

4 Modeling Localities with Fuzzy Sets and GIS

Sungsoon Hwang and Jean-Claude Thill

University at Buffalo – The State University of New York at Buffalo
Department of Geography and National Center for Geographic Information and
Analysis (NCGIA)
105 Wilkeson Quad
Buffalo, NY 14261
{shwang5,jcthill}@buffalo.edu

4.1 Introduction

It is only recently that the fuzzy-set theoretic approach to spatial objects and their concepts has joined the mainstream of geographic information system (GIS) and science (GIScience). Several reasons account for this. Recent research on spatial objects has revealed that spatial vagueness is inherent in some geographic features (Burrough and Frank 1996). For instance, the boundary between a mountain and a valley is not sharply defined. Furthermore, even if a geographic phenomenon is best described as crisp, humans tend not to reason in a precise manner, but rather in an approximate manner (e.g. they live *near* Chicago). Moreover, perception and cognition vary widely between individuals. Furthermore, information can be incomplete or imprecise due to rough measurements or to our incomplete ability to grasp the scope and detail of spatial objects. In other words, there always exists a gap between the reality and its representation.

We use fuzzy set theory (Zadeh 1965) as a mean to reconcile discrepancies existing between reality and its representation. We discern three representational different levels at which fuzzy set concepts can be applied, namely the ontology, perception, and implementation levels. The (spatial) ontology level pertains to generic concepts inherent in spatial objects. The perception level concerns the mental models used to perceive the environment. The implementation level encompasses the errors that have propagated during system implementation. The combination of spatial vagueness, diverse human perceptions, and implementation errors account for the gap existing between reality and its representation. In general, fuzzy set concepts preserve details (Robinson 2002) whereas traditional (crisp) GIS data models overlook the loss of information by forcing reality into a coarse (in the sense of low resolution) representation. Fuzzy set theory can overcome the gap by providing mechanisms for ontologically and cognitively plausible (Worboys 2001) and error-sensitive (Duckham et al. 2001) representation of the reality. In sum, fuzzy set theory provides a means to address various kinds of uncertainty such as spatial vagueness, human perception, and imperfect information.

This study is part of a larger project aimed at geographically referencing the fatal accident data. Our task is to pinpoint the location where a traffic crash is most

likely to have occurred given the limited and imprecise information available on this crash. In our study, georeferencing can be roughly defined as the conversion of the linguistic description of a location to a quantitative specification. As Goodchild (2000) pointed out, effective georeferencing can be a matter of life and death in the case of communication between a caller and an emergency dispatcher. The linguistic description of location is sometimes not clear-cut, not only because many alternate names are used to refer to the same location, but also because the location itself is not well defined. We focus on the problem of determining the location of a certain locality.

We hypothesize that location indeterminacy of localities is caused by spatial vagueness, interpersonal differences in perception, and imperfect information. We compute the value that quantifies location indeterminacy by modeling the indeterminate part of localities by a fuzzy set membership function. We examine the relationship between the value of location indeterminacy and attributes of localities in order to test the stated hypotheses.

The purpose of this research is to show how fuzzy set theory can be properly applied in modeling localities. Also the result will assure whether there exists fuzziness in determining the location of locality. This study develops a fuzzy set membership function for indeterminate boundaries of localities. By testing our hypotheses on the relationship between location determinacy and characteristics of locality, we examine whether fuzzy set theories can capture various kinds of uncertainty at the ontology, perception, and implementation levels.

Modeling localities by fuzzy sets has a definite advantage over a crisp set in that it makes best possible use of sparse information to reconstitute detail. More specifically, fuzzy-set-based localities constitute a closer depiction of reality, such as overlapping memberships of localities. Next, fuzzy set provides a conservative representation tool for individual differences in the perception. Finally, allowing the soft processing (fuzzy set modeling) over the hard data (reference data) can minimize the problems caused by the imperfection of source data.

The remainder of this chapter is organized as follows. In Section 4.2, the specific georeferencing motivating this study is described. We formulate research hypotheses and specify the assumptions on which this study is based. In Section 4.3, we give a brief overview of related research, such as the ontology of spatial objects, the representation of fuzzy regions, and the notion of nearness. In Section 4.4, we define the fuzzy set membership function of localities. The implementation steps in GIS are described in Section 4.5. The analyses of results are given in Section 4.6. We examine if fuzziness is substantial in identifying localities. By looking at the cases that are georeferenced by a fuzzy set modeling, we may or may not find evidence of fuzziness in locality. The hypothesis is examined also. Finally, Section 4.7 concludes this study.

4.2 Locality and the Georeferencing Problem

Figure 4.1 presents the steps involved in the process of georeferencing traffic accidents: from the reporting of accidents, to their storage in the databases, and finally the georeferencing using a geographic information system. Police officers use the coding forms shown in Figure 4.2 to make a permanent record of accident information. The number in the upper right-hand corner of each section of the coding form indicates the column identifier for this information item in the flat files to which the paper-based record is transferred. In Figure 4.2, the locality information of the accident is recorded in the upper section - State (1-2), City (14-17), and County (18-20) - using GSA Geographic Locator Codes (GLC).

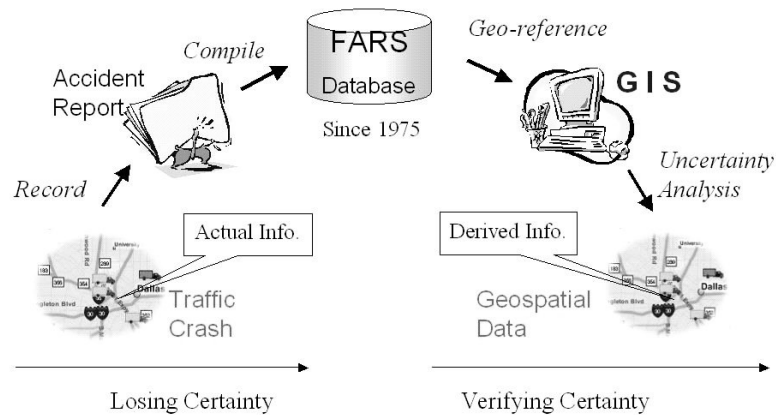


Fig. 4.1. Georeferencing Traffic Accident Data under Uncertainty

As part of the georeferencing process, attention needs to be paid to the manner in which police officers identify the localities. Do they record locality information under the assumption that locality boundaries are determinate? That is, is the location of localities taken as determinate? The answer may be compound. It may depend on the environment (what), on the agent (who), and on the medium (through what).

Environment: It may be much easier to delineate the boundaries if a distinguishable geographic feature, such as a body of water, surrounds a locality (e.g., peninsula, island). Moreover, locality is the result of human conceptualization and demarcation rather than a physical demarcation on the surface of the Earth. Making this kind of spatial concept crisp is part of the repetitive process of conceptualization. It is likely that a person finds it easier to identify the location of a conspicu-

ous locality than that of a less conspicuous locality. Therefore, it can be argued that certain characteristics of localities affect the location indeterminacy, or fuzziness in delineating their boundaries.

100-A EXHIBIT FARS CODING FORMS (ACCIDENT LEVEL)

1998 Fatal Accident Reporting System

ACCIDENT LEVEL

U.S. Department of Transportation
National Highway Traffic Safety Administration

O.M.B. No. 2127-0006
CODED BY: _____
DATE CODED: _____
STATE CASE NO. _____

CASE NUMBER		1	2	CONSECUTIVE NUMBER		3	4	5	6	CITY		14	15	16	17	COUNTY		18	19	20					
STATE (GSA CODES)																									
DATE Month, Day and Year		21	22	23	24	25	26	TIME Military Time 9999—Unknown		29	30	31	32	Number of Vehicle Forms Submitted		33	34	Number of Person Forms Submitted		35	36	Number of Non-Motorist Forms Submitted		37	38
NATIONAL HIGHWAY SYSTEM										39	SPEED LIMIT										78	79			
0—This Section IS NOT on the NHS 1—This Section IS on the NHS 8—Unknown if this Section is on the NHS											Actual Miles Per Hour Except: 00—No Statutory Limit 99—Unknown														
ROADWAY FUNCTION CLASS										40	41	ROADWAY ALIGNMENT										80			
RURAL 01—Principal Arterial-Interstate 02—Principal Arterial-Other 03—Minor Arterial 04—Major Collector 05—Minor Collector 06—Local Road or Street 09—Unknown Rural												1—Straight 2—Curve 9—Unknown													
URBAN 11—Principal Arterial-Interstate 12—Principal Arterial-Other 13—Other Principal Arterial 14—Minor Arterial 15—Collector 16—Local Road or Street 19—Unknown Urban												ROADWAY PROFILE										81			
												1—Level 2—Grade 3—Hillcrest 4—Sag 9—Unknown													
												ROAD SURFACE TYPE										82			
												1—Concrete 2—Blacktop (Bituminous) 3—Brick or Block 4—Slag, Gravel or Stone 5—Dirt 8—Other 9—Unknown													
ROUTE SIGNING										42		ROAD SURFACE CONDITIONS										83			
1—Interstate 2—U.S. Highway 3—State Highway 4—County Road												1—Dry 2—Wet 3—Snow or Slush 4—Ice 5—Sand, Dirt, Oil 8—Other 9—Unknown													
LOCAL STREET 5—Township 6—Municipality 7—Frontage Road 8—Other 9—Unknown												TRAFFIC CONTROL DEVICES (See Instruction Manual)										84	85		
TRAFFICWAY IDENTIFIER (Actual Posted Number, Assigned Number, or Common Name (If No Posted or Assigned Number) Except: None Fill if Unknown)										43															
MILEPOINT (Actual to Nearest .1 Mile (Assumed Decimal) Except: 00000—None 99999—Unknown)										63															
												TRAFFIC CONTROL DEVICE FUNCTIONING										86			
												0—No Controls 1—Device Not Functioning 2—Device Functioning—Functioning Improperly 3—Device Functioning Properly 9—Unknown													

Fig. 4.2. FARS (Fatality Analysis Reporting System) Coding Forms (Accident Level)
(Source: NHTSA 1995)

Agent: Perceiving the environment requires some form of model through which observation can be made. Human beings form their own mental model through their knowledge, experiences, and preferences. The mental model does not necessarily conform to a formal logic (Johnson-Laird et al. 1998). For instance, a person who has spent all their life in the same city is likely to have a less erroneous (i.e., closer correspondence between the mental model and reality) mental model of this locality than new comers. Even if two persons have the same level of familiarity with the city, they might have different mental models depending on their unique experiences and preferences. In short, there exist inter-personal differences in the perception of locality.

Medium: In the traffic accident example reported above, the coding forms control the granularity and structure of the information transferred from human knowledge. Coding forms in coarse granularity would not receive human knowledge in finer granularity. That is, information is lost through the medium whose resolution does not accommodate the resolution at which knowledge is expressed. For exam-

ple, police officers can record the specific location of highway accidents using a milepoint (shown in the lower left of coding forms). But the location of the accident on local streets can only be recorded at a rough scale because local streets are not measured by a milepoint. A poorly designed medium can cause ambiguity as well. For example, the ‘city’ in the coding form can be interpreted as either ‘urbanized area’ or ‘legally recognized and bounded jurisdiction’, even if the latter may be intended. Consequently, an imperfect medium can also be responsible for the location indeterminacy of locality.

In Section 4.3 we review related research to examine solutions to the problem of location indeterminacy. We start with a discussion on the ontology of locality and then review appropriate formalisms of locality. Under the circumstances that the boundaries of localities are indeterminate, the membership of a place to a locality is not necessarily binary (i.e., True or False). We propose an approach that uses the notion of nearness to handle borderline cases (i.e. not sure if it is *in*). Environmental factors that are expected to affect the perception of nearness are reviewed so that they can contextualize the fuzzy set membership function of locality.

4.3 Theoretical Underpinnings of ‘Locality’

4.3.1 Ontology of Locality

Ontology research in the spatial domain has concentrated on the nature of spatial objects and their boundaries, their mereological (i.e., parts and whole), topological (Egenhofer and Franzosa 1991, Allen 1983), and mereotopological structure and their location in space and time. Smith and Mark (1998) argue that the ontological characterization of geographic reality requires considering three basic aspects: (1) the aspects of *what* spatial objects are, (2) the aspect of *where* spatial objects are, (3) aspects of *scale*. We can distinguish between spatial objects and abstract objects. A spatial object is an object that is located in space and time, while abstract objects are not located in space and time (e.g., numbers and prepositions). Thus, location is an inherent property of spatial objects. Spatial objects of geographic scale are larger than the human body and cannot be perceived within a single perceptual act. For example, forest and ocean are spatial objects of geographic scale whereas a human organ and Mars are not. More often than not, the locations of spatial objects are indeterminate.

Smith (1995) pointed out the fundamental distinction between *bona-fide* and *fiat* (spatial) objects. Roughly, *bona-fide* objects are objects which boundaries coincide with discontinuities of the underlying reality such as ‘The planet Earth’, human beings, tennis balls, and so on. *Fiat* objects, on the other hand are the result of human conceptualization and demarcation. Examples of *fiat* objects are countries, Federal states, and land property. They also include objects that have *bona-fide* and *fiat* boundary parts such as Mount Everest, dunes, and the Atlantic Ocean.

When climbing Mount Everest it is perfectly clear where the *bona-fide* boundary between rock and air is, but it is indeterminate where the *fiat* boundaries of Mount Everest among its foothills are. *Fiat* boundaries are not directly observable in reality unless they are explicitly marked as, for example, in built environments (Bittner 2000). Obviously, this causes problems when sharing information about such objects with others or when representing such information on a computer.

In general we need to distinguish at least two different kinds of vagueness of concepts that carve out *fiat* objects: (1) the vagueness of the *identity* condition, and (2) the vagueness of the *unity* conditions. Identity is related to the problem of distinguishing one instance of a class from other instances by means of a characteristic property that is unique for it. Unity is related to the problem of distinguishing the parts of an instance from the rest of the world by means of a unifying relation that binds them together (and not involving anything else) (Guarino and Welty 2000). Mount Everest satisfies the identity condition (is this Mount Everest?), but fails to satisfy the unity condition (does this foothill belong to Mount Everest?).

Now let us look at the ontology of locality based on the formal theory described above. Locality is defined as a surrounding or nearby region (WordNet 1.7.1), or a particular place (Webster-Merriam dictionary). Locality is a spatial object of geographic scale and falls into *fiat* objects. In general, the identity condition of locality holds. However, the unity condition is not necessarily met because the boundaries between localities are not clearly demarcated in mind. Therefore, localities are subject to location indeterminacy. The vague unity condition of locality can suggest that there exist multiple candidates for being the exact spatial region of locality. It allows us to divide the parts of locality in dealing with the multiple candidate situations, which will be discussed in the next section.

4.3.2 Formalism of Locality

A GIS often advocates an entity-oriented view of spatial phenomena. As a result, the vector format has been widely used as a spatial data type. So far, it has been implicitly assumed that the extent, and hence the boundary, of spatial objects is precisely determined. The properties of space are given by attributes whose values are assumed to be constant over the total extent of the objects, whether they are points, lines, or regions (Erwig and Schneider 1997). Increasingly, researchers are beginning to realize that there are many spatial objects in reality that do not have sharp boundaries or whose boundaries cannot be precisely determined. Examples are natural, social, or cultural phenomena with variant properties such as dialectal regions in North America, deserts, vegetation zones, and vernacular localities. We roughly define this kind region as a *fuzzy region*.

According to Schneider (1999), there are at least three possible, related interpretations for a point in a fuzzy region. First, this situation may be interpreted as the degree to which this point *belongs to* some areal feature (being inside or part of). For instance, there is no strict boundary between mountain and valley, and it seems to be more appropriate to model the transition by partial and multiple memberships. Second, the situation may indicate the degree of *compatibility* of the in-

dividual point with the attribute or concept represented by the region with indeterminate boundaries. An example is “warm areas” where we must decide for each point whether and to which grade it corresponds to the concept “warm”. Third, this situation may be viewed as the degree of *concentration* of some attribute associated with a fuzzy region at the particular point. An example is air pollution where we can assume the highest concentration in the direct vicinity of power plant, for instance, and lower concentrations with increasing distance from them.

Let us consider the example of traffic accidents. The upper portion of the accident coding form pertains to where the accident occurs. The issue here is the spatial relation between a point (accident) and a region (locality). GIS can find the binary relation (i.e., in or not in) using a point-in-polygon operation. However, what if the boundary of a region is not crisp in reality, unlike the data stored in the GIS database? In this case, it is hard to state whether the accident occurred *in* the locality. Alternatively, it is also plausible to say that the accident occurred *near* some area. Due to the indeterminate boundary, we need to deal with the borderline cases. A crisp data model cannot provide a realistic solution to this situation because the membership has to be either true or false. Fuzzy set theory can provide a better insight on partial or multiple memberships. Therefore, locality can be interpreted as the *belonging to* case in georeferencing applications.

Now the next natural question is how we can represent a fuzzy region given insufficient knowledge about the grade of indeterminate parts of localities. Locality is the result of human conceptualization unlike natural process such as weather pattern or a soil type. Consequently the semantic import model (Robinson 1988, Fisher 2000) using expert knowledge may not be a plausible option. An empirical model can be properly applied to locality, but it requires a sufficient amount of data from which the generally accepted model can be derived. While the issue of deriving fuzzy set membership is deferred to the next section, we take care here of delimiting the vague parts of a fuzzy region using lower (definitely in) and upper (possibly in) bounds of locality. The approximation of these bounds rests on the categorization of the geographic space into three components. That is, locality can be divided into three parts: *core*, *boundary*, and *exterior*, as shown in Figure 4.3. The three parts respectively relate to those parts that definitely belong, perhaps belong, and definitely do not belong to a specified locality. We can assign the membership value 1 to each point of the *core*, value 0 to each point of the *exterior*, and value in $[0, 1]$ (somewhere between completely true and completely false) to each point of the *boundary*. This model can approximate many different situations as discussed in Section 2. Some of these situations are enumerated as representations based on incomplete information (e.g., missing locality information in the reference data), conflicting information (e.g., conflicting zone types), and changing information (e.g., changing perception of locality, or changing environments such as urban growth, and urban sprawl over time), as well as representations of inherently vague concepts (e.g., the indeterminacy of location) (Worboys 2001).

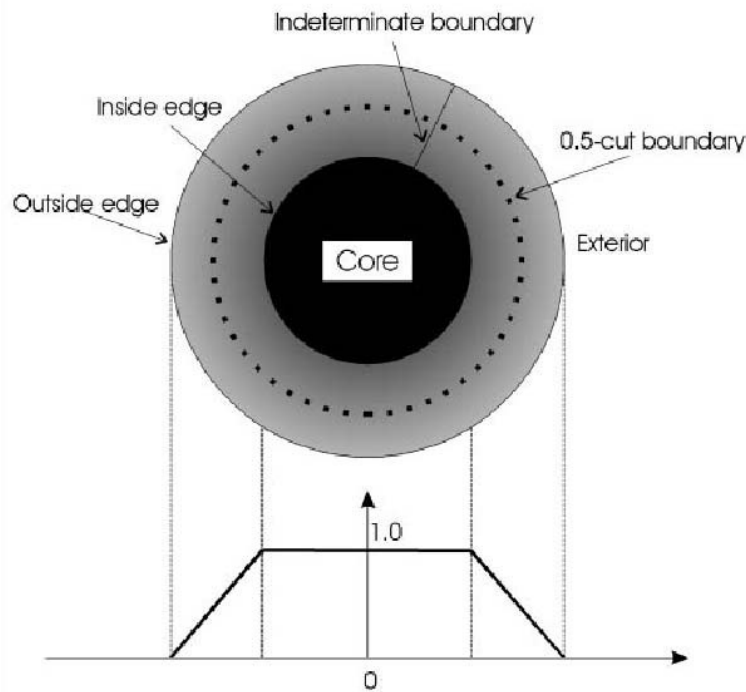


Fig. 4.3. Representation of a Fuzzy Region and its Fuzzy Set Membership Function (Source: Zhan and Lin 2003)

The problem now boils down to delineating those three parts, and defining the grade of fuzzy set membership function. Assigning fuzzy set membership value for the *boundary* part (i.e., perhaps belong) is similar to determining how close the point (accident) is to the area (locality). Therefore, examining the notion of nearness can give important insights into the fuzzy set membership function. More specifically, the next section examines the factors that influence the human perception of nearness. We derive factors (or concepts) from the theory of nearness, and then we apply them to the fuzzy set membership function of locality, which is the main concern of this paper.

4.3.3 Nearness

According to Gahegan (1995), human perception and cognition of nearness (or proximity) is influenced by the following: (1) in the absence of other objects, humans reason nearness in a geometric fashion. Absolute distance is the major factor that affects nearness. Furthermore, the relationship between distance and nearness can be approximated by a linear relationship (some researches suggest an S-

shaped function). (2) When other objects of the same type are introduced, nearness is judged in part by relative distance. For example, London and Milan may be considered to be close in the absence of Paris. But they are considered to be far apart when Paris is introduced. In a linguistic term, Milan is close to London. Milan is far from London compared to Paris. (3) Distance is affected by the size of the area being considered. That is, the reference frame plays a significant role in comparing distances between objects. For instance, Milan may be far from London in a European reference frame, but they are close to each other in a world reference frame. Therefore, the scale of the reference frame influences the perceived distance.

To get further insight into the perception of nearness, we can consider the intervening opportunities (Stouffer 1940) model based on the principle of the least effort. This model states that individuals consider opportunities that are closest to them first, and if they find them unacceptable they will go on to the next closest opportunity or opportunities, and so on. By delineating the opportunity set by their proximity relationship to the reference location, we can obtain a set of sequentially embedded neighborhoods. In the context of the spatial object 'locality', it suggests that nearness can be significantly affected by the order of neighborhoods. In other words, nearness can be qualitatively defined as the order (or lag) of neighborhoods. In Figure 4.4, C is considered to be closer to A than H because C is in the 2nd order neighborhood and H is in the 3rd order neighborhood even though they are apart from A with similar distances. This conceptualization implies that the relative location of neighborhoods determines the extent of nearness surface. Other contextual factors important for judgments on nearness include connection paths between places, attractiveness of objects, type of activity to be undertaken, traffic conditions, transportation mode, as well as personal characteristics (Yao and Thill 2005).

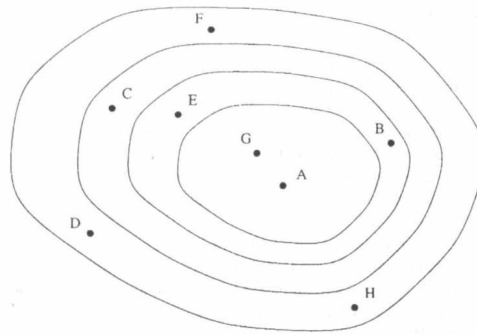


Fig. 4.4. Nearness Illustrated as a Set of Neighborhoods (Source: Guesgen and Albrecht 2000)

4.4 Modeling Localities by Fuzzy Sets

4.4.1 Formal Definition of Locality

Let us consider a locality indexed by l . This locality is a fuzzy region denoted by $\tilde{A}_l$. $\tilde{A}_l$ is composed of the following three parts: *Core*, *Boundary*, and *Exterior*. These parts are defined as crisp regions (regular closed sets) denoted as reg_c . Also let $\mathcal{R}^2$ be the two dimensional geographic space and let $\mu_{\tilde{A}_l}$ be the fuzzy set membership function of $\tilde{A}_l$. $\tilde{A}_l$ and $\mu_{\tilde{A}_l}$ are defined as follows:

$$\begin{aligned}\tilde{A}_l &= Core(\tilde{A}_l) \vee Boundary(\tilde{A}_l) \vee Exterior(\tilde{A}_l) \\ Core(\tilde{A}_l) &= reg_c (\{(x,y) \in \mathcal{R}^2 \mid \mu_{\tilde{A}_l}(x,y) = 1\}) \\ Exterior(\tilde{A}_l) &= reg_c (\{(x,y) \in \mathcal{R}^2 \mid \mu_{\tilde{A}_l}(x,y) = 0\}) \\ Boundary(\tilde{A}_l) &= reg_c (\{(x,y) \in \mathcal{R}^2 \mid 0 < \mu_{\tilde{A}_l}(x,y) < 1\})\end{aligned}$$

The *core* identifies the part that definitely belongs to $\tilde{A}_l$. The *exterior* determines the part that definitely does not belong to $\tilde{A}_l$. The indeterminate character of $\tilde{A}_l$ is summarized in the *boundary* of $\tilde{A}_l$ in a unified and simplified manner. The *core* and *boundary* can be adjacent with a common border, and *core* and/or *boundary* can be empty. When the *boundary* is an empty set, $\tilde{A}_l$ becomes a crisp region. Thus, a crisp region is a special case of a fuzzy region.

The generalized membership function of $Boundary(\tilde{A}_l)$ is a weighted average of neighboring points in the geographic space:

$$\mu_{\tilde{A}_l}(x,y) = \sum_{i=1}^n w_i(x,y) \mu_{\tilde{A}_i}(x,y)$$

where $\sum w_i(x,y) = 1$ and $w_i(x,y) \propto h_i^{-1}$ in which h_i is the distance from (x,y) to $(x,y)_i$. The nonzero weight $w_i(x,y)$ is only given to the points that meet the specific criteria to be described in the next section. The fuzzy set membership value of $Boundary(\tilde{A}_l)$ can be seen as a z-value in the plane that is fitted to *core* and *exterior*. Therefore, delineating *core* and *exterior* determines the extent and grade of fuzzy set membership function in *boundary*. Additionally, *0.5-cut boundary* is introduced as a way to make the interpolation compact. The delineation of *core*, *0.5-cut boundary*, and *exterior* is formalized using predicates that are introduced in the next section.

4.4.2 Formal Properties of Locality

Every locality implicitly has the *Resolution-Level*. Let an instance of locality be x . The *Resolution-Level*(x) (or $RL(x)$) is defined as a spatially hierarchical structure of administrative boundaries or communities. Suppose we have n members of the *Resolution-Level* say $RL_1, RL_2, \dots, RL_n$. Each member of the *Resolution-Level* must be in the hierarchical order; thus the order is consistently assigned to each

member of the *Resolution-Level*(x) as denoted by the subscript. We use $\{State, County, TownorCity, Place\}$ as the members of *Resolution-Level* for our applications, where n is 4, and each member has the resolution order (Liu and Satur 1999). Denote the resolution order of $RL(x)$ as *Resolution-Order*($RL(x)$) (or $RO(RL(x))$). For example, $RO(RL(x))$ is 1 when the *Resolution-Level* of locality x is *State*, and so on. *Difference*($RO(RL(x)), RO(RL(y))$) returns the difference in resolution order between two resolution levels

$$\forall x \text{ Resolution-Level}(x) = \{RL_1, RL_2, \dots, RL_n\}$$

$$\forall x, y \rightarrow \text{Domain of Locality in the United States}$$

$$\forall x \text{ Resolution-Level}(x) = \{State, County, TownorCity, Place\}$$

$$\exists x \text{ Place} \subseteq \text{TownorCity} \subseteq \text{County} \subseteq \text{State}$$

$$\text{Resolution-Order}(RL(x)) = \begin{cases} 1 & \text{if } RL(x) = State \\ 2 & \text{if } RL(x) = County \\ 3 & \text{if } RL(x) = TownorCity \\ 4 & \text{if } RL(x) = Place \end{cases}$$

$$\text{Difference}(RO(RL(x)), RO(RL(y))) = |RO(RL(x)) - RO(RL(y))|$$

Regarding the relationship between two localities x and y , we can distinguish two types of geographic neighborhoods: *horizontal* and *vertical* neighborhood. A horizontal neighborhood refers to a competing neighborhood in the same *Resolution-Level* while a vertical neighborhood is a compatible neighborhood in the different *Resolution-Level*. For example, San Francisco and Los Angeles is the horizontal neighborhood while San Francisco and California is the vertical neighborhood. Denote horizontal neighborhood as *HN* and vertical neighborhood as *VN* in terms of 2-ary predicates. They satisfy symmetric properties, and can be extended to n -ary predicates. If the resolution level of x and y is the same, they are horizontal neighborhood. If the resolution level of x and y is not the same, they are vertical neighborhood.

$$[RL(x) = RL(y)] \Rightarrow HN(x, y)$$

$$[RL(x) \neq RL(y)] \Rightarrow VN(x, y)$$

Locality x has a *Proximity-Order* with respect to other locality y . *Proximity-Order*(x, y) quantifies the ordinal proximity between x and y . Similar to geographic neighborhood, we distinguish two dimensions of proximity order: *Horizontal-Proximity-Order*(x, y) and *Vertical-Proximity-Order*(x, y). *Horizontal-Proximity-Order*(x, y) is 1 when a region x meets other region y where their neighborhood relation is horizontal. The definition of predicate *Meet*(x, y) follows that of Egenhofer and Franzosa (1991).

$$\begin{aligned}
& \text{Horizontal-Proximity-Order } (x, y) \rightarrow \mathbb{N} \\
& HN(x, y) \wedge \text{Meet}(\text{Core}(\tilde{A}_x), \text{Core}(\tilde{A}_y)) \Rightarrow HPO(x, y) = 1 \\
& HN(y, z) \wedge \text{Meet}(\text{Core}(\tilde{A}_y), \text{Core}(\tilde{A}_z)) \Rightarrow HPO(y, z) = 1
\end{aligned}$$

Horizontal-Proximity-Order(x, z) is 2 by transitivity if $HPO(x, y) = 1$ and $HPO(y, z) = 1$.

$$[HPO(x, y) = 1] \wedge [HPO(y, z) = 1] \Rightarrow HPO(x, z) = 2$$

The function *FirstOrderHNGroup* of x is defined as the sum of all y whose horizontal proximity order with respect to x is 1. Similarly, the function *SecondOrderHNGroup* of x is defined as the sum of all z whose horizontal proximity order with respect to x is 2.

$$\begin{aligned}
\text{FirstOrderHNGroup}(x) &= \sum_{i=1}^n y_i \text{ where } \forall y_i [HPO(x, y_i) = 1] \\
\text{SecondOrderHNGroup}(x) &= \sum_{i=1}^n z_i \text{ where } \forall z_i [HPO(x, z_i) = 2]
\end{aligned}$$

Vertical-Proximity-Order(x, y) is 1 when a region x is inside another region y , and y is the one-level-higher vertical neighborhood of x . The predicate *Inside*(x, y) is not the same as *Inside*(y, x), thus the predicate is non-symmetric unlike *Meet*(x, y).

$$\begin{aligned}
& \text{Vertical-Proximity-Order } (x, y) \rightarrow \mathbb{N} \\
& VN(x, y) \\
& RO(RL(y)) > RO(RL(x)) \\
& \text{Difference}(RO(RL(x)), RO(RL(y))) = 1 \\
& \text{Inside}(\text{Core}(\tilde{A}_x), \text{Core}(\tilde{A}_y)) \\
& \Rightarrow VPO(x, y) = 1 \\
& VN(y, z) \\
& RO(RL(z)) > RO(RL(y)) \\
& \text{Difference}(RO(RL(y)), RO(RL(z))) = 1 \\
& \text{Inside}(\text{Core}(\tilde{A}_y), \text{Core}(\tilde{A}_z)) \\
& \Rightarrow VPO(y, z) = 1
\end{aligned}$$

Vertical-Proximity-Order(x, z) is 2 by transitivity if $VPO(x, y) = 1$ and $VPO(y, z) = 1$.

$$[VPO(x, y) = 1] \wedge [VPO(y, z) = 1] \Rightarrow VPO(x, z) = 2$$

The function *FirstOrderVNGroup* of x is defined as the sum of all y whose vertical proximity order with respect to x is 1. Similarly, the function *SecondOrderVNGroup* of x is defined as the sum of all z whose vertical proximity order with respect to x is 2.

$$[VPO(x, y) = 1] \Leftrightarrow FirstOrderVN(x, y)$$

$$[VPO(y, z) = 1] \Leftrightarrow FirstOrderVN(y, z)$$

$$[VPO(x, z) = 2] \Leftrightarrow SecondOrderVN(x, z)$$

$$FirstOrderVNGroup(x) = \sum_{i=1}^n y_i \text{ where } \forall y_i [VPO(x, y_i) = 1]$$

$$SecondOrderVNGroup(x) = \sum_{i=1}^n z_i \text{ where } \forall z_i [VPO(x, z_i) = 2]$$

To illustrate the point, let us consider locality 'Syracuse' as a fuzzy region. *Resolution-Level* of Syracuse is *TownorCity*. Vertical neighborhoods of Syracuse are 'Onondaga' in the *Resolution-level* of *County*, and 'New York' in the *Resolution-level* of *State*. Horizontal neighborhoods of 'Syracuse' are other localities in the same *Resolution-level* such as 'Clay', 'De Witt', and 'Pompey' shown in Figure 4.5. Let 'Syracuse' be x , then *Horizontal-Proximity-Order*(x , 'Salina') is 1 whereas *Horizontal-Proximity-Order*(x , 'Clay') is 2. Thus, *FirstOrderHN*(x , 'Salina') and *SecondOrderHN*(x , 'Clay') hold. In Figure 4.5, it can be seen that *FirstOrderHNGroup*(x) is {'Salina', 'Geddes', 'De Witt', 'Onondaga'} and *SecondOrderHNGroup*(x) is {'Clay', 'Lysander', 'Van Buren', 'Camillus', 'Marcellus', 'Otisco', 'La Fayette', 'Pompey', 'Manlius', 'Cicero'}.

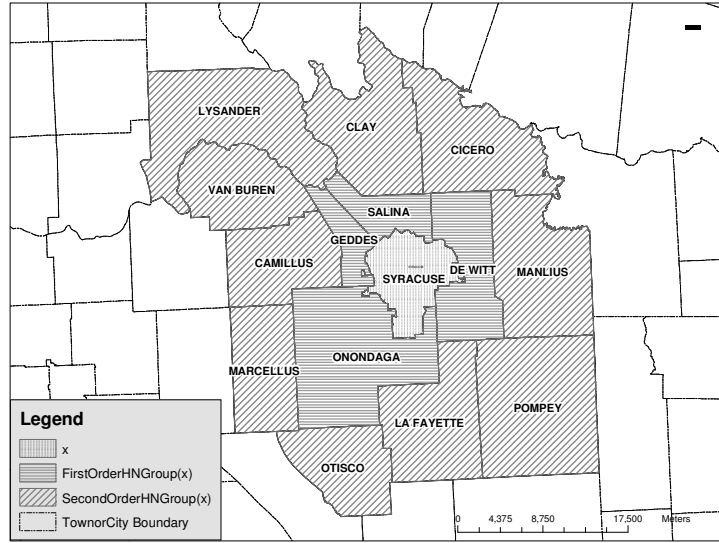


Fig. 4.5. Illustration of *FirstOrderHNGroup*(x) and *SecondOrderHNGroup*(x)

Likewise, an example of vertical neighborhoods is given in Figure 4.6. Let 'Syracuse' be x , then *Vertical-Proximity-Order*(x , 'Onondaga') is 1 whereas *Vertical-Proximity-Order*(x , 'New York') is 2. Thus, *FirstOrderVN*(x , 'Onondaga') and *SecondOrderVN*(x , 'New York') hold. *FirstOrderVNGroup*(x) is 'Onondaga' and *SecondOrderVNGroup*(x) is 'New York'.

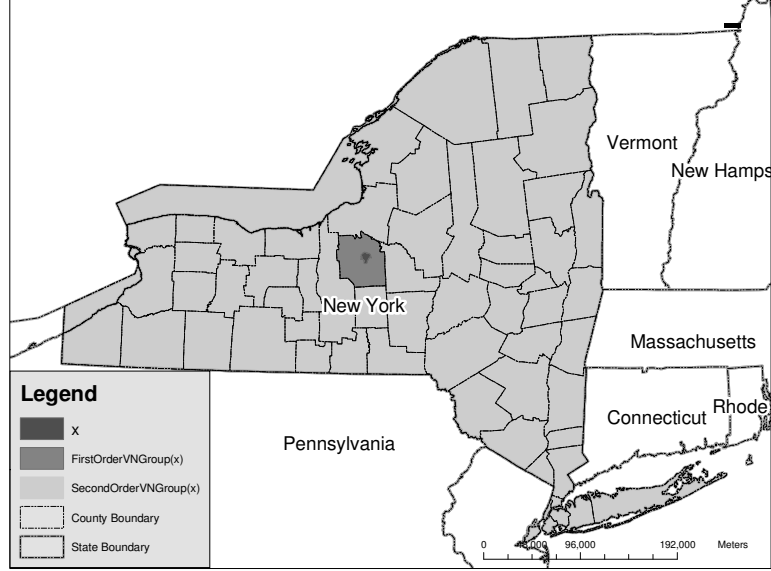


Fig. 4.6. Illustration of *FirstOrderVNGroup*(x) and *SecondOrderVNGroup*(x)

4.4.3 Fuzzy Set Membership Function of Locality

The indeterminate parts of locality l can be replaced with predicates defined above. The combinations of *FirstOrderHNGroup*(l) and *SecondOrderHNGroup*(l) comprise the (indeterminate) *boundary* of locality as a fuzzy region. Within *boundary*, the outside edge of *FirstOrderHNGroup*(l) serves as *0.5-cut boundary* is defined as. Additionally, the extent of locality as a fuzzy region is delimited by *FirstOrderVNGroup*(l).

$$\begin{aligned} \text{Boundary}(\tilde{A}_l) &= \text{FirstOrderHNGroup}(l) \vee \text{SecondOrderHNGroup}(l) \\ \text{where } 0.5\text{-cut boundary } (\tilde{A}_l) &= \text{outside-edge } (\text{FirstOrderHNGroup}(l)) \\ \tilde{A}_l &\subseteq \text{FirstOrderVNGroup}(l) \end{aligned}$$

Accordingly, the fuzzy set membership value can be redefined as follows:

$$(x,y) \in \mathcal{R}^2$$

$$\mu_{\tilde{A}_l}(x,y) = \begin{cases} 0 & \text{if } (x,y) \text{ is at } \text{Exterior}(\tilde{A}_l) \\ [0, 0.5] & \text{if } (x,y) \text{ is in } \text{SecondOrderHNGroup}(l) \\ [0.5, 1] & \text{if } (x,y) \text{ is in } \text{FirstOrderHNGroup}(l) \\ 1 & \text{if } (x,y) \text{ is in } \text{Core}(\tilde{A}_l) \end{cases}$$

To compute the fuzzy set membership value in $\text{Boundary}(\tilde{A}_l)$, we create Delaunay Triangulation whose nodes are comprised of any vertices on *core*, *0.5-cut boundary*, and *exterior*. The membership value is obtained by intersecting a vertical line with the plane defined by the three nodes of the triangle (Figure 4.7). In Figure 4.7, any location within this plane is guaranteed to have a fuzzy set membership value between 0.5 and 1.

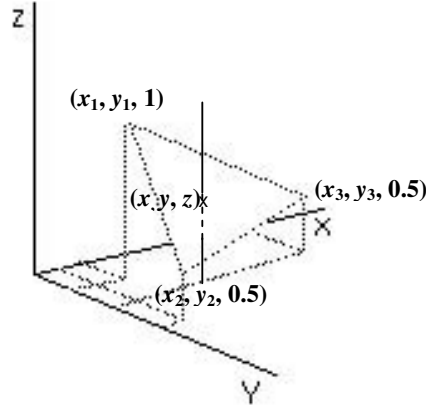


Fig. 4.7. TIN Surface Created to Interpolate the Membership Value

In Figure 4.7, the generalized equation for linear interpolation of a point (x, y, z) in a triangle facet is:

$$Ax + By + Cz + D = 0$$

where A, B, C, and D are constants determined by the coordinates of the triangle's three nodes. Thus, the fuzzy set membership value can be obtained from the following equation given the x- and y- coordinates:

$$\mu_{\tilde{A}_l}(x,y) = (-Ax - By - D) / C$$

4.5 Implementing the Fuzzy Locality Model in GIS

In this section, we describe the procedures used to georeference 8631 fatal accident records in New York State during the period 1996-2001. The accident records are matched against the reference data such as road network and locality layers. The records are matched on the basis of how similar two features (trafficway and locality) recorded in the accident database are to those in the reference data. Fuzzy locality layers are created (derived) from (crisp) locality layers according to the formalism presented in Section 4.4. The overall quality of a match is computed as the average of two similarity scores. The road segment with the best score is chosen as the best candidate. If the score of the best candidate is above a preset threshold, the georeferencing result is accepted. Otherwise, we reject the result.

4.5.1 Data Sets

Source Data

The Fatality Analysis Reporting System (FARS) contains data on a census of fatal traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. To be included in FARS, a crash must involve a motor vehicle traveling on a trafficway customarily open to the public and result in the death of a person (occupant of a vehicle or a non-occupant) within 30 days of the crash.

In Table 4.1, we identify the data fields that may be of use for georeferencing purposes. The rightmost column of the table labels them. More particularly, RG stands for “relevant to georeferencing” while RGBI indicates “relevant to georeferencing, but incomplete.” For instance, data field TRAFFICWAY IDENTIFIER is definitely used for geo-referencing purposes, whereas SPECIAL JURISDICTION is not likely to be used because most values are left blank.

Table 4.1. Location-related Fields in the FARS Accident Data

ACCIDENT LEVEL DATA FIELDS	TYPE	START	LENGTH	GEORef.
STATE	N	1	2	RG
CITY	N	13	4	RG
COUNTY	N	17	3	RG
TRAFFICWAY IDENTIFIER	A/N	42	20	RG
MILEPOINT	N	62	5	RG
SPECIAL JURISDICTION	N	67	1	RGBI
RELATION TO JUNCTION	N	71	2	RG
LATITUDE	A/N	116	8	RGBI
LONGITUDE	A/N	124	9	RGBI

Reference Data

Table 4.2 lists the reference data that we used for georeferencing accident records of the FARS database. The selection of reference data is based on two primary considerations. First, we consider how relevant/compatible the data is to the source data. Second, we compromise between data accessibility and data quality. For instance, TIGER/LINE roads are chosen over other road network data that may be more accurate because TIGER is in the public domain.

Table 4.2. Sources of Reference Data

Relevant FARS Field	Organization	Layer Name	Data Source (URL or product)
			Metadata Documentation
Tway_id	DOT-FHWA	NHPN	http://www.fhwa.dot.gov/planning/nhpn http://www.fhwa.dot.gov/planning/nhpn/docs/metadata.html
	US Census Bureau	TIGER	http://arcdata.esri.com/data/tiger2000/tiger_download.cfm http://www.census.gov/geo/www/tiger/rd_2ktiger/tlrdmeta.txt
	NY State GIS clearinghouse: DOT	CLASS	https://www.nysgis.state.ny.us/s_dot/regional.html http://www.nysgis.state.ny.us/repository/dotlist.htm
County	NY State GIS clearinghouse: DOT	Nyshore	https://www.nysgis.state.ny.us/s_dot/data/state/nyshore.zip (user login required) http://www.nysgis.state.ny.us/gis3/data/dot.nyshore.html
City	NY State GIS clearinghouse: DOT	Nybndry	https://www.nysgis.state.ny.us/s_dot/data/state/nybndry.zip (user login required) http://www.nysgis.state.ny.us/gis3/data/dot.nybndry.html
	Caliper	Ccplacec	Academic TransCAD [®] version 3.5d (program CD) N/A
	GSA	GSA GLC table	http://www.gsa.gov/attachments/GSA_PUBLICATIONS/extpub/glcout_1.zip N/A

4.5.2 Data Preprocessing

It is necessary to preprocess the reference data from multiple sources. Major tasks include (1) transforming different coordinate systems (for example, NHPN and Nybndry are stored in UTM whereas TIGER and Ccplacec are in decimal degrees) into a single unified one, (2) creating locality layers against which the locality

code in FARS is matched, (3) populating the field used to join FARS table to a spatial data (i.e., writing GSA code to locality layer), and (4) assembling necessary spatial data. We only describe the second and third ones due to the relevance to the content in this chapter.

Creating Locality Layers

We define four different *Resolution-Levels* in Section 4.2, including *State*, *County*, *TownorCity*, and *Place*. It is necessary to create locality layers that correspond to those levels respectively (Liu and Satur 1999). Three locality layers are created: COUNTY for the *Resolution-Level County*, PLACE_PL for *TownorCity*, and PLACE_PT for *Place* (Layer names are capitalized to tell them apart from other names). Nyshore (see Table 4.2) is renamed to the layer COUNTY.

PLACE_PL is extracted from Nybdry. The field CITY in FARS accidents (see Table 4.1) loosely refers to a place name. The field does not distinguish between different types of zones. On the contrary, the reference data Nybdry distinguishes them. The *region* (i.e., Arc/Info feature class) coverage Nybdry consists of multiple layers (including county, town, city, village, airport, and so on) to which the value of the field CITY can be related. For example, the Town of Amherst (156) and the City of Buffalo (750) are coded in the same field CITY in FARS accidents without regard for the fact that these localities belong to different classes of administrative units. On the other hand, the reference data Nybdry stores towns and cities as distinct layers. For matching purposes, the Town and City layers in Nybdry are merged into a new polygon layer called PLACE_PL. They can be merged because those layers do not overlap.

Layer PLACE_PT is extracted from Ccplacec, which contains the centroids of places. Only features appropriate to our applications are selected to populate this new layer. First, features within New York State are selected. In a second stage, features that are included in the GSA geographic code table are selected out of this first set. Since PLACE_PT does not have a polygon topology, Thiessen polygons are created from the point features in order to approximate their polygon boundaries. Locality layers are illustrated in Figure 4.8.

When it comes to the difference between PLACE_PL and PLACE_PT, PLACE_PL refers to administrative boundaries whereas PLACE_PT refers to the place with vague boundaries. For example, 'New York' can be found in PLACE_PL while 'Greenwich Village' can be found in PLACE_PT.

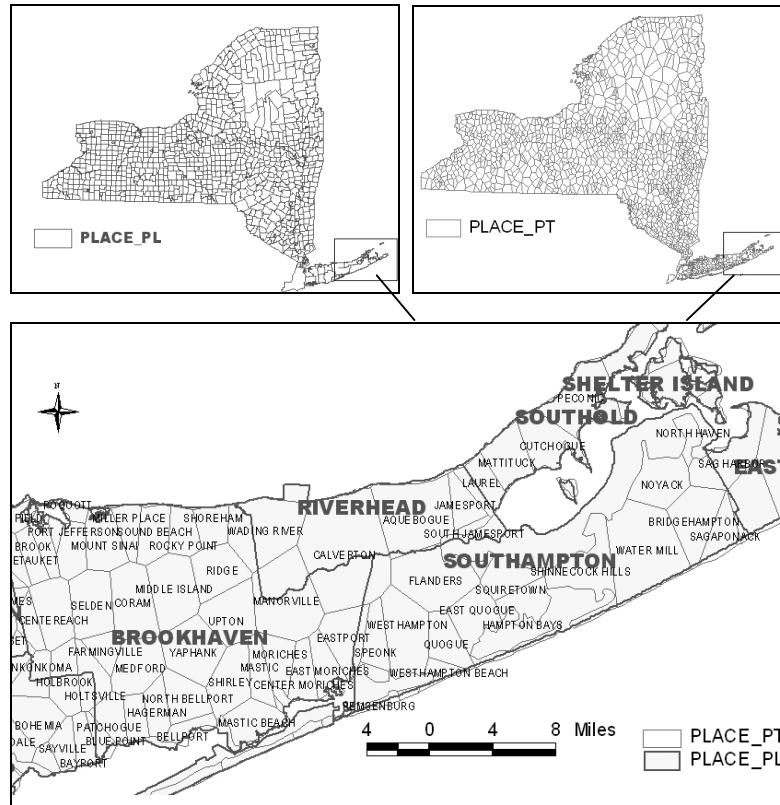


Fig. 4.8. Locality Layers: PLACE_PL and PLACE_PT

Writing GSA Codes to Locality Layers

The locality-related fields of the FARS database are populated with GSA codes, but entities of the locality layers do not contain a GSA code. In order to match each FARS data record to entities of the locality layers, we need to create a key filed for joining. The GSA code table is used to relate locality (in FARS) to locality layers, as depicted in Figure 4.9. In this figure, GSA City code 0010 can be derived and written to locality layers using the match through the field NAME between the GSA code table and the Reference GeoData layer.

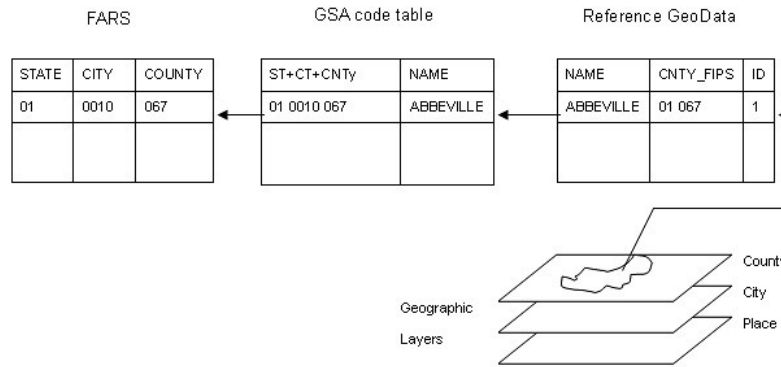


Fig. 4.9. Writing GSA Codes to Locality Layers

4.5.3 Georeferencing Procedures

Given the intrinsic differences between the coding schemes of accidents on highways and on local streets, the georeferencing problem is divided into two procedures implemented in parallel. One is to use dynamic segmentation tied to the highways' linear referencing system (LRS) of mile points, while the other involves a similarity-based matching on local streets. The two types of matching lead to different spatial resolutions under which crashes can be positioned. Dynamic segmentation identifies the exact point of location while the local street matching merely identifies the road segments on which the crash is most likely to have occurred. (Accidents have not been reported in an address-style for more than two decades in FARS databases. NHTSA is in the process of switching to a longitude/latitude geographic coding system that is fully compatible with contemporary GPS-based data collection techniques). Simplified matching rules are specified in Figure 4.10. In this figure, the notation " $\cong$ " is roughly defined as equivalence, similarity, consistency, or fuzzy proximity, depending on the case in hand.

Since the fuzzy set modeling of locality is only used in local streets matching, we exclusively focus on the local streets matching hereunder. The whole procedure of local streets matching is fully automated using Arc/Info 8.1 AML. Due to the rather low quality of FARS accident data (i.e. there are not many exact matches), similarity-based procedures have been implemented to account for the reliability of the matches. Each potential match is rated on the basis of a similarity score. The similarity score accounts for two features, namely the trafficway and the locality. In order to compute a trafficway similarity score, we have developed a thesaurus and a complete set of associational rules that account for the incompatibility between source and target (or reference) datasets. See Hwang and Thill

(2003) for more detail. With regard to the locality similarity score, the locality code in FARS is matched to that of fuzzy locality layers with a certain fuzzy set membership value. In contrast to a (crisp) locality layer where a single membership is assigned to each feature, a fuzzy locality layer can have n -possible memberships where n is the total number of features (Stefanakis et al. 1999). Thus, the match to the fuzzy locality layers yields any similarity score between 0 and 1 whereas a match to the (crisp) locality layer would yield a score that is either 0 or 1.

■ *Dynamic Segmentation*

```
[(FARS.TWAY_ID ≡ NHPN.SIGN1) OR (FARS.TWAY_ID ≡ NHPN.SIGN2) OR
(FARS.TWAY_ID ≡ NHPN.SIGN3) OR (FARS.TWAY_ID ≡ NHPN.LNAME)]
AND (NHPN.BEGMPT ≤ FARS.MILEPT ≤ NHPN.ENDMPT)
AND (FARS.CITY ≡ PLACE_PT.CITY_GSA)
AND (FARS.COUNTY = COUNTY.COUNTY_FIPS)
```

■ *Local Streets Matching*

```
[FARS.TWAY_ID ≡ TIGER.(FEDIRP+FENAME+FETYPE+FEDIRS)] OR
(FARS.TWAY_ID ≡ NHPN.LNAME) OR (FARS.TWAY_ID ≡ CLASS.NAME)
AND (FARS.CITY ≡ PLACE_PL.CITY_GSA) OR (FARS.CITY ≡ PLACE_PT.CITY_GSA)
AND (FARS.COUNTY = COUNTY.COUNTY_FIPS)
```

Fig. 4.10. Simplified Matching Rules Used in Georeferencing FARS Accidents

Fuzzy locality layers are created from the original locality layers PLACE_PL and PLACE_PT. For instance, when the original locality layer consists of 10 localities, ten fuzzy locality layers are created, each of which corresponds to an original locality feature. Each fuzzy locality layer has the fuzzy set membership value 1 for the core of the locality, and any value between 0 and 1 for the boundary of the locality. To illustrate this point, suppose that we create a fuzzy locality layer for locality l , say Buffalo. First, a fixed membership value is set to the *core*, the *0.5-cut boundary*, and the *exterior* of l . According to the conditions stated in Section 4.3, *core* is derived from the outside edge of l ; the *0.5-cut boundary* is the outside edge of the set of first-lagged localities surrounding l ; *exterior* is the outside edge of a set of second-lagged localities surrounding l . Next, the fuzzy set membership value is set to 1 for the *core*, 0.5 for the *0.5-cut boundary*, and 0 for the *exterior*. Third, a triangulated irregular network (TIN) is built on the vertices of these three line features. Fourth, the fuzzy set membership value is interpolated on the TIN facet. The steps of creating fuzzy locality layers are illustrated in Figure 4.11. Consequently, a continuous membership value between 1 and 0 is computed along the *boundary* region (Wang and Hall 1996) as shown in Figure 4.12.

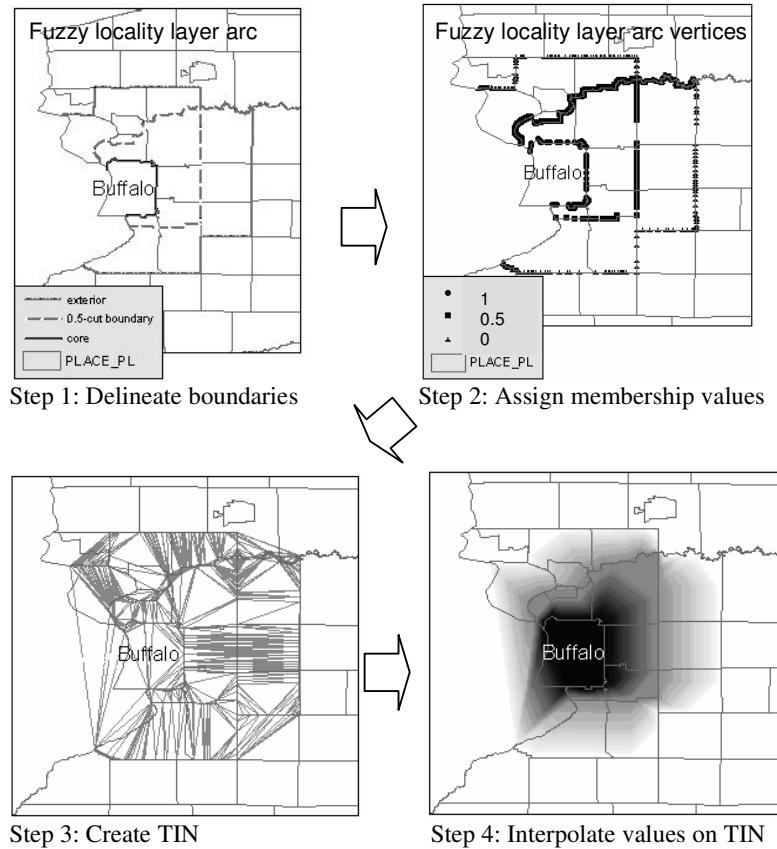


Fig. 4.11. Creating Fuzzy Locality Layers

The membership value is not only determined by the Euclidean distance, but also by the proximity order of neighboring localities. Because horizontal neighborhoods that compose the fuzzy parts of a locality are in the same resolution level, the membership value is also scale-dependent. In sum, Euclidean distance, neighborhood relation, and scale determine the fuzzy set membership value of locality. In Figure 4.12, the locality of Buffalo has a full membership in its *core*, a partial membership in its *boundary* (e.g. Amherst), and has no membership beyond its *exterior* (e.g. Alden).

We can illustrate the similarity-based georeferencing procedure using the example in Figure 4.12. Let us consider a FARS record where TWAY_ID is “Millersport Hwy”, and CITY corresponds to “Buffalo”. The reference data layers contain a record named Millersport Hwy, but this entity is in the nearby Town of

Amherst instead of the City of Buffalo. For this best match, the matching score is 0.5, which is the average of the trafficway score of 1 and of the locality score of 0. The latter score would be assigned by virtue of Millersport Hwy being outside of the Buffalo city limits. However, the matching score will be at least 0.5 (average of 1 and [0.5, 1]) with the similarity-based matching described above since the locality score is greater than zero considering the nearness of Amherst to Buffalo.

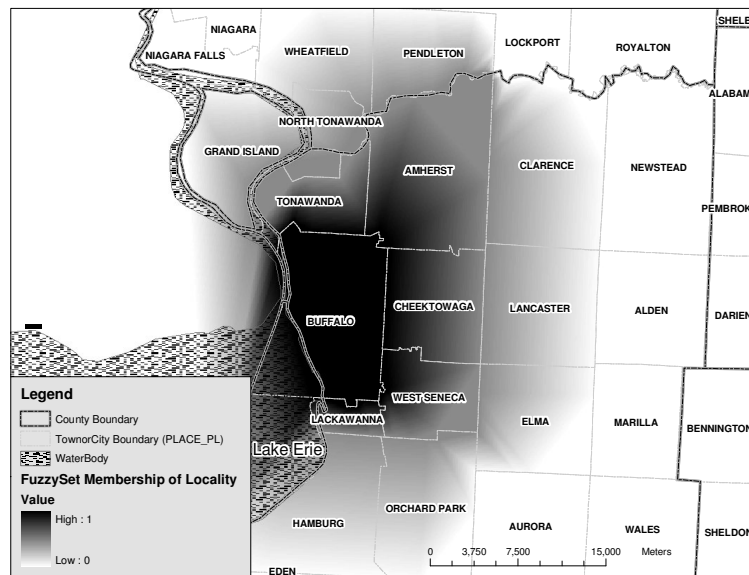


Fig. 4.12. Fuzzy Set Membership of the Locality of Buffalo

It may be useful to compare the grade of fuzzy proximity membership function used in the computation of the similarity score with other possible measures. Figure 4.13 illustrates the setting where the arrow that stretches outwards from the geographic center of the locality of Buffalo is depicted as an x-axis in Figure 4.14. The first option is to use distance as a measure of dissimilarity. Another option is to consider buffers delineated around the *Core*.

To measure the distance between a polygon (locality) and a point (accident), a representative point of polygon, such as the centroid, is required. This leads to a loss of geographic detail. Moreover, distance option ignores the anomalous proximity surface that is created by neighborhood relations. For instance, if the arrow (in Figure 4.13) is drawn along a different direction, the membership values will be adjusted to new neighborhood relations in the case of fuzzy proximity whereas the distance option maintains the membership value regardless of the direction. Creating multiple buffers around the *Core* does not require a centroid, but proximity decreases in an abrupt manner along the buffer boundary. As for the distance

option, it does not take into account the anomalous geometry of proximity. Moreover, the buffer interval has to be set based on the absolute distance while the fuzzy proximity can be determined by relative distances. Figure 4.14 shows the differences in membership grades under each of the three options.

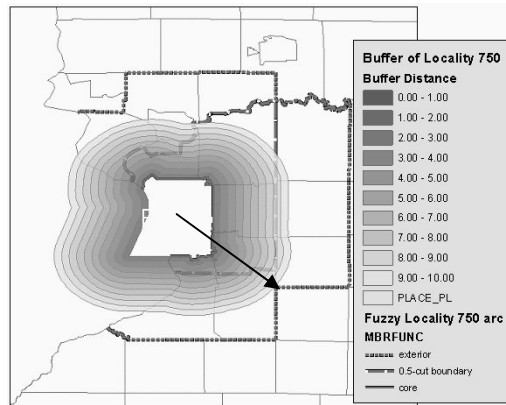


Fig. 4.13. The Setting for the Proximity Measures in a Geographic Space Where the Arrow is the Transect Depicted in Figure 4.14.

In Figure 4.14, the x-axis represents the distance from the centroid of the *Core* of the locality while the y-axis represents the fuzzy set membership value. If the arrow is drawn along another direction in Figure 4.13, the membership value will be adjusted to the new neighborhood relations only in the case of fuzzy proximity, while distance- and buffer-based measures keep the same uniform membership value.

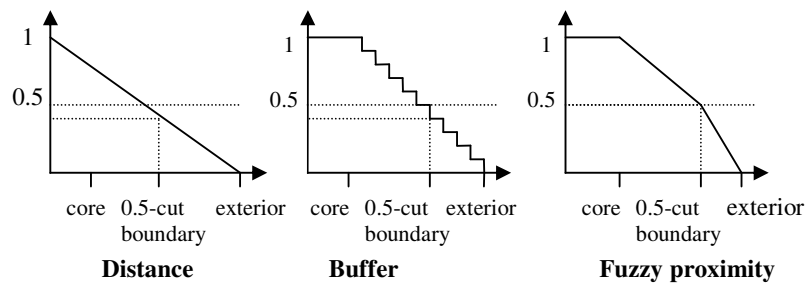


Fig. 4.14. Differences in Membership Grade between Three Proximity Measures

The grades of the first two measures are given by the Euclidean distance in a non-scalable way without regard for qualitative relations to neighboring regions. On the contrary, the grade of fuzzy proximity is determined by the Euclidean distance that has been accommodated to the size and qualitative relations to surrounding localities in the same resolution level.

4.6 Analysis of Results

4.6.1 Georeferencing Results

Table 4.3 is the cross-tabulation of georeferencing results by year and matching score range. 'DS' refers to instances geocoded by dynamic segmentation on the highway system. Score 1 means an exact match, whereas scores less than 1 indicate that a less than perfect match was achieved. The proportion of DS cases has noticeably increased since 1998, and so have overall similarity scores. This suggests that a matching procedure based on similarity outperforms an exact-match procedure when data to be georeferencing is of rather low quality.

Table 4.3. Cross-tabulation of Georeferenced Cases by Year and by Matching Score

Score_Class	1996	1997	1998	1999	2000	2001
DS	85	55	435	484	460	534
1	227	429	519	516	542	520
0.90 - 0.99	111	188	205	204	169	154
0.80 - 0.89	130	342	59	93	75	60
0.70 - 0.79	33	71	35	35	34	26
0.60 - 0.69	14	34	33	27	18	18
0.50 - 0.59	168	103	39	34	19	33
0.40 - 0.49	66	63	17	22	11	17
0.30 - 0.39	176	64	9	15	11	6
0.20 - 0.29	64	11	2	1	0	1
0.10 - 0.19	173	11	0	0	0	1
0.00 - 0.09	16	0	0	0	0	0
Unmatched	187	139	51	42	30	55
Total	1450	1510	1404	1473	1369	1425

4.6.2 The Evidence of Fuzzy Locality

A locality similarity score (from now we will call this the locality score for simplicity) is the fuzzy set membership value of locality, as discussed in Section 4.5. Therefore, we can examine the evidence of fuzzy locality by looking at locality scores less than 1. During the six-year period 1996-2001, 8631 fatal accidents occurred in the State of New York. Of this total, we now focus on 5460 cases to examine the contribution of the notion of fuzzy locality to the geocoding process. (The remaining 3171 cases are accounted for by DS georeferencing or are unmatched cases which do not use fuzzy locality modeling at all). In Figure 4.15, we distinguish georeferenced cases by reference data (either Place_PL or Place_PT) and by crispness of locality (either In or Near), for illustration purposes. We depict crisp cases (score = 1) on the left side of the figure, and fuzzy cases (score less than 1) on the right side. Fuzzy locality accounts for 12.4% (677/5460) of the total number of georeferenced cases. It shows there is fuzziness involved in identifying locality, but not too prominently. Anyhow, the match rate has increased from 86% (7450/8631) to 94% (8127/8631) thanks to fuzzy locality modeling.

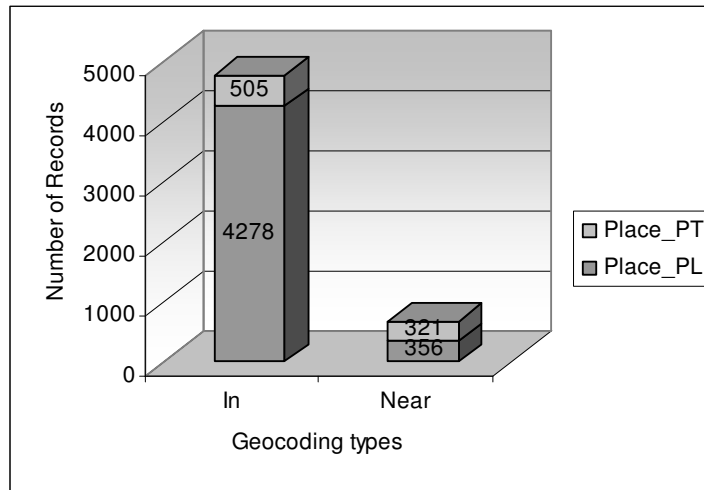


Fig. 4.15. Proportion of Fuzzy Locality Relative to Crisp Locality

We also examine how the locality scores are distributed in case of fuzzy locality. To control for the effect of poor data quality, we only consider cases whose matching quality is reasonably high (matching score is above 0.7). As expected, in the majority of instances, locality is perceived with little fuzziness and only a short distance from the target locality (Figure 4.16). Of the 329 fatal crashes for which fuzzy locality modeling enhances georeferencing, the location of the target local-

ity x is identified within the $FirstOrderHNGroup(x)$ 75% (247/329) of the time, and within the $SecondOrderHNGroup(x)$ 25% (82/329) of the time.

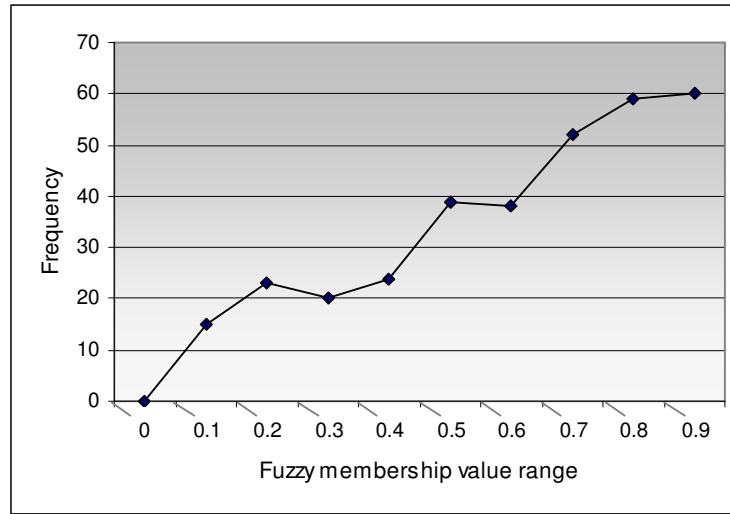


Fig. 4.16. Histogram of Locality Fuzzy Set Membership Values below 1 (Fatal Crashes with a Matching Score over 0.7)

4.6.3 The Effect of Locality Characteristics on Location Indeterminacy

We examine if the location determinacy is affected by certain characteristics of localities. For this purpose, we test two hypotheses. (1) How does the degree of urbanization affect the location determinacy of a locality? Is a big city more location determinate or not? (2) Is the location determinacy of a locality the outcome of human conceptualization? For instance, do people find it easier to identify the location of a well-established city?

The location indeterminacy value of each locality is computed as follows: First we compute the total number of accident cases (t) for each locality. Next we compute the number of fuzzy locality cases (f) for each locality as the number of accidents whose locality fuzzy membership value is less than one. Then divide f by t . For example, suppose that ten fatal traffic crashes occurred in Buffalo. Among them, suppose three cases turn out to be georeferenced outside of the legal boundary of the stated locality. In which case, the location indeterminacy value is 0.3. Thus, the location indeterminacy value, which is a characteristic specific to the locality, is the proportion of fuzzy locality cases to total cases. The higher the value

is, the more indeterminate its location is. Sixty-four percent of all localities (657/1031) have zero location indeterminacy. To test the stated hypotheses, we only consider the remaining 374 cases for which some indeterminacy is revealed by the fuzzy set model of locality.

For the localities corresponding to entities of the PLACE_PL data layer (211 cases), the location indeterminacy value is graphed against population in Figure 4.17. The scatter plot suggests that larger localities are more location determinate. People may find it easier to identify the location of urban areas than rural areas. It is likely that urbanization strengthens location determinacy whereas the location indeterminacy value in smaller and more recently settled areas follow more of a “hit or miss” pattern. Curve fitting to the data reveals a clear inverse and non-linear relationship ($R^2 = 42.71\%$). Table 4.4 lists the 15 largest localities, and their population as well as their location indeterminacy value.

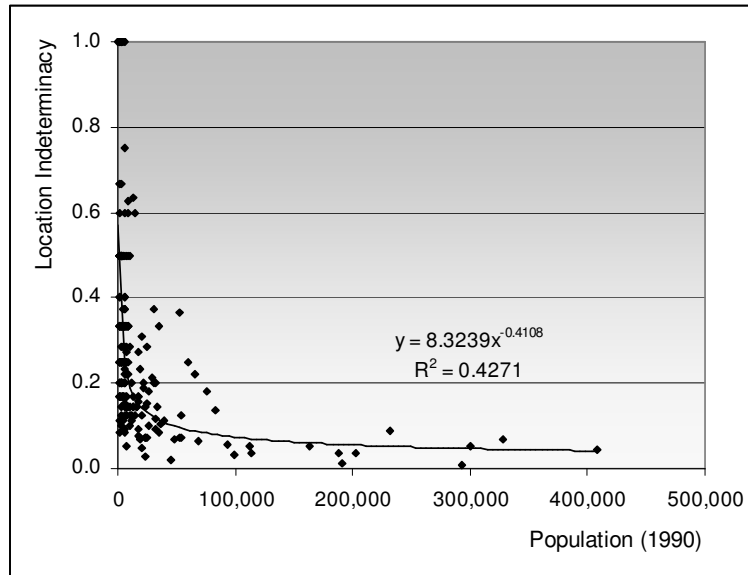


Fig. 4.17. Scatter Plot between Location Indeterminacy and Population

We need to be careful when interpreting the location indeterminacy index because location indeterminacy is measured on the basis of a very small number of accidents, possibly because a large proportion of recorded accidents were georeferenced by means of dynamic segmentation. Alternatively, localities with large population may yield a low indeterminacy index simply due to the large denominator, just as localities with small population may yield a high value due to small denominator. However, when we control for the number of accidents per locality,

the pattern remains pretty much the same. We can not produce an equivalent indeterminacy-population graph for localities that are matched against PLACE_PT because there is no consistent population statistic for these types of spatial entities.

Table 4.4. Location Indeterminacy Index for Selected Localities

NAME (PLACE_PL)	1990 Population	#TotalCases	#FuzzyCases	Location Indeterminacy
HEMPSTEAD	725639	283	8	0.028
BROOKHAVEN	407779	201	9	0.045
BUFFALO	328123	102	7	0.069
ISLIP	299587	137	7	0.051
OYSTER BAY	292657	106	1	0.009
ROCHESTER	231636	78	7	0.090
BABYLON	202889	81	3	0.037
HUNTINGTON	191474	93	1	0.011
YONKERS	188082	55	2	0.036
SYRACUSE	163860	37	2	0.054
SMITHTOWN	113406	55	2	0.036
AMHERST	111711	38	2	0.053
CHEEKTOWAGA	99314	29	1	0.034
RAMAPO	93861	18	1	0.056
TONAWANDA	82464	22	3	0.136

It may be insightful to provide maps showing the pattern of location indeterminacy index grouped by region. The 211 localities considered above are grouped into 11 regions that correspond to New York State Department of Transportation (NYSDOT) administrative boundaries. Location indeterminacy is averaged for each region, which is tabulated in the column labeled AVELOCINDET in Table 4.5. According to the map in Figure 4.18, regions 1, 6, and 9 show a higher degree of location indeterminacy compared to regions encompassing the New York metropolitan area. The map confirms the existence of a direct relationship between location determinacy and urbanization. The more urbanized the area, the more location-determinate the locality because urban areas provide more landmarks on which the judgment on location can be made. In addition, it is less likely that a location would be indeterminate when clearly defined bodies of water abound as on Long Island and the rest of the New York metropolitan area. Finally, metropolitan areas have been settled and developed for a longer period of time so that it is likely that a stronger sense of community identity has established than in other areas. The sense of community that has been formed for centuries may contribute to

the higher degree of location indeterminacy. In that sense, locality as a *fiat* spatial object seems to find some empirical validity here.

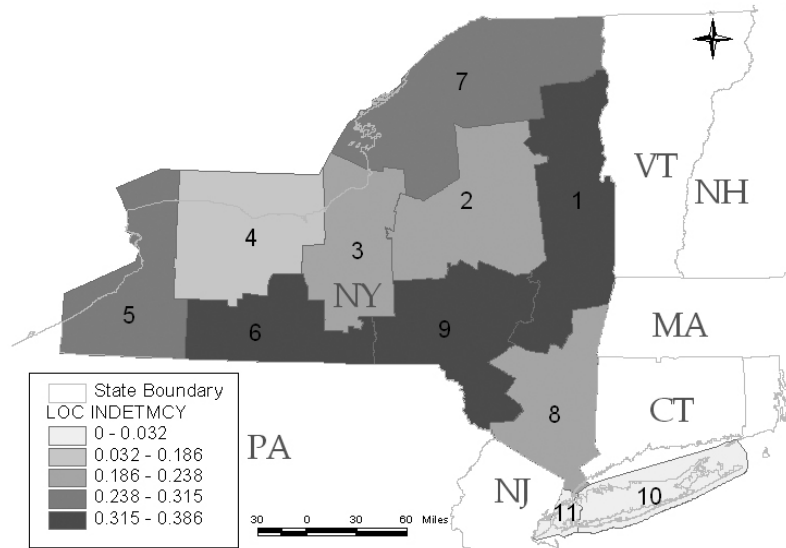


Fig. 4.18. Location Indeterminacy Index Grouped by NYSDOT Regions

Table 4.5. Average Location Indeterminacy Index Grouped by NYSDOT Regions and Number of Localities Included in Fuzzy Cases in Each Region

REGION	OFFICE SITE	AVELOCINDET	#LOCALITY
1	Albany	0.3860	84
2	Utica	0.2376	52
3	Syracuse	0.2192	67
4	Rochester	0.1856	82
5	Buffalo	0.2829	67
6	Hornell	0.3734	54
7	Watertown	0.3147	64
8	Poughkeepsie	0.2195	76
9	Binghamton	0.3576	67
10	Hauppauge	0.0316	14
11	New York	0.0000	0

4.7 Conclusions

Fuzzy set theory has been used to improve the georeferencing of traffic accident data. Georeferencing has been performed using a similarity measure due to imperfect information available in source datasets. To obtain the similarity measure of locality features, locality (e.g. city, town, and place) is modeled by fuzzy sets such that membership values can be set to the similarity measures. Thanks to this approach, the rate of success of the georeferencing procedure is enhanced from 86 % of all fatal accidents to 94 %.

Locality is seen as a fuzzy region and is represented as eggs in the *egg-yolk* representations (also known as *core-boundary* representation). The *Yolk* (or *core*) has a full membership of locality whereas the remaining parts (egg minus yolk or wide boundary) have partial membership. The vague parts are composed of a set of the first- and second-order surrounding localities at the same resolution level. Consequently neighborhood relation and scale determine the vague parts of locality.

To implement the concept of fuzzy locality in GIS, fuzzy locality layers are created from the GIS databases such as city/town boundary and place centroid. With a crisp conceptualization of locality, a single membership value would be assigned to each feature of the locality layers. In a fuzzy set approach, however, multiple membership values can be assigned to each feature of fuzzy locality layers. The fuzzy set membership value is set to 1 for the *core* part, and 0 for the *exterior* part of locality. The membership value is set to 0.5 on the edge of a set of the first-order surrounding localities (spatial lag of one). Then, the membership value is linearly interpolated between the core and the 0.5-cut boundary (so $[0.5, 1]$), and between the 0.5-cut boundary and the exterior (so $[0, 0.5]$). Therefore, the fuzzy set membership function of locality is defined as the combination of 4 different piecewise linear functions (0, $[0, 0.5]$, $[0.5, 1]$, 1) that depend on proximity with respect to neighboring localities. Uncertainty involved in determining the location of locality is thus captured in the vague parts through the notion of partial membership.

Fuzzy perception and cognition of localities accounts for 12.4% of all accident cases examined in this study. That is, people actually designated a locality outside of their legal boundary 12.4% of the time. Fuzziness in determining the location of locality turns out to be concentrated in areas close to the legal boundary. The magnitude of fuzziness is modeled to decrease linearly with proximity.

The contribution to the determination of locality can be explained at three different levels: (1) Ontology level: Some spatial objects may be inherently vague. The spatial extent of a locality can be indeterminate because this locality is the result of human conceptualization. (2) Perception level: There exist individual differences in perceiving the environment. Individuals have different mental maps (Gould and White 1986, Thill and Sui 1993) depending on their experiences, preferences, and knowledge. Maybe their decision to determine the location of a certain locality has relied on their error-prone mental maps rather than survey knowledge (Golledge et al. 1995). (3) Implementation level: Imperfect information such

as imprecise measurements, ambiguity, and incomplete value induces more fuzziness.

We attempted to address several questions with regard to the location determinacy of locality. Is the location determinacy inherent in locality? If so, can a certain characteristic of locality actually affect the location determinacy? Or is it just the result of rough measurements (e.g., poorly designed coding form)? The indicator of location indeterminacy is obtained from the proportion of fuzzy locality cases to the total cases in each locality. The relationship between population and location indeterminacy follows a negative exponential curve in general. The more urbanized the locality is, the more determinate its location is. This may be due to the fact that urban settings provide more landmarks on which perception and cognition of spaces and places can be made. This result actually fits the perception level hypothesis: mental maps are less erroneous in urban area than in rural area. Acquiring spatial knowledge in urban settings is quite different from rural settings. It can also be noted that natural environments such as bodies of water help determine the boundaries of localities, as evidenced by the low location indeterminacy index on Long Island and the rest of the New York metropolitan area. A sense of community can also play an important role in location determinacy even though it is hard to measure. Generally speaking, a historical city is more likely to satisfy identity condition with regard to location determinacy than other cities, not only because it has long been recognized as a distinct spatial entity, but also because it helps build a sense of community.

This study is significant by the wide range of domains of Geographic Information Science to which the methods and concepts discussed can be applied, including georeferencing, modeling fuzzy regions in GIS, and data quality control. Most importantly, the study attempts to formalize how to assign the fuzzy set membership value to a fuzzy region such as locality, which is often pointed out as a weakness in applying fuzzy set theory (Duckham 2001). Furthermore, the study supports the proper application of fuzzy set theory to spatial concepts, such as the indeterminacy of location. In sum, fuzzy set theory provides a mechanism to address various kinds of uncertainty by preserving the detail that would have been truncated in a crisp set.

This study has several limitations. We introduce a resolution level to make a fuzzy proximity measure scale-dependent, but the use of administrative boundaries, as a member of resolution level, does not necessarily represent the concrete existence of localities. Moreover, the assumption that a higher resolution level would exhibit a lower magnitude of fuzziness does not necessarily hold. The assumption turns out to exclude many appropriate parts of the locality entity. Furthermore, algorithms have not been tested for the cost that may trade off the effectiveness obtained from fuzzy locality modeling. When it comes to testing hypotheses, it was hard to distinguish between different forces that cause fuzziness in ontology, perception, and implementation level. They are treated as rather mixed. Finally, in the analysis of results, some cases presented may be prone to misinterpretation because of the small number problem.

Acknowledgements

This research was supported in part by the Center for Transportation Injury Research (CenTIR) in Buffalo, NY with funding from the Federal Highway Administration (FHWA) and technical oversight by the National Highway Traffic Safety Administration (NHTSA).

References

- Allen JF (1983) Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11): 832-843
- Bittner T (1999) On ontology and epistemology of rough location. In: Freksa C, Mark D (eds) *Spatial information theory: A theoretical basis for GIS (COSIT '99) Lecture Notes in Computer Science (LNCS) Vol 1661*, Springer, Hamburg, pp 433-448
- Bittner T (2000) A qualitative formalization of built environments. In: *Proceedings of DEXA2000, LNCS Vol 1873*, Springer-Verlag, Berlin-Heidelberg
- Burrough P, Frank AU (1996) (eds) *Geographic Objects with Indeterminate Boundaries. GISDATA Series II*. Taylor and Francis, London
- Duckham M, Mason K, Stell JG, Worboys MF (2001) A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems* 25: 89-103
- Egenhofer M, Franzosa R (1991) Point-set topological spatial relations. *International Journal of Geographical Information Systems* 5(2): 161-174
- Fisher P (2000) Sorites paradox and vague geographies. *Fuzzy Sets and Systems* 113: 7-18
- Erwig M, Schneider M (1997) Vague regions. In: *5th International Symposium on Advances in Spatial Databases (SSD'97)*, LNCS Vol 1262, Springer, Berlin, pp 298-320
- Gahegan M (1995) Proximity operators for qualitative spatial reasoning. In: Frank, AU, and Kuhn W (eds) *Spatial Information Theory: A Theoretical Basis for GIS. LNCS Vol 988*, Springer, Berlin, Germany, pp 31-44
- Golledge RG, Dougherty V, Bell S (1995) Acquiring spatial knowledge: survey versus route-based knowledge in unfamiliar environments. *Annals of the Association of American Geographers* 85(1): 34-158
- Goodchild M (2000) GIS and transportation. *Geoinformatica* 4(2): 127-140
- Gould P, White R (1986) *Mental Maps*, Harmondsworth, Penguin
- Guarino N, Welty C (2000) Ontological analysis of taxonomic relationships. In: Laender A, Storey V (eds) *Proceedings of ER-2000: The International Conference on Conceptual Modeling, Lecture Notes in Computer Science (LNCS) Vol 1920*, Springer-Verlag, Berlin, pp. 210-224
- Guesgen HW, Albrecht J (2000) Imprecise reasoning in geographic information systems. *Fuzzy Sets and Systems* 113: 121-131

- Hwang S, Thill J-C (2003) Georeferencing historical FARS accident data. A preliminary report. National Center for Geographic Information and Analysis, University at Buffalo, Amherst, NY
- Johnson-Laird PN, Girotto V, Legrenzi P (1998) Mental Models: A Gentle Guide for Outsiders <http://www.si.umich.edu/ICOS/gentleintro.html>
- Liu Z-Q, Satur R (1999) Contextual fuzzy cognitive map for decision support in geographic information systems. *IEEE Transactions on Fuzzy Systems* 7(5): 495-507
- NHTSA (1995) FARS 1996 Coding and Validation Manual. National Center for Statistics and Analysis, National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C.
- Robinson VB (1988) Some implications of fuzzy set theory applied to geographic databases. *Computers, Environment, and Urban Systems* 12: 89-97
- Robinson VB (2002) A perspective on geographic information systems and fuzzy sets. In: *Proceedings of North American Fuzzy Information Processing Society*, IEEE, New Orleans, LA, p 1-6
- Schneider M (1999) Uncertainty management for spatial data in databases: fuzzy spatial data types. In: Guting RH, Papadias D, Lochovsky F (eds) *SSD'99*, LNCS Vol 1651, Springer-Verlag, Heidelberg, pp 330-351
- Smith B (1995) On drawing lines on a map. In: Frank AU, Kuhn W (eds) *Spatial Information Theory: A Theoretical Basis for GIS*. LNCS Vol 988, Springer, Berlin, Germany, pp 475-484
- Smith B, Mark DM (1998) Ontology and geographic kinds. In: *Proceedings of international symposium on spatial data handling (SDH'98)*, Vancouver, Canada
- Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. *American Sociological Review* 5: 845-867
- Stefanakis E, Vazirgiannis M, Sellis T (1999) Incorporating fuzzy set methodologies in a DBMS repository for the application domain of GIS. *International Journal of Geographical Information Science* 13(7): 657-675
- Thill J-C, Sui DZ (1993) Mental maps and fuzzy preferences. *Professional Geographer* 45: 264-276
- Wang F, Hall GB (1996) Fuzzy representation of geographic boundaries in GIS. *International Journal of Geographical Information Systems* 10(5): 573-590
- Worboys MF (2001) Nearness relations in environmental space. *International Journal of Geographical Information Science* 15(7): 633-651
- Yao X, Thill J-C (2005) How far is too far? – A statistical approach to context-contingent proximity modeling. *Transactions in GIS* 9(2), forthcoming
- Zadeh LA (1965) Fuzzy sets. *Information and Control* 8: 338-353
- Zhan FB, Lin H (2003) Overlay of two simple polygons with indeterminate boundaries. *Transactions in GIS* 7(1): 67-81