangulated irregular network (TIN) or raster representation.

Using a computer programming analogy, the modeling language is described in terms of two interpretative components: a *low-level* language that is directly translatable into a probability surface, and a *high-level* language that encapsulates common concepts and translates them into the low-level language.

A. Low-Level Language

The low-level language requires specification of two nested polygons, which are referred to as the *core region* (CR) and the *bounding region* (BR). The CR is defined as the set of locations at which the probability density function reaches its maximum value, and may be a true or degenerate polygon (i.e. a line or point). The BR is defined as the set of all locations within which the value of the probability density function is greater than zero. The BR cannot be degenerate, and must completely contain the core.

An example of a core/boundary specification and its resultant probability density surface is shown in Figure 1.

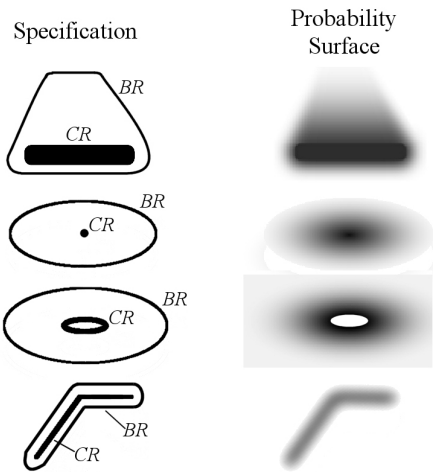


Figure 1. Illustration of sample core region (CR) and bounding region (CR) pairs and resultant probability surfaces.

Within the CR the probability density function is uniform. The intersection of the BR and the complement of the CR forms a transition zone, within which the probability density function decreases in a linear fashion from the maximum value to zero. Note that the edges of the CR and BR may be coincident.

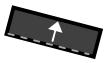
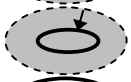
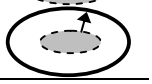
Calculation of the probability density at any point within the BR is a two-step process. Following Kronenfeld (2007), an initial surface is created by assigning values of 0 and 1 to the edges of the BR and CR, respectively, and of ½ to all points along the medial axis of the transition zone; in-between values are interpolated along straight lines connecting each medial axis point to the nearest edge. Once the initial surface has been created, the second step is to rescale the height of the surface so that the cumulative probability sums to unity.

B. High-Level Language

Although the low-level language is designed to be intuitive, specification of the CR and BR may be cumbersome and/or redundant. Text descriptions often include offset distances and spatial relationships to other elements; using these relationships, the BR can be derived geometrically from the CR or vice versa. For example, suppose that a specimen is believed to have been found within 100m of a road. In this case, the modeler should only need to manually sketch one element (the road) representing the CR; the BR can be derived automatically by applying a 100m buffer.

With this in mind, a high-level language is proposed to define each CR and BR from a base feature and an optional predicate with associated parameters. Eight predicates are proposed initially: *center*, *skeleton*, *endpoint*, *expand*, *expandToSide*, *contract*, *insideEdge*, *outsideEdge* (Table 1). Some predicates can only be applied to certain types of base features; for example, the *skeleton* predicate can only be applied to a polygon feature. Predicates may have required input parameters (not shown); for example, the *expandToSide* predicate requires a distance input as well as a

TABLE I. PREDICATES FOR DERIVING CORE REGION AND BOUNDING REGION FROM BASE FEATURES

	Name	Input	Example[a]
Dimensionality Reduction	center	line, polygon	
	skeleton	polygon	
	endpoint	line	
Buffer	expand	any	
	expandToSide	line	
	contract	line, polygon	
	insideEdge	polygon	
	outsideEdge	any	

a. base feature: dashed outline, light fill; derived feature: bold outline and fill

specification of which side of the line to expand. Eventually we envision development of a richer predicate vocabulary, including feature extraction (e.g. clipping) and predicate chaining, embedded within a visual interface environment.

IV. CASE STUDY

Five herbarium specimens of *Kosteletzkya pentacarpos*, were selected for analysis from a larger collection of specimens currently on loan to the Ted R. Bradley herbarium (standardized acronym: GMUF) for a taxonomic revision of the species. Specimens selected derive from the state of Virginia, USA and represent the range in specificity of text-based locality information that is typically found in the labels of herbarium specimens. The collector's name and collection number, herbarium of origin, and locality information for each specimen is as follows: 1) Crouch 459 (UNC) College of William and Mary; N side Papermill Creek, at mouth ofthe westernmost of the 3 tributary ravines N of creek between S Henry St and Colonial Parkway; 2) Fernald 12736 (PH) Surry Co., Cobham Bay, James River, NW of Chippokes; 3) Salle 522 (BRIT) York Co., ca 0.25 mile east of Indian Field Creek on northern side of Yorktown Colonial Parkway; 4) Schuyler 7146 (PH) Stafford Co., north of Widewater Beach, east side of Aquia Creek; 5) Wright s.n. (ODU); Seashore State Park, western edge of park.

Sketch-based models representing the uncertainty of the original location of each specimen were developed using a three-step process. Approximate latitude/longitude coordinates and the associated error radius for each locality description were estimated using Biogeomancer Workbench (http://bg.berkeley.edu/latest/). These coordinates were then used to access four geographical datasets simultaneously in Topofusion (http:// www.topofusion.com). The relevant United States Geological Survey topological maps, aerial photographs, U.S. Census data with road names, and land ownership categories were displayed for each coordinate. Once the relevant map region was displayed, the probable location of the specimen was sought. The "draw track" function was then used to create either a polygon or a line to encompass/delineate the probable location of the specimen based on the textual information, manmade and naturally-occuring boundaries apparent from map data, and our existing knowledge about the ecological limitation of *K. pentacarpos* to brackish or saline wetland areas. Once drawn, tracks were saved as ESRI shapefiles.

To compare levels of uncertainty using the numerical error distance, possibility region, and uncertainty field models, four reference areas (*RA*) were measured for each specimen. Two circular *RA*s were calculated: one using the numerical error distances returned by Biogeomancer, and a second determined from the smallest circle fully enclosing the manual sketch for each specimen. A third *RA* was calculated from the BR to represent the possibility region that would be produced from a simple polygon sketch. A possibility region can be defined as a homogeneous uncertainty field where the probability density function p is everywhere equal to $1/RA$. Therefore, we defined an *RA* for the uncertainty field as:

$$RA = \frac{1}{p_{max}}$$

where p_{CR} is the probability density function within the CR. The value of p_{max}, and therefore *RA*, depends on the specific configuration of the CR and BR, but is bounded by:

$$\frac{A_{BR} + \sqrt{A_{BR}A_{BR}}}{3} \leq RA \leq \frac{A_{BR} + A_{CR}}{2}$$

The lower bound occurs when the CR and BR are concentric circles, while the upper bound occurs when the transition zone is rectangular.

V. RESULTS

*RA*s for each of the five specimens using automated and manual error distances, possibility regions and uncertainty fields are listed in Table 2. Biogeomancer was unable to incorporate all information from the text descriptions, which generally resulted in extremely large *RA*s from automated georeferencing. In the most extreme example, the Crouch specimen's description had to be reduced to "Williamsburg; Papermill Creek, Virginia, USA" because Biogeomancer was not able to interpret street names or tributary ravines. This resulted in an error distance of nearly 5km, compared to a similarly defined error distance of 132m derived from the manual sketch. The human interpreter was also able to take advantage of contextual information to reduce uncertainty. For example, the interpreter determined that the Schuyler specimen, found along a shoreline, would not have been located past a train bridge because Schuyler would almost certainly have noted the train bridge if s/he had crossed it. In one case (Figure 2), the BioGeomancer *RA* was smaller than that of the manual sketch, but this appears to be the result of a data error in the recorded area of an input feature.

TABLE II. REFERENCE AREAS (*RA*S) FOR *KOSTELETZKYA* SPECIMENS

ID	Reference Area (km^2)			
	Error Distance (auto)	***Error Distance (manual)***	***Possibility Region***	***Uncertainty Field***
Crouch	70.972	0.055	0.022	0.009
Fernald	29.417	19.074	0.107	0.054
Salle	1.161	0.102	0.036	0.036
Schuyler	38.815	2.125	0.036	0.018
Wright	0.913	3.067	1.426	0.777

On average, the *RA* for the possibility region was 75% smaller than that of the error distance model (Table 2). This reduction in uncertainty varied considerably from specimen to specimen, however, and was highest for specimens located along linear features such as shorelines and rivers. The *RA* of the uncertainty field was an additional 41% smaller than that of the possibility region (Table 2). This reduction in uncertainty resulted from being able to assert a higher likelihood of occurrence in certain parts of the possibility region than others. For example, the Wright specimen (Figure 2) was believed to occur along a shoreline in the western part of a state park; although a wide area was possible, the middle ground between two shorelines was assigned a lower probability than the shorelines themselves. In one of the five cases (Salle), no reduction in the *RA* was achieved by the uncertainty field model because the specimen was considered equally likely to occur anywhere within the possibility region.

VI. CONCLUSIONS

Using the proposed modeling language, we were able to specify uncertainty fields for five herbarium specimens with relative ease. The resulting effective *RA* of the resulting representations was smaller by nearly 40% on average than that of the corresponding possibility region. This reduction in area may potentially reduce error in ecological niche models by reducing variance in predicted environmental characteristics at each specimen's location of origin. Whatever gain is accrued should come at low cost because specification of the uncertainty field often requires only a single sketched feature, which in some cases is the possibility region itself. However, implementation will require custom software development because the required functionality is not contained in standard geographic information systems.

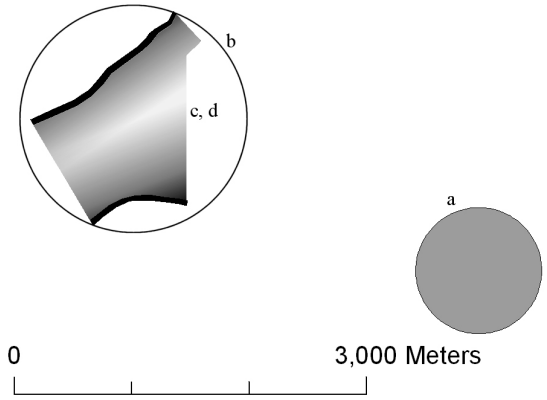


Figure 2. References areas for the Wright specimen: (a) Biogeomancer results; (b) manual sketch uncertainty distance, (c) possibility region, and (d) uncertainty field.

REFERENCES

Beaman, R., J. Wieczorek and S. Blum. 2004. Determining space from place for natural history collections. *D-Lib Magazine*, 10(5). 8pp.

BioGeomancer Working Group. 2007. *BioGeomancer*. http://www.biogeomancer.org (last accessed Sep. 21, 2009)

Blum, H. A transformation for extracting new descriptors of shape. In W.W. Dunn (ed.), *Models for the Perception of Speech and Visual Form*. MIT Press, Cambridge, USA.

Chapman, A. 2005. *Principles of Data Quality*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 58 pp. ISBN: 87-92020-03-8 (available as a standalone PDF from http://www.gbif.org)

Chapman, A.D. and J. Wieczorek (eds.). 2006. *Guide to Best Practices for Georeferencing*. Copenhagen: Global Biodiversity Information Facility

Cohn, A.G. and N.M. Gotts. 1996. The 'egg-yolk' representation of regions with indeterminate boundaries. In P.A. Burrough and A.U. Frank, eds., *Geographic Objects with Indeterminate Boundaries*. GISDATA Series, Taylor and Francis, London.

Fernandez, M.A., S.D. Blum, S. Reichle, Q. Guo, B. Holzman and H. Hamilton. 2009. Locality uncertainty and the differential performance of four common niche-based modeling techniques. *Biodiversity Informatics*, 6:36-52.

Ferreira de Siqueira, M. G. Durigan, P. de Marco, Jr., A. T. Peterson. 2009. Something from nothing: Using landscape similarity and ecological niche modeling to find rare plant species. *Journal for Nature Conservation*. 17: 25-32.

Fisher, P., T. Cheng and J. Wood. 2007. Higher order vagueness in geogrpahical information: Empirical geographical population of type *n* fuzzy sets. *Geoinformatica* 11:311-330.

Graham, C.H., J. Elith, R.J. Hijmans, A. Guisan, A.T. Peterson, B.A. Loiselle and The Nceas Predicting Species Districutions Working Group, 2008. *Journal of Applied Ecology*, 45:239-247.

Kronenfeld, BJ. 2007. Triangulation of gradient polygons: A spatial data model for categorical fields. *Proceedings of the 8th International Conference On Spatial Information Theory, COSIT 2007, Melbourne, Australia, Sep. 19-23, 2007*. Springer Lecture Notes in Computer Science. Berlin, Germany.

Liu, Y., Q.H. Guo, J. Wieczorek and M.F. Goodchild. 2009. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, 23(11):1471-1501.

Murphey, P.C., R.P. Guralnick, R. Glaubitz, D. Neufeld and J.A. Ryan. 2004. Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiateive (Mapstedi). *PhyloInformatics* 3:1-29.

Proctor, E.J., S.D. Blum and G. Chaplin. 2004. *A Software Tool for Retrospectively Georeferencing Specimen Localities using ArcView®. http://researcharchive.calacademy.org/research/informatics/GeoRef/index.html (last accessed Mar. 30th, 2010).*

Téllez-Valdés, O., P. Dávida-Aranda R. Lira-Saade. 2006. The effects of climate change on the long-term conservation of *Fagus grandifolia* var. *mexicana*, an important species of the cloud forest in eastern Mexico. *Biodiversity and Conservation* 15: 1095–1107.

Thibault, D. and C.M. Gold. 2000. Terrain reconstruction from contours by skeleton construction. *Geoinformatica*, 4(4):349-373.

Zhan, F.B. and H. Lin. 2003. Overlay of two simple polygons with indeterminate boundaries. *Transactions in GIS*, 7(1):67-81.