Toward Quantitative Geocode Accuracy Metrics

Daniel W. Goldberg[1], John P. Wilson[1, 2], Myles G. Cockburn[3]
Departments of Computer Science[1], Geography[2], and Preventive Medicine[3]
University of Southern California
Los Angeles, CA USA
{dwgoldbe, jpwilson, mylesc}@usc.edu

***Abstract*—Existing geocode quality metrics provide little utility for those interested in the spatial uncertainty associated with a geocoded location. The per-geocode metrics describe aspatial characteristics of individual aspects of the geocoding process, while the per-dataset spatial metrics provide only general information that may not apply to a single geocode of interest. In this paper we develop a method for describing the certainty of a geocoded datum as a spatial probability surface based on an uncertainty propagation model which takes into account the certainty stemming from each portion of the geocoding process. This surface-based geocode output structure provides a more truthful view of the uncertainty present in these data and will enable more realistic estimates of information derived from them in such tasks as environmental exposure modeling.**

Keywords: geocode; uncertainty, spatial probabilty distributions

I. INTRODUCTION

It is well known that spatial data derived through any process will involve uncertainty (Veregin 1995; Heuvelink 2002; Fisher and Tate 2006). Research has shown that the process of geocoding, converting aspatial textual information such as a postal address into a spatial location, is especially subject to uncertainty because error can be introduced at each of the many components of the process (Bichler and Balchak 2007; Zandbergen 2009). This situation is problematic because geocoding is a fundamental geospatial operation essential to many diverse fields such as epidemiology, marketing, and planning. Spatial data derived from geocoding typically form the underlying data from which geographic mapping and visualization can occur and spatially-based research questions can be posed and investigated (Zandbergen 2009). Although varied and diverse in terms of their applications and usages, this wide set of geocode users all require spatially accurate geocoded results as well as metrics capable of describing the accuracy. However, there remains a lack of research and technology capable of clearly describing, calculating, and/or predicting the spatial accuracy and uncertainty associated with a geocode result. In this article, we make two contributions: (1) we propose a flexible model of the geocoding process that enables the modeling, assessment, and propagation of uncertainty across the components in different geocoding approaches; and (2) we develop a novel geocoding output structure as an uncertainty surface which more truthfully describes the certainty of a geocode.

This work was supported by a supplement to contract N01-PC-35139 with the US National Cancer Institute, by grant number U55/CCU921930-02 from the US Centers for Disease Control and Prevention, and NIEHS grants P30 ES07048 and R01 ES015552.

II. CURRENT GEOCODING QUALITY METRICS

The current metrics used to describe the quality of geocoded data are most often qualitative classifications – in the best case describing characteristics and result codes from one or more components of the geocoding process. Specifically, the metrics used to express the accuracy / certainty of geocodes are: (1) the *match-rate* – the number of input addresses that a geocoding system was able to match; (2) the *match type* – the level of geographic object matched to e.g., parcel centroid, postal code, street address; and (3) the *match certainty* – a value describing the level of similarity and/or likelihood of a match between the input address and the address associated with the matched feature input derived either probabilistically or deterministically. In addition, a fourth error metric, *spatial accuracy*, is often determined after the fact as average values of distance and direction "from truth" by comparing computed output locations and known locations for a subset of the data (Bichler and Balchak 2007; Zandbergen 2009).

Although these metrics represent the status quo, scientists and other users should use caution when utilizing them to determine fitness-for-use for a geocoded dataset and a particular study because, fundamentally, they do not describe a true geographic area for a geocode, nor any form of confidence interval, spatial or otherwise. Of these, the *match type* is particularly troubling because it is typically the primary metric used to determine quality, yet it is a crude aspatial metric which assumes global relationships of relative accuracies between reference data layer types and implies spatial accuracies with resolutions that are in fact nonstationary and in many cases contradictory. The *match rate* provides quantitative information about the precision and recall of a geocoding system as a whole, but is aspatial and thus falls short in quantitatively describing the potential spatial error and/or uncertainty. *Match certainty* is a true quantitative value, but is again aspatial and thus provides no insight for those interested in spatial certainty. *Spatial accuracy* values computed for a set of geocoded data do provide some confidence in the overall character of a dataset, but are of limited utility when one seeks to quantify the error of any particular output. Furthermore, this metric often describes error in a radial pattern (Wieczorek *et al.* 2004), often as "the true location is within a 100 m buffer output geocode", even though the actual underlying geographic shapes of the matched reference features clearly preclude such descriptions as in the case of linear street segment interpolation where the predominant uncertainty in the output is along the axis of the line, not in a radius extending outward from it, (notwithstanding the fact that centerline dropbacks do account for some small portion of uncertainty) (Zandbergen 2009). Finally, not one of the current metrics available for describing geocode accuracy considers the temporal aspects of either the input data or the underlying geographic reference data from which an output is ultimately computed, as would be normal practice in uncertainty modeling (Gahegan and Ehlers 2000).

The reasons for these shortcomings are fundamentally twofold. First, the geocoding process has traditionally been cast as a subclass of the record linkage problem: given an input and a set of reference features, determine the most likely match between the two (Bichler and Balchak 2007). In this context, the geographic component of spatial error and uncertainty have been considered as an afterthought, as evidenced by the qualitative *match type* being the primary metric and the computation of spatial accuracy and uncertainty as a post processing step (Zandbergen 2009). Second, current geocoding metrics focus on describing the quality of a single output location resulting from the geocoding process (Zandbergen 2009). This practice of defining the structure of geocoding output as a single "best" location implies absolute certainty in the outcome and ignores all alternative outcomes that may have had nearly the same likelihood of being correct.

III. Representations and Sources of Geocode Uncertainty

The appropriate representation and computation of error and uncertainty propagation in geospatial models and operations is critical to their successful utilization of research (Veregin 1995). The literature is rich with descriptions of geoprocesses as complex models with interactions between numerous components, all with their own error and uncertainty that must be accounted for in concert (Heuvelink 2002; Smith and Fuller 2002; Fisher and Tate 2006), which combined together complicate a simple computation of a single uncertainty value (Heuvelink 2002). No exception to this rule, the sources and magnitudes of spatial error and uncertainty in the geocoding process have been well documented on numerous occasions (Bichler and Balchak 2007; Zandbergen 2009), which has, to date, hindered the creation of a single unified and consistent uncertainty metric for geocoded output. Each component of the process is subject to uncertainty including: (1) the input data which is the text describing a location; (2) the address parsing and normalization algorithms which identify the pieces of the input text and transform them to standard values; (3) the feature matching algorithms which identify candidate matching geographic features in the reference data sources; and (4) the feature interpolation algorithms (Zandbergen 2009). Each of these components transforms some portion of the data within the system and thus needs to be considered and modeled in an error propagation framework to ensure the final output fully represents what happened within the system (Veregin 1995; Gahegan and Ehlers 2000).

Existing research has made strides toward this goal, most notably the Geocoding Certainty Indicator (GCI) which attempts to quantify an overall propagated uncertainty value derived from the level of information inherent in the type of input data, the match probability, and the interpolation uncertainty (Davis Jr. and Fonseca 2007). This is an important first step, but falls short of truly describing spatial certainty of geocoded information because: (1) it utilizes a limited set of parameters in its model of uncertainty in the geocoding process; (2) does not completely capture the different meanings of uncertainty within each geocoding process component and across their various data models; and (3) focuses on describing a single output. With regard to the first two, what is missing is a generalizable way to describe the data transformations and uncertainty flows through the steps of the geocoding process. With regard to the third, an uncertainty computation for an output geocode should take into account the multiple candidate output locations and their respective uncertainty values.

IV. Unceratainty Modeling in the Geocoding Process

The framework proposed by Gahegan and Ehlers (2000) enables us to define the geocoding process in a manner which captures, represents, and propagates uncertainty across the different data models inherent to the set of chained data transformations which are the components of the geocoding system. This framework provides the tools necessary for achieving the first steps toward defining an error budget with regard to a single geographic datum, datasets, and data transformations. In addition, the flexibility of this framework allows the inclusion of different geocoding techniques and error analysis procedures, such as different feature interpolation techniques and the model proposed in the GCI. Complete details of this framework are provided in Gahegan and Ehlers (2000). In brief, each data model in the geocoding process is a cast as a transformation ψ from an input dataset $A(\,)$ to an output dataset $A^{'}(\,)$, which optionally uses some additional data Q. Datasets are described by a series of values where D is the value, S is the spatial extent, T is the temporal extent, each of which has its own uncertainty, α, β, and χ, respectively. The context in which the data are relevant is included (C_x), as are values for consistency and completeness, δ and ε, respectively. Applying this framework to the geocoding process defines model transformations from raw textual input data (C_R) to an address data model (C_A) and a geographic object data model (C_G). We additionally include the transformation to a surface model (C_S) which is our new structure for representing the geocode output.

$$A(C_R) \rightarrow A^{'}(C_A) \rightarrow A^{''}(C_G) \rightarrow A^{'''}(C_S) \tag{1}$$

Space limitations prevent a complete derivation of the transformations inherent to all of these components. This would be of limited use in any case because each implementation of the geocoding process will include different parameters and quantifications of uncertainty. As one example of how a process is mapped to this framework, consider the transformation from the raw input data (C_R) to the address data model (C_A). Here, uncertainty may arise because the input datum may be incomplete, inaccurate, outdated, and/or contain terms that are ambiguous. Machine learning approaches to this problem require the inclusion of $Q_1 \dots Q_k$ classifiers as input, each with its own spatial and temporal extent used for training.

$$A(D,S,T,\alpha,\beta,\chi,\delta,\varepsilon):(C_R) \xrightarrow{\psi_{Parsing}}$$

$$\left.\begin{matrix} Q_1(D_1,S_1,T_1,\alpha_1,\beta_1,\chi_1,\delta_1,\varepsilon_1) \\ Q_2(D_2,S_2,T_2,\alpha_2,\beta_2,\chi_2,\delta_2,\varepsilon_2) \dots \\ Q_k(D_k,S_k,T_k,\alpha_k,\beta_k,\chi_k,\delta_k,\varepsilon_k) \end{matrix}\right\}$$

$$A'(D',S',T',\alpha',\beta',\chi',\delta,\varepsilon) \tag{2}$$

Given the ample derivation of uncertainty sources in other components available in the literature (Bichler and Balchak 2007; Zandbergen 2009), our generalized model provides a methodology for uncertainty representation and propagation for the geocoding process as a whole regardless of the specific geocoding implementation strategy utilized.

V. Geocode Structure as a Certatinty Surface

Using the above approach we can redefine the structure of geocode output to describe the likelihood that the true location resides at all possible output locations across a region by representing geocode output as a certainty surface. Doing so moves us away from the notion of geocode quality as local uncertainty (describing a single point) and toward the multi-point, or spatial uncertainty, which describes the uncertainty associated with multiple locations (Goovaerts 2001). This structure is more representative of what the geocoding process actually tells us – that there is a body of evidence supporting and/or refuting the proposition that the true geocode exists at any particular location.

To accomplish this, we overlay a grid G to rasterize and discretize the region containing the output so we can assign a certainty value $u(c_{ij})$ to each cell $c_{ij} \in G$ within the region, as a fuzzy membership value in a single class $incorrectLocation \in [0..1]$ representing the likelihood that any cell contains the correct location. Selecting the appropriate grid resolution is of course an issue (Heuvelink 2002; Fisher and Tate 2006), but will not be discussed here as our immediate goal is to simply develop a representative model. The certainty value for each c_{ij} is given by two quantities.

$$u(c_{ij}) = 1 - (\lambda + \gamma) \tag{4}$$

The first, $\lambda = Area(c_{ij})/Area(r_k)$, is the base certainty implied by c_{ij} being covered by the geographic reference feature because the true output is assumed to be somewhere within, along, or at the feature. This value is uniformly distributed to each $c_{ij} \in r_k$. The second, γ, is the output certainty resulting from the process that generated the candidate. For example, this includes the certainty of the interpolation function $\gamma = f(a, c_{ij})$ at each c_{ij} for a given input datum a. In weighted areal interpolation using a population density surface, the uncertainty would be the population value at each c_{ij}. In address range interpolation, uncertainty would be calculated using a rasterized version of the street segment (Rueda *et al.* 2004) (Figure 1).

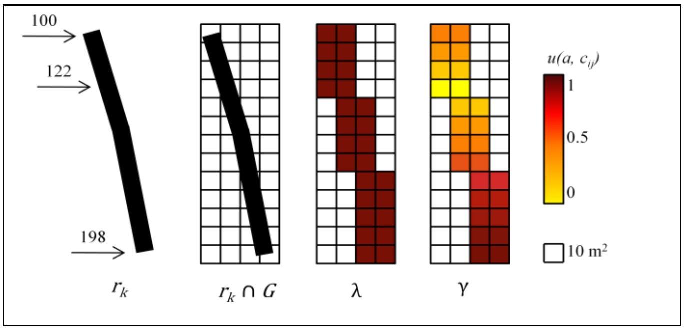

Figure 1. Simplified example of rasterization and interpolation spatial certainty computation for address range interpolation with $n_A = 122$ along street segment with $n_F = 100$ and $n_T = 198$

Here, the address range associated with the street is proportionally applied to each c_{ij} along the segment to produce the set of address numbers associated with each cell, $n_{c_{ij}}$. In this case, $u(c_{ij})$ is expressed as a linear function which decreases outward from the location containing the input address (n_A) and the distance from the address range of the street in terms of its from and to addresses, n_F and n_T, respectively, creating epsilon error bands for each address interval from the rough approximation of the shape (Zhang and Goodchild 2002).

$$\gamma = Abs\left(\frac{n_{c_{ij}} - n_A}{n_T - n_F}\right) \tag{5}$$

Other interpolation techniques will need to be modeled in a similar manner before they can be incorporated into our error propagation framework and applied to produce a certainty surface describing the likelihood of correctness at each location. Fundamentally however, all interpolation techniques utilize a geographic feature as a basis and therefore can be handled in this way.

VI. Composite Certainty for Multiple Candidates

As shown in Figure 2, the geocoding process is actually a decision tree with multiple potential outcomes at each level (transformation) that guide the set of choices available to all subsequent levels. Each complete path in this tree ultimately leads to a geographic reference feature and interpolated output with an associated degree of belief that has been propagated by the framework presented earlier. Typical geocoding systems toss aside the uncertainty information inherent in these alternative outcomes, providing the user an incomplete picture of the uncertainty in their data. These alternative outcomes should be presented to the user if they are to understand their data (Mowrer 2000). To do so we first note that each of the $R_1 \dots R_n$ reference data layers can provide one or more geographic reference data features ($r_{n1} \dots r_{nk}$) it believes matches the input. In the US this would include such objects as street centerlines, postal regions, cities, counties, and states. In our approach, every possible candidate output is fully realized for each input datum. The spatial uncertainty of each of these is modeled as a certainty surface as described in the last section using a grid of uniform resolution.

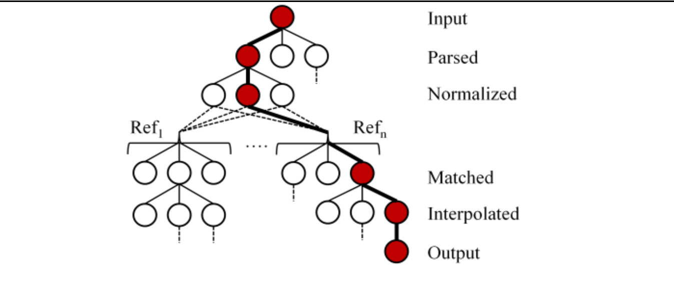

Figure 2. The alternative paths in geocode production

This results in a comparable set of surfaces describing the degree of belief that the true location occurs in a number of locations $s_m \in S, m = 0 \ldots n$. Together, this set provides the full evidence for all locations at which the system has reason to believe the true output resides.Using Dempster's rule of combination we can combine the $u(c_{ij})$ associated with each c_{ij} across each $s_m \in S$. This is evidence-based approach determines a multi-source composite certainty value (Zhang and Goodchild 2002; Bi et al. 2007). The combined certainty surface output spans the full region of all reference features used in producing all candidate geocodes, and will be representative of the certainty inherent in each as propagated through the framework. Because the resulting surface is based on the independent probability of each, we must apply a normalization procedure to proportionally scale these independent certainties to account for the overall certainty of the combined surface. One or more matched features at the highest administrative level will implicitly define the applicable region, e.g., in the US the state or states the output is believe to be within. From this we obtain the maximum spatial extent (Z) which can be used to compute the normalized values for each cell on the combined output surface, $u'(c_{ij})$.

$$u'(c_{ij}) = \frac{u(c_{ij})}{\sum u(c_{ij})}, \forall c_{ij} \in Z \tag{4}$$

VII. Implications for Geocode Consumers

Reporting the results of the geocoding process as spatial probability distributions will change the way that geocoded information is utilized in scientific studies and applications. The representation structure that we propose herein will enable consumers to model the interactions between spatial processes and geocoded locations with a new and novel spatial uncertainty value. For example, epidemiological investigations linking environmental exposures and health outcomes will be able to associate confidence intervals with exposure estimates derived from the intersection of chemical dispersion surfaces and the potential location of an individual. Such a certainty surface can be directly incorporated into epidemiological spatial modeling and will enable exposure assessments to be normalized to account for the probability of a geocode being in the correct location. In contrast to simply classifying a subject as un/exposed based on the point representation of their location, scientists will be able to assign classification probabilities. Further, quantitative estimates of ambient environmental exposure can be assessed with regard to confidence intervals around the position of the individual.

VIII. Future Work

We have presented a method for combining what is known about a geocode to produce a certainty surface describing the likelihood that a particular output should be placed at a particular location guided by insight about the characteristics of each component of the geocoding process. This method assumes that several quantities are available to describe the spatial-temporal aspects of accuracy and uncertainty for each component which is in many cases still not well defined or understood because of the complex assumptions (parameters) which are incorporated and/or relevant. We plan to investigate methods for identifying unknown model parameters and values such as those employed in complex modeling with high degrees of model and parameter uncertainty (Wiegand *et al.* 2004).

References

Bi, Y., D. Bell, H. Wang, G. Guo, and J. Guan. (2007). Combining multiple classifiers using Dempster's rule of combination for text categorization. Applied Artificial Intelligence. 21 (3), 211-239.

Bichler, G., and S. Balchak. (2007). Address matching bias: Ignorance is not bliss. PIJPSM. 30 (1), 32-60.

Davis Jr., C. A., and F. T. Fonseca. (2007). Assessing the certainty of locations produced by an address geocoding system. GeoInformatica. 11 (1), 103-129.

Fisher, P. F., and N. J. Tate. (2006). Causes and consequences of error in digital elevation models. Progress in Physical Geography. 30 (4), 467-489.

Gahegan, M., and M. Ehlers. (2000). A framework for the modelling of uncertainty between remote sensing and geographic information systems. ISPRS Journal of Photogrammetry and Remote Sensing. 55 (3), 176-188.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. Geoderma. 103, 3-26.

Heuvelink, G. B. M. (2002). Analysing uncertainty propagation in GIS:Why is it not that simple? In G. M. Foody and P. M. Atkinson (Eds) Uncertainty in Remote Sensing and GIS (pp. 156-165). West Sussex, UK: John Wiley & Sons.

Mowrer, H. T. (2000). Uncertainty in natural resource decision support systems: sources, interpretation, and importance. Computers and electronics in agriculture. 27, 139-154.

Rueda, A. J., R. J. Segura, F. R. Feito, and J. R. de Miras. (2004). Rasterizing complex polygons without tessellations. Graphical Models. 66 (3), 127-132.

Smith, G. M., and R. M. Fuller. (2002). Land Cover Map 2000 and meta-data at the land parcel level. In G. M. Foody and P. M. Atkinson (Eds) Uncertainty in Remote Sensing and GIS (pp. 143-153). West Sussex, UK: John Wiley & Sons.

Veregin, H. (1995). Developing and testing of an error propagation model for GIS overlay operations. International Journal of Geographical Information Science. 9 (6), 595-619.

Wieczorek, J., Q. Guo, and R. J. Hijmans. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science. 18 (8), 745-767.

Wiegand, T., E. Revilla, and F. Knauer. (2004). Dealing with uncertainty in spatially explicit population models. Biodiversity and Conservation. 13, 53-78.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. Geography Compass. 3 (2), 647-680.

Zhang, J., and M. F. Goodchild. (2002). Uncertainty in Geographic Informatoion. London: Taylor & Francis.