

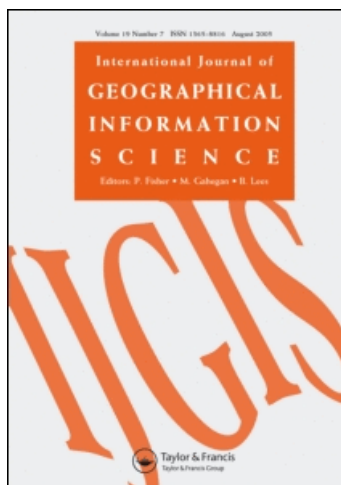
This article was downloaded by: [CDL Journals Account]

On: 26 August 2008

Access details: Access Details: [subscription number 785022370]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t71359799>

### Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach

Q. Guo <sup>a</sup>; Y. Liu <sup>ab</sup>; J. Wieczorek <sup>c</sup>

<sup>a</sup> School of Engineering, University of California Merced, Merced, USA <sup>b</sup> Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, PR China <sup>c</sup> Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, USA

First Published:2008

**To cite this Article** Guo, Q., Liu, Y. and Wieczorek, J.(2008)'Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach',International Journal of Geographical Information Science,22:10,1067 — 1090

**To link to this Article:** DOI: 10.1080/13658810701851420

**URL:** <http://dx.doi.org/10.1080/13658810701851420>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## Research Article

# Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach

Q. GUO\*†, Y. LIU†‡ and J. WIECZOREK§

†School of Engineering, University of California Merced, PO Box 2039, Merced, CA 95344, USA

‡Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, PR China

§Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720, USA

(Received 17 November 2007; in final form 17 November 2007)

Locality information for specimens of geological, biological, and cultural objects is traditionally stored as textual descriptions. With an increasing demand for natural and cultural information, the lack of spatially explicit descriptions has become a major barrier to the management and analysis of these data using geographic information systems. In this paper, we propose a method to georeference descriptive data, using an uncertainty field model to represent the distribution of a locality based on two types of uncertainties: uncertainty of reference objects, and the uncertainty of spatial relationships. We propose probability distributions for each known form of these two types of uncertainties and present a probabilistic method to georeference localities based on the integration of different uncertainty sources.

*Keywords:* Geographical information system; Spatial positioning; Georeferencing; Probability; Uncertainty; Textual descriptions

## 1. Introduction

It has been estimated that there are more than 2500 million specimens in natural history collections (Duckworth *et al.* 1993). With the increasing interest in understanding changes in environmental, biological, and cultural resources due to human disturbance and climate change, specimen collections have become ever more important, since they can provide baseline information on the environment and the factors driving change. Before the advent of geographical information systems (GISs) and global positioning systems (GPSs), occurrence information for most specimens was stored as textual descriptions without explicit geographic coordinates. This is a major obstacle for managing and analysing specimen data in a GIS. For example, at the beginning of the ‘Mammal Networked Information System’ Project (MaNIS 2001), which consists of a growing distributed database network of mammal collections data, 97.8% of the 296 737 distinct digitized collecting localities from the 17 participating collections had no coordinates. Descriptive localities have numerous sources of imprecision and other kinds of uncertainty (Wieczorek *et al.* 2004). Assessing and recording these uncertainties

---

\*Corresponding author. Email: qguo@ucmerced.edu

during the georeferencing process is arguably as important as determining coordinates for the locality, because only with the uncertainty can one determine if the location information is suitable for a particular analysis. Many efforts have been made to study the positional accuracy of spatial data (Goodchild and Hunter 1997, Leung and Yan 1998, Veregin 2000, Van Niel and McVicar 2002, Bonner *et al.* 2003). These studies try to assess the difference between test data and higher-accuracy 'true data'. These methods cannot be directly applied to estimate uncertainty while georeferencing specimen localities, however, because it is impossible to acquire higher accuracy 'true' spatial data for the millions of historically collected specimens that need retrospective georeferencing. Consequently, several practical georeferencing methods have been proposed. For example, a common approach is to use a bounding box (a rectangle) to encompass the locality being georeferenced. Recently, Wieczorek *et al.* (2004) proposed an alternative, point-radius method, which describes each locality as a circle where the radius represents the maximum error. One major advantage of this method over the bounding box is that the uncertainties can be readily combined into one attribute that is independent of geographic location, whereas the bounding box method requires contributions to uncertainty to be calculated independently in each of the two dimensions. Since the point-radius method describes a locality as a circle, this method has difficulty in dealing with some more complicated shapes and tends to overestimate the uncertainty in order to encompass completely the area in which the collection occurred. Therefore, in this research, we developed a new georeferencing method based on probability distributions of error sources that takes into consideration the shape of the locality being described.

### 1.1 *Related work on uncertainty in georeferencing*

Uncertainty is an inherent attribute of geographic information (Goodchild 2001). Shi (1998) believes that there are four aspects of uncertainty in GIS: positional, attribute, topological, and temporal. Positional uncertainty is the focus of this study on retrospective georeferencing based on textual descriptions.

It is possible that a textual description could contain errors, for example, '50 miles north of Merced' might mistakenly be recorded as '5 miles north of Merced'. In this research, we make the initial assumption that descriptions are correctly recorded in their textual form. During or after georeferencing, this assumption can be partially tested; some original errors may be detectable through spatial validation by testing for spatial consistency in the description, and through environmental outlier detection using statistical methods with related GIS layers.

In addition to measurement errors, recording errors, and the positional uncertainties associated with quantitative georeferencing (e.g. precision) uncertainties from the descriptive references to locations are inevitable and challenging to quantify. Referencing a named place (e.g. Yosemite National Park) is a typical example of qualitative georeferencing; it is an efficient way to communicate location information in everyday life, despite its inexact and sometimes ambiguous nature (Longley *et al.* 2005). In addition to named places alone, locality descriptions often combine place names and spatial relationships. For example, the description '5 kilometres north of Merced (CA) on Highway 99', includes two named places 'Merced, CA' and 'Highway 99', one metric relationship instance '5 kilometres', and one direction relationship 'north'. Compared with quantitative geographic

Table 1. Commonly encountered classes of locality descriptions based on occurrences in specimen records in the MaNIS project.

Locality type	Description	Example	Frequency (%)
F	Feature	Springfield	51.0
FOH	Offset from a feature (or a path) at a heading	10 km N of Kuala Lumpur	18.2
P	Path or linear feature	Hwy. 1	8.6
NF	Near a feature or path	Big Bay vicinity	6.2
FS	Subdivision of a feature or path	N part of Mono Lake	7.2
FOO	Orthogonal offsets from a feature	1 miles N, 3 miles W of Fairview	5.2
FH	Heading from a feature, no offset	W of Tucson	3.2
J	Junction	Confluence of Labarge Creek & South Labarge Creek	0.8
FO	Offset from a feature, no heading	5 km outside Calgary	0.4
BF	Between features or paths	Between Point Reyes and Inverness	0.2

coordinates, qualitative locality descriptions bear additional uncertainties that need to be addressed during the georeferencing process.

The most common types of georeferencible locality descriptions encountered in specimen records in the MaNIS project are shown in table 1. In Wieczorek *et al.* (2004), six factors were identified as contributing to the uncertainty of a georeferenced locality description and grouped into two categories:

Uncertainty of the referenced object:

1. extent of the locality;
2. unknown datum;
3. imprecision in coordinate measurements;
4. map accuracy.

Uncertainty of the spatial relationship:

5. imprecision in distance measurements;
6. imprecision in direction measurements.

Note that place name ambiguity (e.g. Springfield) is not in the list; in the absence of a definitive reference object, the locality references were considered too ambiguous for quantitative georeferencing. In practice, this problem is sometimes relatively easy to address with related evidence that suggests which of multiple possible reference objects to use. For example, an expedition would generally produce records from the same region, which could be used to isolate an appropriate named place.

One of the major challenges in calculating uncertainty for a locality is that many of the uncertainties in the above-mentioned list can affect a given description. An error-propagation method is necessary to deal with two or more uncertainty sources. Traditionally, the deterministic error propagation rule (Thapa and Bossler 1992) applies when the errors are in the form of standard errors. Standard error is not applicable to georeference locality descriptions retrospectively because there is no way to reconstruct a meaningful standard error without 'true' original data. In principle, one could construct patterns of error for specific locality types by accumulating sufficient locality descriptions of

that type that also had associated coordinates. Unfortunately, there are no systematic rules for how places with coordinates should be described. As a result, coordinates associated with a locality may be for the nearest reference object rather than the actual place. Without the original recorder of the data, we have no way to determine what data are 'true'. Because of this interesting situation, the point-radius method (Wieczorek *et al.* 2004) estimates the maximum uncertainty  $U_{\max}$  taking into consideration the interaction of all uncertainty sources  $\sum u_i + \sum u_d$ , where  $u$  is the uncertainty for independent (i) or dependent (d) sources of error. The point-radius method provides a liberal, but overestimated representation of the locality uncertainty.

There are two ways in which to reduce the overestimation and provide a potentially more specific and therefore more useful georeference. First, the point-radius method uses a circle to describe the possible distribution of a locality, which, in reality, is often irregular in shape. For example, the least bounding shape that satisfies the description 'five kilometres north of a point feature  $A$ ' is an arc rather than the circular region circumscribing that arc as prescribed by the point-radius method. In order to be more specific, therefore, we propose to develop a 'shape method' that has the potential to produce refined quantitative spatial descriptions of qualitative textual locality descriptions. The second way to improve upon the point-radius method is related to the fact that the point-radius method presents no means to distinguish any of the points within the circle as being any more or less likely to be a part of the locality; nor does the method take into account the probability distribution of different uncertainty sources. More often than not, the probability that an event occurred at any given point within the circle given by the point-radius method will not be uniformly distributed in the circle because uncertainties from different sources have different probability distributions. With the point-radius method, for example, points within  $45^\circ$  of north from points in  $A$  are considered to be 'north of  $A$ '. In reality, the further away from actual north a point lies, the less likely it would have been described as 'north'. We recognize that these improvements are complicated to implement without the aid of digital maps and specialized GIS software. This study describes the methods and software we developed to overcome this difficulty.

## 2. Conceptual framework for probabilistic georeferencing

### 2.1 Uncertainty field, reference object, and target object

In order to develop a probabilistic georeferencing approach, we first introduce the concept of the uncertainty field to represent the localities and their associated uncertainties. Field and object models have been widely accepted as two alternative approaches for conceptualizing and modelling geographic phenomena (Goodchild 1992). The field model is more suitable for determining uncertainty in the georeferencing process than the object model because the uncertainty boundaries are not crisp, and the probabilities may vary within the boundaries (Goodchild 1989, Couclelis 1996). For convenience of information storage, Tøssebro and Nygård (2002) proposed a discretization approach for uncertain features. If the uncertainty associated with a feature cannot be modelled in a deterministic way, as is the case for the georeferencing process, the raster model is a more appropriate way to represent the spatial variance in uncertainty. We can populate the region of the

uncertainty field with probability values and use these to generate an estimated shape as an implementation of the ‘shape method’ described by Wieczorek *et al.* (2004).

Most locality descriptions are based on at least one specific named place, which acts as the Reference Object (RO) for positioning a locality. The RO may be a point, linear, or areal feature, such as a junction, highway, or city. For simplicity, these objects are sometimes represented using a circle (Wieczorek *et al.* 2004) or a bounding box (Hill 2006), by which the actual shape of RO is circumscribed. The final shape containing the described locality is called the Target Object (TO). The objective of the georeferencing process is to estimate the TO based on the positions, shapes, and uncertainty of its ROs and spatial relationships.

## 2.2 Error propagation

There are two major error propagation approaches to calculate uncertainty from multiple sources: the analytical approach and the numerical approach (Burrough and McDonnell 1998). The square root of the sum of the squared errors is a commonly used analytical approach to model errors resulting from different uncertainty sources (Thapa and Bossler 1992). This approach requires the errors to be represented as standard deviations (or standard errors) and to follow a normal distribution. However, in the retrospective georeferencing process, the errors often do not conform to normal distributions or cannot be represented as standard errors (Wieczorek *et al.* 2004). Therefore, we opt to use a numerical approach, applying the Monte Carlo method, to calculate a locality’s uncertainty using the following steps:

1. Develop probability distribution functions for each uncertainty source.
2. Choose a starting-point in the reference object (this can be done randomly or systematically—with enough simulations the result will be similar).
3. Use the probability distribution function of each uncertainty source to generate the contribution to the location of the resulting point in the target object for this simulation.
4. Increment the count of simulation results for the raster cell corresponding to the point resulting from step 3.
5. Repeat the simulations (steps 2–4) many times to produce an uncertainty field for the target object.

## 2.3 Probability distributions of uncertainties

We use uncertainty fields to represent the probabilistic spatial distribution of TOs. For localities confined to a two-dimensional surface, the uncertainty field  $z$  of a TO can be defined as a two-dimensional probability density function (PDF):

$$z = p(x, y) \quad (1)$$

The probability of the TO inside a region  $R$  is  $\iint_R p(x, y) dx dy$ . Across the entire domain  $D$  of the PDF  $p(x, y)$ , the probability is 1 as given by the following equation:

$$\iint_D p(x, y) dx dy = 1 \quad (2)$$

For an uncertainty field, we can define a representative point  $O_R$  as:

$$O_R = (x_R, y_R) = \left( \iint_D p(x, y)x \, dx \, dy, \iint_D p(x, y)y \, dx \, dy \right) \quad (3)$$

In some cases, the probability density at  $O_R$  may be 0. There are three methods for the uncertainty measurement: maximum error, mean error, and a region with probability  $P$ .

**2.3.1 Maximum error.** Centring at  $O_R$ , a circular region covering all non-zero positions can be found. The radius of this circle is the maximum uncertainty:

$$U_{\max} = \max\{\text{dist}((x, y), (x_R, y_R)) | p(x, y) > 0\} \quad (4)$$

where  $\text{dist}$  is a function to calculate the Euclidean distance between two points.

**2.3.2 Mean error.** Mean error is the weighted average value of the distances between all non-zero positions and  $O_R$ :

$$U_{\text{mean}} = \iint_D \text{dist}((x, y), (x_R, y_R)) p(x, y) \, dx \, dy \quad (5)$$

where  $\text{dist}$  is a function to calculate the Euclidean distance between two points.

**2.3.3 Region with probability  $P$ .** It is possible to determine a region  $R_P$  of minimum area in an uncertainty field within which the position of the  $TO$  is distributed with a given probability  $P$ . The probability densities of every point inside  $R_P$  should be greater than or equal to the probability densities of the points outside  $R_P$ . The following two conditions should be satisfied for such a minimum area region  $R_P$ :

$$\iint_{R_P} p(x, y) \, dx \, dy = P \quad (6)$$

$$\forall (x_1, y_1) \in R_P, \forall (x_2, y_2) \notin R_P, p(x_1, y_1) \geq p(x_2, y_2) \quad (7)$$

$R_P$  need not be singly bounded, and there is no guarantee that there is a unique region satisfying these conditions. It is difficult to calculate  $R_P$  analytically for a continuous field, but if the field is represented in a raster format,  $R_P$  can be obtained by sorting the probability density values of all pixels, then summing the top probability pixel values iteratively until the net probability is greater than or equal to  $P$ .

Many factors (e.g. coordinate precision, map accuracy, datum) can contribute to the uncertainty of the reference object. We can use the probabilistic approach to determine the uncertainty of the RO and take into account error propagation. Assuming a point RO has  $n$  ( $n$  may be infinite) possible positions with probabilities  $q_i$  ( $1 \leq i \leq n$ ), an uncertainty field for the TO can be obtained from each of these starting positions. The probability of a field is  $q_i p_i(x, y)$ , where  $p_i(x, y)$  is the  $i$ th field associated with the  $i$ th possible position of the RO. By summing these  $n$  fields using equation (8), the uncertainty field of the TO can be obtained:



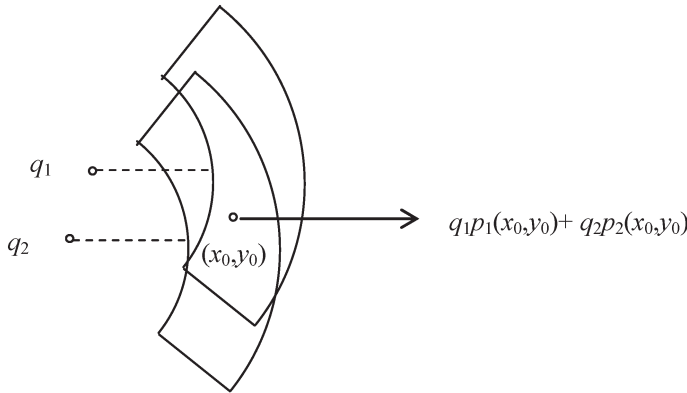


Figure 1. Summing uncertainty fields to represent probability.

$$p(x, y) = \sum_{i=1}^n q_i p_i(x, y) \quad (8)$$

Figure 1 shows such an overlay operation including two uncertainty fields. In the result field, the probability density at  $(x_0, y_0)$  is  $q_1 p_1(x_0, y_0) + q_2 p_2(x_0, y_0)$ . If there are infinite possible positions for the RO, equation (8) becomes:

$$p(x, y) = \iint_D q(u, v) p(x, y, u, v) du dv \quad (9)$$

where  $D$  is the distribution range of the RO.

### 3. Implementation of the probabilistic georeferencing approach

In the previous section, we presented a theoretical framework for the probabilistic georeferencing approach by introducing the uncertainty field, which provides an approach for georeferencing a locality and estimating the uncertainty. In the following discussion, we will present the detailed derivation and implementation of the proposed method. We use the specific example of a ‘distance at a heading’ (e.g. ‘10 km E of Berkeley’) to better explain and demonstrate the probabilistic referencing method. We choose the ‘distance at a heading’ case because: (1) it is represented in a large proportion (18.2%) of all locality descriptions and (2) it is one of the most complicated cases, as it can be affected by all six different uncertainty sources. In section 3.1, we will first investigate the uncertainty sources of a reference object and establish the uncertainty field ( $F_1$ ) of the starting-point. Then, the uncertainty field ( $F_2$ ) associated with the spatial relationship will be discussed in section 3.2. Finally, the uncertainty field of the TO will be obtained by integrating  $F_1$  and  $F_2$  using corresponding operations on the uncertainty field.

#### 3.1 Uncertainty associated with the reference object

Four major uncertainty sources contribute to the uncertainty of the reference object: spatial extent, map accuracy, coordinate precision, and unknown datum. Detailed discussions of these uncertainty sources are given in the following sections.



**3.1.1 Uncertainty due to spatial extent.** Though often represented as points in gazetteers, reference objects in reality have a non-point spatial extent, and the point of reference may be located at any position inside the reference object. In the example ‘10 km N of Merced city, CA’, the reference object (i.e. the starting-point) is the city of Merced within California. As is usually the case with descriptions of this type, the exact point of reference within Merced is not specified in this description; it could be the centre of Merced, the post office, the courthouse, some intersection, the northern border, or any other location in the city. Without further documentation, it is impossible to determine what the author really meant when the textual locality was written, and we therefore model the starting-point by assuming a uniform distribution within the bounds of the city of Merced. Note that, in most cases, it would be best to use the areal region of Merced from the time the locality description was written, but this kind of information can be difficult to obtain. Usually, a recent spatial representation of a RO is a reasonable approximation, especially for a populated place, since it is likely to contain the region of the RO at the time of recording, and the resulting TO will simply be an overestimation. It is a good idea to be cognizant of this problem and check that any particular case is not an exception. Note that if we do have documentation (metadata, field notes) suggesting a specific starting-point, then that point can be used as the RO and the spatial extent uncertainty can be reduced or non-existent. In general, the uniform PDF of the RO  $A$  is:

$$p_e(x, y) = \begin{cases} 1/S(A), & (x, y) \in A \\ 0, & (x, y) \notin A \end{cases} \quad (10)$$

where  $S(A)$  is the area of  $A$ , and  $(x, y)$  is a point with coordinates  $x$  and  $y$  in two-dimensional Euclidean space  $\mathbf{R}^2$ . If the shape of the RO cannot be determined (e.g. we may not be able to find the boundary, historical or otherwise, for a small town), we can still use some simplified measurements such as an estimated radius for the RO. In this case, the PDF of starting-point in the reference object is expressed as:

$$p_e(x, y, u, v) = \begin{cases} 1/\pi S^2, & \text{dist}((x, y), (u, v)) \leq S \\ 0, & \text{dist}((x, y), (u, v)) > S \end{cases} \quad (11)$$

where  $(u, v)$  is the coordinate of the central point of the RO and  $S$  is the radius of the RO as in figure 2.

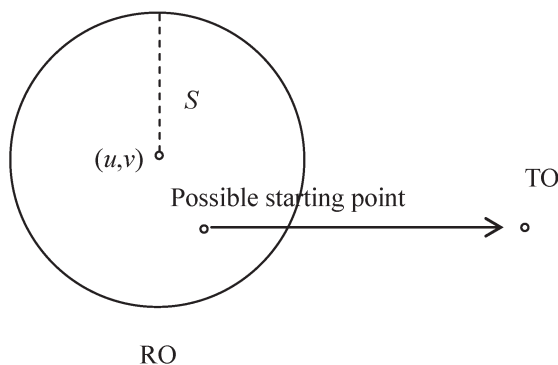


Figure 2. Spatial distribution of an RO estimated by a radius  $S$  from a central location. Every possible starting-point within the extent of the RO maps to a point in the TO.

Though we will focus on the simplified case of a point RO for demonstration purposes in the following section, the methods developed in this paper are suitable for any arbitrary shape by rasterizing the probability distribution, and have been implemented in the software described in section 4.

**3.1.2 Uncertainty due to map accuracy.** Maps have an inherent and sometimes specified level of accuracy. Generally, a large-scale map has greater accuracy than a small-scale map. According to the mapping standard of the United States Geological Survey (USGS 1947), ‘for maps on publication scales larger than 1:20 000, not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20 000 or smaller, 1/50 inch’. This specification does not give an analytical representation of the error distribution associated with map scale. Let  $E$  be the maximum uncertainty caused by the map accuracy at a given scale. For a 1:40 000 map,  $E$  would be 800 inches (i.e. 20.32 m). The probability distribution of a point derived from the map could be modelled as a circle with the radius  $E$ . The PDF of uncertainty due to map accuracy can be defined as a function of  $(u, v)$ .

$$p_a(u, v) = \begin{cases} 1/\pi E^2, \text{dist}((u, v), (o_x, o_y)) \leq E \\ 0, \text{dist}((u, v), (o_x, o_y)) > E \end{cases} \tag{12}$$

where  $(o_x, o_y)$  is the coordinate based on the locality description. When combining the uncertainty from the spatial extent (equation(11)) and map accuracy (equation (12)), the PDF of the starting-point can be expressed as:

$$p(x, y) = \iint_D p_e(x, y, u, v) p_a(u, v) du dv \tag{13}$$

where the integral domain is actually a circle defined as:

$$\text{dist}((u, v), (o_x, o_y)) \leq E \tag{14}$$

Finally, we obtain the PDF of the starting-point as:

$$p(x, y) = \begin{cases} \frac{3}{2\pi(E^2 + S^2 + |E^2 - S^2|)}, \text{dist}((x, y), (o_x, o_y)) \leq |E - S| \\ \frac{3((E+S) - \text{dist}((x, y), (o_x, o_y)))}{2\pi(E^2 + S^2 + |E^2 - S^2|)((E+S) - |E - S|)}, |E - S| < \text{dist}((x, y), (o_x, o_y)) \leq E + S \\ 0, \text{dist}((x, y), (o_x, o_y)) > E + S \end{cases} \tag{15}$$

A two-dimensional PDF can be discretized and stored in raster format to be managed in an information system. In this paper, we choose the raster model to store a two-dimensional PDF. Suppose a cell covers a square region  $C$ ; the value  $v$  of this cell is:

$$v = \iint_C p(x, y) d\sigma \tag{16}$$

where  $p(x, y)$  is the PDF and the pixel values vary in the interval  $[0,1]$ . For a raster field, the constraints given by equation (2) can be expressed as:

$$\sum p(i, j) = 1 \tag{17}$$

where  $p(i,j)$  is the pixel value at the  $i$ th row and  $j$ th column in the raster data. Let  $2S$  be the spatial extent of the reference object, and let  $E$  be the maximum error due to the map accuracy. Based on the discretized raster data, the PDF defined in equation (15) can be seen in figure 3.

**3.1.3 Uncertainty due to coordinate precision.** In Wieczorek *et al.* (2004), uncertainty due to the imprecision with which the original coordinates were recorded was estimated as follows:

$$\text{uncertainty} = \sqrt{\text{lat\_uncertainty}^2 + \text{long\_uncertainty}^2} \quad (18)$$

where

$$\text{lat\_uncertainty} = \pi \times R \times (\text{coordinate precision}) / 180.0$$

and

$$\text{long\_uncertainty} = \pi \times X \times (\text{coordinate precision}) / 180.0$$

where  $R$  is the radius of curvature of the meridian at the given latitude,  $X$  is the distance from the point to the polar axis, orthogonal to the polar axis, and coordinate precision is the precision with which the coordinates were recorded. Detailed calculations on  $R$  and  $X$  can be found in Wieczorek *et al.* (2004). Using equation (18), the maximum uncertainty due to coordinate precision is assumed to be due to rounding error and therefore is the same in all directions (in coordinate space). Suppose a locality, such as '30.1°N, 124.2°W', is recorded in decimal latitude and longitude to 0.1° of precision, the true latitude lies in the interval  $[\text{lat} - 0.05, \text{lat} + 0.05]$ , and the true longitude lies between  $[\text{long} - 0.05, \text{long} + 0.05]$ , where  $\text{lat}$  and  $\text{long}$  are the geographic coordinates of the locality. Thus, the PDF of the true position, based on precision alone, is uniformly distributed in a quadrangle with four vertices:  $(\text{lat} - 0.05, \text{long} - 0.05)$ ,  $(\text{lat} - 0.05, \text{long} + 0.05)$ ,  $(\text{lat} + 0.05, \text{long} - 0.05)$ , and  $(\text{lat} + 0.05, \text{long} + 0.05)$ .

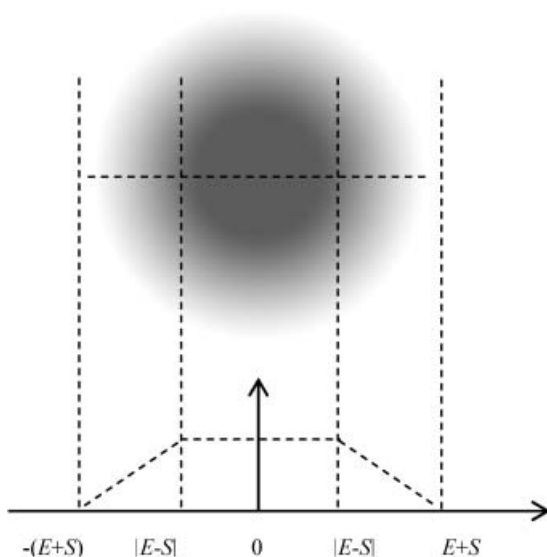


Figure 3. Probability distribution of the starting-point considering the uncertainty of spatial extent and map accuracy.

$-0.05$ ), and (lat  $+0.05$ , long  $+0.05$ ). Figure 4(a) demonstrates the uncertainty of a locality ( $-120^\circ$ ,  $38^\circ$ ) with  $1^\circ$  precision overlaying the county boundaries of the state of California. Note that because actual distances represented by  $1^\circ$  in latitude and longitude usually differ at any given point on the Earth, the shape of the uncertainty region caused by the latitude and longitude precision will not be a square in most projections. In practice, ROs defined by coordinate of low precision are of little utility and are generally discarded in favour of better sources; therefore, the latitude and longitude precisions are usually relatively high, and we can use a rectangle as the approximation of the possible distribution range of the true position. Figure 4(b) shows the PDF of the starting-point by combining the three sources of uncertainty: spatial extent of the RO, map accuracy, and coordinate precision.

**3.1.4 Uncertainty due to unknown datum.** A missing datum reference introduces a complicated ambiguity, which varies geographically (Welch and Homsey 1997). A simple example of the complications can be demonstrated for the USA. Since many currently available maps of North America are based on the North American Datum of 1927 (NAD27) or North American Datum of 1983 (NAD83), if a geographic coordinate record lacks datum information, the true datum could reasonably be either of them. Therefore, an error may be generated if we wrongly assign a datum to a locality with a missing datum. Wiczorek *et al.* (2004) point out that the uncertainty from not knowing which of these datums was used to determine the coordinates varies in the contiguous USA up to 104 m. In our current example, there are only two possible outcomes: either we assign the correct datum or we assign the wrong datum. We are not likely to have a reason to prefer one possibility to another, and so we would assign a probability of 0.5 to each point. Figure 5 illustrates the probability distribution of the starting-point of the RO by combining the four uncertainty factors, spatial extent of the RO, map accuracy, coordinate precisions, and datum. Note that figure 5 is exaggerated for demonstration purposes. In reality, the errors caused by unknown datum are often much smaller than errors caused by other sources, which would make the two quadrangles in

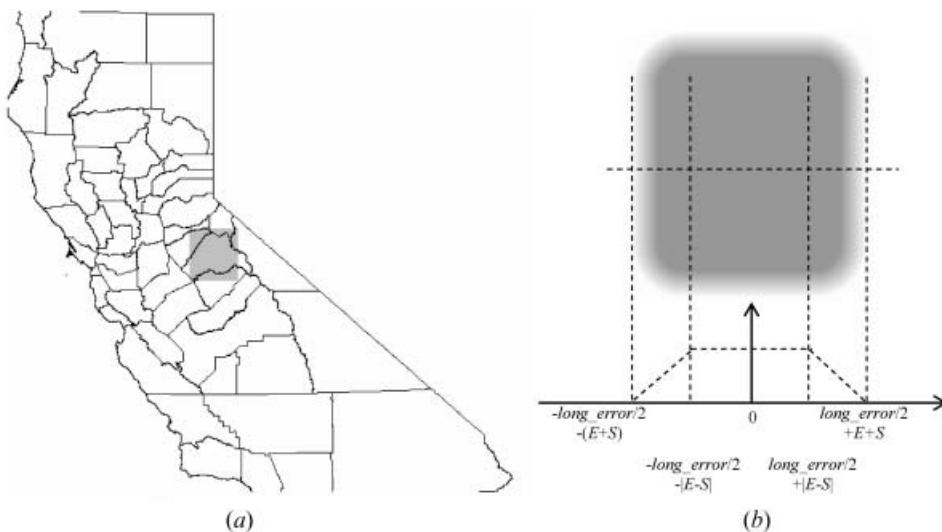


Figure 4. (a) Uncertainty from coordinate precision (b) Combining three sources of uncertainty of RO (spatial extent of RO, map accuracy, and coordinate precision).



Figure 5. Probability distribution of the starting-point when combining four uncertainty sources (spatial extent of RO, map accuracy, coordinate precision, and datum).

figure 5 overlap each other more closely. Moreover, the orientation of these two positions with respect to each other will be location-dependent.

### 3.2 *Uncertainty due to spatial relationships*

In addition to the reference object, spatial relationships used in the locality description are also important sources of the TO's uncertainty. The uncertainty caused by spatial relationships includes two factors: (1) uncertainty of direction and (2) uncertainty of distance.

Dutta (1991) and Du *et al.* (2004) described the vagueness of cardinal direction relationships and qualitative distance relationships. Topological relationships, although they are theoretically determinate (Egenhofer and Herring 1991, Randell *et al.* 1992), may be uncertain (Mark and Egenhofer 1994) in the context of natural language description. Egenhofer and Shariff (1998) propose quantitative indices to discriminate further between the topological relationships belonging to the same category, such as 'overlap'. Precision of distance relationships will also lead to uncertainty in georeferencing a locality. In most locality descriptions, distances are recorded with few or no significant digits to the right of the decimal, or even with fractions (Wieczorek *et al.* 2004). In the description '9 km N of Bakersfield', the true distance might be an arbitrary value in [8.5, 9.5) due to the rounding operation or due to human estimation, which we take to have the same effect. Note, however, that '500 miles North of Merced' may not represent the same precision as '5 miles North of Merced'. The first case might have 100 miles' uncertainty, while the latter only has 1 mile's uncertainty (half a mile to either side of 5 miles). Detailed discussion on how to treat precision can be found in Worboys Clementini (2001) and Wieczorek *et al.* (2004).

**3.2.1 Uncertainty of direction.** As shown in table 1, direction relationships play an important role in locality descriptions. Much research has been done on

representing and reasoning about direction relationships (Frank 1991, Freksa 1992, Montello and Frank 1996, Goyal and Egenhofer 2001, Skiadopoulos *et al.* 2004) as well as positioning based on direction (Clementini *et al.* 1997, Dehak *et al.* 2005). In addition to the work presented by these authors, Wieczorek *et al.* (2004) identified a complicating issue that must be considered in dealing with textual locality descriptions—whether a direction should be taken ‘by air’ and ‘by road’ should be distinguished during the georeferencing process. In the example ‘10 miles north of Merced’, if there is further evidence to suggest that the original textual description meant ‘10 miles north of Merced by road’ (or if the locality were to explicitly state that the directions are by road), then we would not use the probabilistic direction distribution. Instead, we would use the road as a constraint on the directional component of the locality. The same treatment would apply to descriptions that state, imply, or have further evidence to suggest a direction by any type of path, such as a river. In the case of ‘by air’, we do need to take into consideration direction uncertainty. For example, ‘north’ may not mean ‘due north’, but might rather mean ‘roughly north as opposed to east’. Wieczorek *et al.* (2004) describe north of the RO with a direction imprecision (e.g. 45°) where the actual locality lies anywhere in the region bounded by the given direction +/- the direction imprecision. One solution for a PDF is to make the distribution uniform within the region just described. Alternatively, the TO could be unevenly distributed, reflecting a greater likelihood that, for example, any point due north of a RO would be more likely to be called north than any points further away to either side. For this study, we developed a conceptual PDF associated with cardinal directions based on the eight-sector partition scheme (i.e. N, NE, E, SE, S, SW, W, and NW):

$$p = \begin{cases} p_{\max}, & \alpha < \pi/16 \\ (2 - 16\alpha/\pi)p_{\max}, & \pi/16 \leq \alpha \leq \pi/8 \\ 0, & \alpha > \pi/8 \end{cases} \quad (19)$$

where  $\alpha$  is the angle between that direction and the central axis of the corresponding cone, and  $p_{\max}$  is a constant depending on the search domain. Once the search domain is determined,  $p_{\max}$  can be computed using equation (2) to ensure that the sum of the PDF is 1. Figure 6 illustrates an example PDF for ‘north’. Note that the search domain is an important concept in georeferencing; auxiliary information associated with the locality description can sometimes constrain the TO. For example, it would be unreasonable to locate any part of the TO of a locality for an endemic California species inside Washington State based on the expression ‘north of San Francisco’. Normally, an explicit distance is used in a locality description to constrain the search domain, such as in the example ‘10 miles north of Merced’.

**3.2.2 Uncertainty of distance.** The distribution range of the TO resulting from the distance relationship uncertainty is an arc or band. If the distance uncertainty is due to measurement error, the error band model proposed by Tong *et al.* (2003) is suitable. Though direction relationships or distance relationships alone can provide only rough constraints on the target object, their combination can provide a more refined estimate (Clementini *et al.* 1997). Consequently, the uncertainty field can be obtained for a given starting-point (figure 7) based on the product of two associated uncertainty fields.

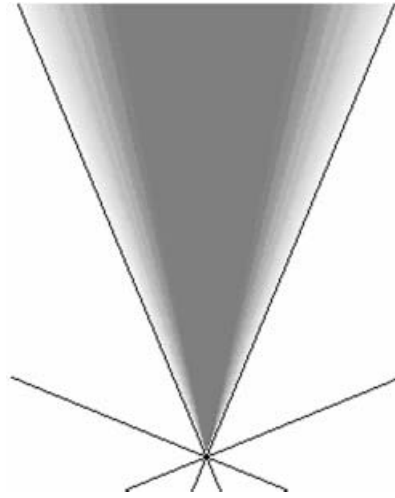


Figure 6. Probabilistic distribution of the direction relationship using the direction ‘north’ as an example.

#### 4. Case study and software implementation

Suppose we are trying to georeference the locality description ‘5.0 km NE of Colfax, CA’ using the probabilistic shape method. We have been given the coordinates  $-120.95, 39.10$  for Colfax from a 1:24 000 map without datum information, and from satellite imagery we can see that 1 km from that point seems to be the limit of the extent of the town. A summary of the essential information needed for georeferencing follows:

- Geographic coordinate: longitude= $-120.95$ , latitude= $39.10$ .
- Coordinate precision:  $0.01^\circ$ ; the error associated with this coordinate information is 1408 m (measured using the diagonal of the quadrangle shown in figure 4(a)).
- Spatial extent of the RO: 2 km (i.e. radius=1000 m).

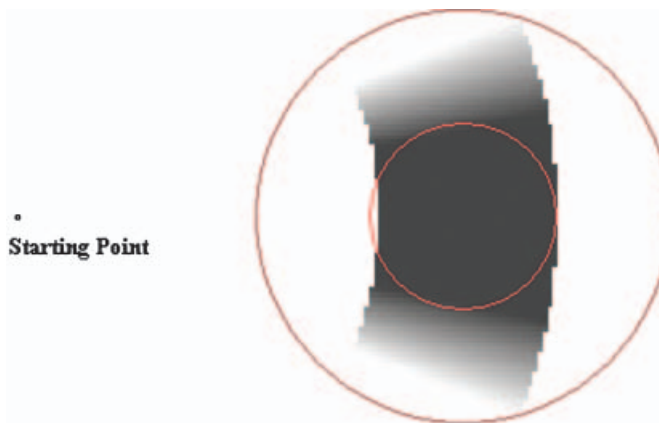


Figure 7. Uncertainty field derived from an uncertain direction relationship (east) and an uncertain distance relationship.



- Map scale and associated uncertainty: 1:24 000 and 12 m.
- Datum: unknown; assuming the possible datums could be NAD27 or NAD 83, the distance between the real positions of the point  $(-120.95, 39.10)$  in these two datums is 92 m using NADCON (National Geodetic Survey 1992).
- Direction: north-east.
- Distance and distance precision: 5000 m and 100 m.

Following section 3.1, the probability distribution of the starting-point is shown in figure 8(a). Meanwhile, figure 8(b) illustrates the probability distribution of the TO based on the distribution of the RO, the spatial relationship between them, and the associated precision of the relationship. The spatial distribution looks like a 'bean' with a diffuse boundary.

Figure 8 is based on calculations using the following projection parameters: 'Projection name=Albers; False\_Easting=0.00; False\_Northing=-4000000.00;

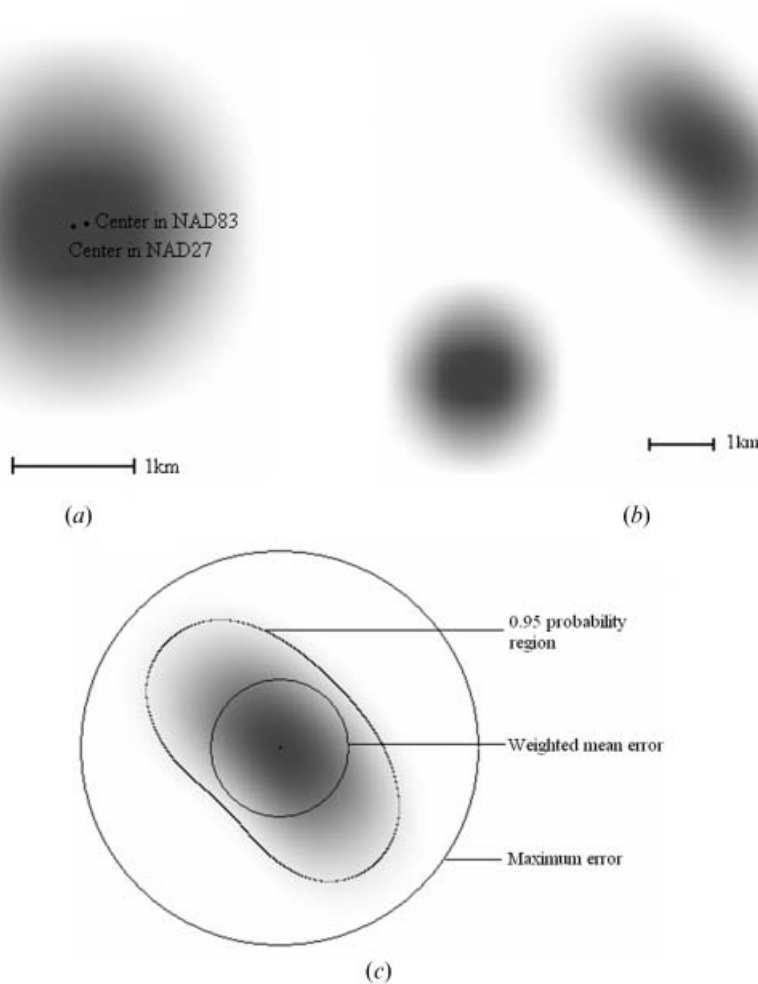


Figure 8. Probability distribution of reference object (a) with extent, coordinate, and datum uncertainty, (b) shown in relation to the target object taking into account distance and direction uncertainty, and (c) showing various derivative shapes to express the uncertainty.

Central\_Meridian=-120.00; Standard\_Parallel\_1=34.00; Standard\_Parallel\_2=40.50; Geographic Coordinate System=GCS\_North\_American\_1927'. The resolution of the raster data is 5.26 m. According to equations (4) and (5), the maximum error is 3.705 km (43.125 km<sup>2</sup> area), and the mean error is 1.213 km (4.622 km<sup>2</sup> area) (figure 8(c)). The area of the 0.95 probability region is 10.234 km<sup>2</sup>, while the whole non-zero probability region (i.e. 1.0 probability region) is 16.795 km<sup>2</sup>. If the maximum uncertainty is calculated using the method proposed by Wieczorek *et al.* (2004) with the same parameters, the result is 4.618 km (66.996 km<sup>2</sup> area). Clearly, the method developed in this paper leads to a smaller uncertainty associated with the TO.

The six uncertainty sources play different roles for determining the probability distribution of a TO. In this case study, the uncertainty resulting from the distance relationship depends on the distance precision (0.1 km), which affects the width of the 'bean'. Meanwhile, the uncertainty resulting from the direction relationship affects the length of the arc of the 'bean', which also increases with increasing distance from the RO. The coordinate precision and spatial extent of the RO result in the diffuse boundary of the 'bean' and contribute both to its length and to its width. Compared with the other four sources, the uncertainties resulting from an unknown datum and from map accuracy based on scale in this case study are small. Since they are usually less than 100 m in practice, their effect only becomes pronounced if the RO is relatively small.

#### 4.1 Toolbox for georeferencing and estimating associated uncertainty

In order to facilitate the use of the proposed method, we developed a software toolbox to georeference localities and estimate the associated uncertainties. The program was developed based on C++. The georeferencing steps are as follows:

1. Load reference maps: One can use one or more reference maps in ESRI shape format to position a locality in the toolbox. Features in the maps can be points, lines, and polygons (e.g. administrative units, rivers, roads, cities). These features are used as the reference objects to determine the final shapes of the target objects.
2. Select or enter necessary parameters: The toolbox also provides functions to retrieve important parameters automatically from the reference maps, such as the map scale and spatial extent of a RO. However, if a parameter does not exist, users are required to set them during the georeferencing process. A dialogue box has been designed to help users enter the necessary parameters for georeferencing (figure 9). In the dialogue box, users can select an RO based on the reference maps by querying the name or geographic coordinates. In addition, users need to enter the associated spatial relationship and other related parameters to calculate the position of the TO. As mentioned in section 3.1.1, if a linear or areal feature is selected using its name, then the shape method is employed, and the starting-point is assumed to be uniformly distributed within the feature. If the shape of the reference object (e.g. the boundary of a city) is not available, then the extent of the reference object (i.e. the radius) is needed to derive the geographic coordinate of a locality and its uncertainty.
3. Calculate associated uncertainties: After all the parameters have been entered into the program, the probability distribution of the target point will be

**Input the information of the target object**

Step1: Input the reference object:

Place name       Geographic coordinates

Hayfork      Search

Search result:

(Census Name Place-Hayfork)      Highlight it

The coordinates:

Longitude: -123.13545      Latitude: 40.588702       Use actual shape  
 Use point-radius

Datum: \_\_\_\_\_

Extent of RO: 39179.711194      Coordinate precision: \_\_\_\_\_ Meter

Step2: Input the spatial relationship to the object:

In the reference object (eg Merced County, 99 Highway)  
 Distance to the reference object (eg 5 miles to Berkeley)  
 Direction to the reference object (eg North to Berkeley)  
 Internal direction to the areal reference object (eg North of Merced County)  
 Direction and Distance to the reference object (eg 10 miles North of Merced County)

Distance: 20      Precision: 1      Mile      Direction: East

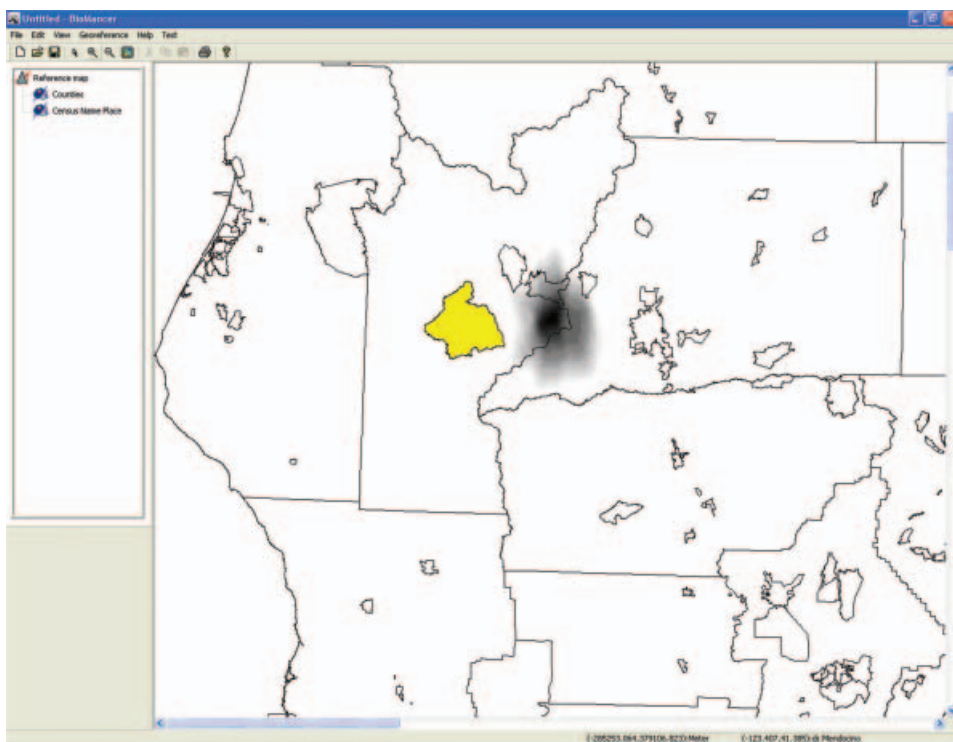
Calculate new...      Add to existing...      Close

Figure 9. Dialogue box for entering information necessary to georeference a locality description.

generated and presented as shown in figure 10(a). Figure 10(b) depicts the uncertainty field and the uncertainty measures based on the description '20 miles E of Hayfork, Trinity County, California'. Users can select from among various units to represent uncertainties.

## 5. Discussion

Two major differences can be identified between the point-radius method in Wieczorek *et al.* (2004) and the probabilistic method proposed in this paper. First,



(a)

Error of target point (Unit: meter)	
Center point x=	-233640.23898943
y=	288197.58253568
Longitude:	-122.75964
Latitude:	40.58099
<input checked="" type="radio"/> Meter <input type="radio"/> Kilometer <input type="radio"/> Mile	
Maximum error:	27294.371
Mean error:	11237.188

(b)

Figure 10. Uncertainty field and the associated measures for the locality '20 miles E of Hayfork, Trinity County, California': (a) probability distribution of the target object; (b) the target object summary showing location, and maximum and mean errors.

the former uses circles to represent both ROs and the maximum uncertainty of TOs. Because the method determines a maximum uncertainty, it often exaggerates the uncertainty, as demonstrated in the case study. Second, the TO is assumed to be uniformly distributed in a circle in the point-radius method, while the probabilistic

method show the differences in likelihood of occurrence across the TO. Table 2 provides a comparison between these two methods based on the common locality types listed in table 1.

As shown in table 2, the probabilistic method will result in smaller uncertainties in most of the cases that have been described for both methods. In addition, the probabilistic method can handle cases such as NF, FS, FH, FO, and BF that were not described by Wieczorek *et al.* (2004) for the point-radius method. In the case of F (i.e. feature), if the feature is a point, the point-radius and probabilistic methods will have the same result; however, if the feature is a region, the latter method will maintain the original shape, while the former method will simplify the shape into a point with a radius. In the real world, no features are points—they all have spatial extent. Therefore, unless the feature is a circle, the probabilistic method will provide a smaller uncertainty estimate than the point-radius method. The point-radius method does have some advantages compared with the probabilistic method. For example, because additional steps are needed to calculate the probabilistic distribution as well as to store the full shapes of both the TOs and ROs in raster format, the point-radius method will outperform the probabilistic method from both computational and storage perspectives. Yet with ever-increasing computational power and inexpensive storage, these concerns about the probabilistic method will become less important.

The importance of spatial uncertainty in GIS analysis and environmental modelling is well recognized, and results reported that do not include the consideration of uncertainty could be of limited use (Fisher 1999). Although the analysis of georeferencing uncertainty in ecological studies is beyond the scope of this study, researchers have demonstrated that documentation of locality uncertainty for museum collection data is important for their studies (Rowe 2005, Waltari *et al.* 2007). There are two major applications of the locality uncertainty. The first is to filter out from analyses those data having large uncertainties and aid in selecting data according to fitness for a specific application. For example, Waltari *et al.* (2007) used the locality uncertainty to select species occurrence data that have geographic uncertainty less than 15 km to reconstruct Pleistocene refugia. The second major use of locality uncertainty is to incorporate the locality uncertainty directly into environmental modelling to understand the sensitiveness of locality uncertainty on model results. For example, Rowe (2005) evaluated the impact of the locality uncertainty on the analysis of patterns of species richness and species range overlap along elevation gradients, and found that failing to assess spatial errors would result in misleading estimates of species richness and community composition. In general, a smaller locality uncertainty will result in smaller uncertainty in modelling results. However, this may not always be true or significant enough to be noticed because the importance of locality uncertainty also depends on specific models in use, characteristics of spatial data, and the questions that need to be addressed. For example, in studying the relationship between species distribution and environment, if the environmental layers are relatively homogeneous, then a slight difference in uncertainty measurements may not be as important as for those highly heterogeneous areas.

Although this study primarily focuses on georeferencing locality description and estimating uncertainty for the museum collection data, the proposed probabilistic method is also suitable to other geographic applications. One example could be geographic information retrieval (GIR) techniques, which parse a variety of textual

Table 2. Comparisons between the point-radius method (PRM) and the probabilistic method (PM) for different locality types.

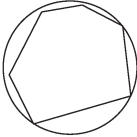
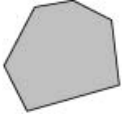
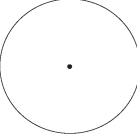

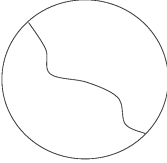



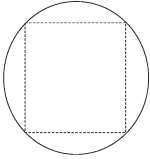

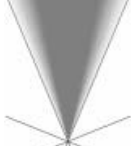
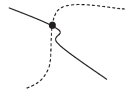
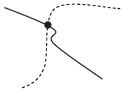
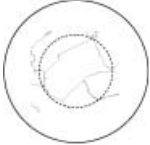


Locality type	Description	Comparison	Demonstration figures	
			Point-radius method	Probabilistic method
F	Feature	Same if the RO is a point, or better result from PM if the RO is a region since the shape is maintained		
FOH	Offset from a feature (or a path) at a heading	Result is better constrained by PM		
P	Path or linear feature	Similar to areal cases of 'F' locality type; result is usually much better constrained by PM		
NF	Near a feature or path	Original PRM does not handle this case	NA	
FS	Subdivision of a feature or a path	Original PRM does not handle this case	NA	

Table 2. (Continued.)

Locality type	Description	Comparison	Demonstration figures	
			Point-radius method	Probabilistic method
FOO	Orthogonal offsets from a feature	Result is better constrained by PM		
FH	Heading from a feature, no offset	Original PRM does not handle this case	NA	
J	Junction	Same result		
FO	Offset from a feature, no heading	Result is usually much better constrained by PM		
BF	Between features or paths	Original PRM does not handle this case	NA	



information (e.g. Web documents) to retrieve references to locations and assign geographic coordinates to them (Jones *et al.* 2003). Although these coordinates provide the basis for search and retrieval engines, published research does not use estimates of the spatial uncertainty of the textual locality information. Uncertainty sources described in this paper may not be the same as those faced by GIR studies; however, the probabilistic method based on uncertainty source distributions to describe both ROs and TOs could provide useful guidance for GIR techniques in estimating locality uncertainty from textual documents.

## 6. Conclusions

There is increasing demand for techniques to integrate spatial information from sources that are traditionally qualitative in nature, especially among natural history collections, which have a legacy of written descriptions of habitat, environment, observations, occurrences, and collections. Also, increasingly, the value of data quality documentation in the form of measures of data uncertainty is being recognized. Wieczorek *et al.* (2004) developed a point-radius method to estimate the maximum uncertainty associated with a locality description by summing the maximum errors from all uncertainty sources. The point-radius method provides a relatively easy and practical solution for georeferencing localities and estimating uncertainties. However, this method tends to overestimate the uncertainty, since it is essentially additive and does not consider the probability distribution for each uncertainty source. In this study, we introduced the uncertainty field, a two-dimensional PDF, to represent both the reference and target objects. We then created the probability distributions for uncertainty sources normally encountered during the georeferencing process. The spatial distribution of the target object can be computed and visualized by discretizing the uncertainty field into a raster form. Using the ‘distance at a heading’ as a case study, the results indicated that the proposed method provides a much more constrained, and hopefully, therefore, a more realistic and useful uncertainty estimate than the point-radius method.

## Acknowledgements

We thank Dr. Goodchild, Dr. Hill, and the reviewers for their constructive comments that helped strengthen the paper. This research is partially supported by the BioGeomancer Project funded by the Gordon and Betty Moore Foundation.

## References

- BONNER, M.R., HAN, D., NIE, J., ROGERSON, P., VENA, J.E. and FREUDENHEIM, A.L., 2003, Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, **14**, pp. 408–412.
- BURROUGH, P.A. and MCDONNELL, R.A., 1998, *Principles of Geographical Information Systems* (Oxford: Oxford University Press).
- CLEMENTINI, E., FELICE, P. and HERNÁNDEZ, D., 1997, Qualitative representation of positional information. *Artificial Intelligence*, **95**, pp. 317–356.
- COUCLELIS, H., 1996, Towards an operational typology of geographic entities with ill-defined boundaries. In *Geographic Objects with Indeterminate Boundaries*, P.A. Burrough and A.U. Frank (Eds), pp. 45–55 (London: Taylor & Francis, 1996).
- DEHAK, S.M.R., BLOCH, I. and MAÎTRE, H., 2005, Spatial reasoning with incomplete information on relative positioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, pp. 1473–1484.

- DU, S., WANG, Q. and YANG, Y., 2004, Fuzzy description of fuzzy direction relations and their similarities. In *Proceedings of 12th International Conference on Geoinformatics—Geospatial Information Research: Bridging the Pacific and Atlantic*, pp. 496–502.
- DUCKWORTH, W.D., GENOWAYS, H.H. and ROSE, C.L., 1993, *Preserving Natural Science Collections: Chronicle of Our Environmental Heritage* (Washington, DC: National Institute for the Conservation of Cultural Property).
- DUTTA, S., 1991, Approximate spatial reasoning: Integrating qualitative and quantitative constraints. *International Journal of Approximate Reasoning*, **5**, pp. 307–330.
- EGENHOFER, M.J. and HERRING, J., 1991, Categorizing binary topological relations between regions, lines and points in geographic databases. In *Technical Report, Department of Surveying Engineering, University of Maine, Orono, ME*.
- EGENHOFER, M.J. and SHARIFF, A.R., 1998, Metric details for natural-language spatial relations. *ACM Transactions on Information Systems*, **16**, pp. 295–321.
- FISHER, P.F., 1999, Models of uncertainty in spatial data. In *Geographical Information Systems*, P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Eds) (New York: Wiley), pp. 191–205.
- FRANK, A.U., 1991, Qualitative spatial reasoning about cardinal directions. In *Proceedings of the 7th Austrian Conference on Artificial Intelligence*, D. Mark and D. White (Eds) (Baltimore, MD: Morgan Kaufmann), pp. 157–167.
- FREKSA, C., 1992, Using orientation information for qualitative spatial reasoning. In *Proceedings of the International Conference GIS-From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Lecture Notes in Computer Science*, vol. 639 (Berlin: Springer), pp. 162–178.
- GOODCHILD, M.F., 1989, Modeling error in objects and field. In *Accuracy of Spatial Databases*, M.F. Goodchild and S. Gopal (Eds), pp. 107–114 (London: Taylor & Francis).
- GOODCHILD, M.F., 1992, Geographical data modeling. *Computers & Geosciences*, **18**, pp. 401–408.
- GOODCHILD, M.F., 2001, A geographer looks at spatial information theory. In *Proceedings of COSIT 2001, Lecture Notes in Computer Science*, vol. 2205 (Berlin: Springer), pp. 1–13.
- GOODCHILD, M.F. and HUNTER, G.J., 1997, A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, **11**, pp. 299–306.
- GOYAL, R. and EGENHOFER, M.J., 2001, Similarity of cardinal directions. In *Advances in Spatial and Temporal Databases*, C.S. Jensen, M. Schneider, B. Seeger and V.J. Tsotras (Eds), *Lecture Notes in Computer Science*, vol. 2121 (Berlin: Springer), pp. 36–55.
- HILL, L.L., 2006, *Georeferencing—The Geographic Associations of Information* (Cambridge, MA: MIT Press).
- JONES, C.B., ABDELMOTY, A.I. and FU, G., 2003, Maintaining ontologies for geographical information retrieval on the web. In *Proceedings of ODBASE'03*, LNCS 2888, pp. 934–951.
- LEUNG, Y. and YAN, J.P., 1998, A locational error model for spatial features. *International Journal of Geographical Information Science*, **12**, pp. 607–620.
- LONGLEY, P.A., GOODCHILD, M.F., MAGUIRE, D.J. and RHIND, D.W., 2005, *Geographic Information Systems and Science, Second Edition* (New York: Wiley).
- MANIS (Mammal Networked Information System), 2001, rev. 16 Apr 2007. Available online at: <http://manisnet.org> (accessed 16 November 2007).
- MARK, D.M. and EGENHOFER, M.J., 1994, Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. *Cartography and Geographical Information Systems*, **21**, pp. 195–212.
- MONTELLO, D.R. and FRANK, A.U., 1996, *Modeling Directional Knowledge and Reasoning in Environmental Space: Testing Qualitative Metrics. The Construction of Cognitive Maps*, J. Portugali (Eds), pp. 321–344 (Dordrecht, Netherlands: Kluwer Academic).

- NATIONAL GEODETIC SURVEY, 1992, NADCON—Version 2.1, [http://www.ngs.noaa.gov/PC\\_PROD/NADCON/](http://www.ngs.noaa.gov/PC_PROD/NADCON/) (accessed 16 March 2008).
- RANDELL, D.A., CUI, Z. and COHN, A.G., 1992, A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference (KR'92)*, pp. 165–176.
- ROWE, R.J., 2005, Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. *Journal of Biogeography*, **32**, pp. 1883–1897.
- SHI, W., 1998, A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographic Information Science*, **12**, pp. 131–143.
- SKIADOPOULOS, S., GIANNOUKOS, C., VASSILIADIS, P., SELLIS, T. and KOUBARAKIS, M., 2004, Computing and handling cardinal direction information. In *Proceedings of EDBT 2004*, E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Bohm and E. Ferrari (Eds), *Lecture Notes in Computer Science*, vol. 2992 (Berlin: Springer), pp. 329–347.
- THAPA, K. and BOSSLER, J., 1992, Accuracy of spatial data used in geographic information-systems. *Photogrammetric Engineering and Remote Sensing*, **58**, pp. 835–841.
- TONG, X., SHI, W. and LIU, D., 2003, An error model of circular curve features in GIS. In *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, pp. 141–146.
- TØSSEBRO, E. and NYGÅRD, M., 2002, An advanced discrete model for uncertain spatial data. In *Proceedings of WAIM 2002, Lecture Notes in Computer Science*, vol. 2419 (Berlin: Springer), pp. 37–51.
- USGS (United States Geological Survey), 1947, *United States National Map Accuracy Standards*, <http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/nmas/NMAS647.PDF> (accessed 2 March 2008).
- VAN NIEL, T.G. and McVICAR, T.R., 2002, Experimental evaluation of positional accuracy estimates from a linear network using point- and line-based testing methods. *International Journal of Geographical Information Science*, **16**, pp. 455–473.
- VEREGIN, H., 2000, Quantifying positional error induced by line simplification. *International Journal of Geographical Information Science*, **14**, pp. 113–130.
- WALTARI, E., HIJMANS, R.J., PETERSON, A.T., ÁRPÁD S. NYÁRI, PERKINS S.L. and ROBERT, P.G., 2007, Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE*, **2**: e563. doi:10.1371/journal.pone.0000563.
- WELCH, R. and HOMSEY, A., 1997, Datum shifts for UTM coordinates. *Photogrammetric Engineering and Remote Sensing*, **63**, pp. 371–375.
- WIECZOREK, J., GUO, Q. and HIJMANS, R.J., 2004, The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, **18**, pp. 745–767.
- WORBOYS, M.F. and CLEMENTINI, E., 2001, Integration of imperfect spatial information. *Journal of Visual Languages and Computing*, **12**, pp. 61–80.