Mapping Vernacular Geography: Web-based GIS Tools for Capturing “Fuzzy” or “Vague” Entities.

Abstract

Most people don’t use a formal geographical vocabulary, however they do use a wide variety of geographical terms on a daily basis. Identifiers such as “Downtown” are components of a vernacular geography which is vastly more used than the coordinates and scientifically defined variables beloved of most professional analysts. Terms like these build into the jointly-defined world-views within which we all act. Despite its importance for policy making and quality of life, attention is rarely paid to this vernacular geography because it is hard to capture and use. This paper presents tools for capturing this geography, an example of the tools’ use to define “High Crime” areas, and an initial discussion of the issues surrounding vernacular data. While the problems involved in analyzing such data are not to be underestimated, such a system aims to pull together professional and popular geographical understanding, to the advantage of both.

Keywords: Vague; Fuzzy; GIS; Vernacular; Geography

Dr Andrew Evans is a Senior Lecturer in the University of Leeds, UK. He is a member of the Centre for Computational Geography. His interests are in developing new technologies to facilitate democratic decision making, chiefly through the internet, and in the use of artificial intelligence in socio-economic modeling.

Mr Tim Waters is a researcher within the Policy and Research Unit of City of Bradford Metropolitan District Council, UK. He works on the crime prevention unit, mainly on issues of public engagement and crime reduction.

1.0 Introduction

Every day, billions of people exist in a vernacular geography very different from that captured by standard geographical techniques. Millions of us “go uptown for the evening” or “go to the shops on Saturday”, meaning particular geographical areas, but without a clear definition of where or what they are. We avoid “the

rough end of town" late at night or park away from "high crime areas" without clear definitions of what these terms mean geographically, despite their links with our behaviour. Such vernacular geographical terms are a good thing: the use of descriptors like "Downtown" or "the grim area down by the station" allows us to communicate geographical references that often include information on associated environmental, socio-economic, and architectural data. They place us within a network of socio-linguistic communities with shared understandings and, less fortunately, prejudices. These vernacular geographical terms are not simply indicative - they often represent psychogeographical areas in which we constrain our activities, and they convey to members of our immediate socio-linguistic community that this constraint should be added to their shared knowledge and acted upon. This private and shared vernacular geography influences billions of people every day, and yet, because of its difficult and subjective nature, it is hard to tie directly to objective data so we can use it to make scientific models or policy.

In this paper we first outline the nature of such vernacular areas and define the conditions under which they gain problematic characteristics – particularly with respect to 1) capturing them, and 2) using them with standard datasets. We then address these two topics in turn: we present a Geographical Information System (GIS) developed to capture such areas, and then discuss some of the methods and issues that are key to utilizing such data in combination with standard datasets. As an example data-capture exercise, and to show the potential power of such data when used in combination with a traditional dataset, we also present a use of the GIS system to capture and analyze "high crime areas" within the city of Leeds, Britain.

2.0 The Nature of Vernacular Areas

Vernacular areas were first identified as a strand of geographical conceptualization by Schwartzberg in 1967 (Zelinsky, 1980). Much of the subsequent work has focused on sub- and super-State scale regions in the US

(for a review, see Shortridge, 1987), though as the examples above show there is also a great richness to be found at the community and small-group level. Plainly such features are a subset of the more general division of cognitive spatial objects and therefore fit well within the current research trend that aims to explain and manipulate all geographical features as perceived entities given a semantic solidity and manipulated within an ontological framework (for examples and reviews, see Cross and Firat, 2000; Mennis, 2003; Agarwal, 2005). However, as the commonest public mode of geographical reference, vernacular features are worth their own specific attention. Such terms stand in for a complex and interesting set of variable values and they have considerable potential, particularly within informed policy making. It should be noted that we do not deal here with the other component of a vernacular geography: relative or absolute vernacular positioning. Issues around positioning terms such as “near” and “far” are discussed by Frank (1996), while a mechanism for capturing and quantifying them is outlined by Robinson (2000). Here we are solely interested in vernacular objects themselves.

The chief problem with such objects is that they are not usually discretely delimited. Conditions encouraging this include :

1. **Indifference:** It is often the case that people refer to geographical areas without reference to their boundaries (Hadzilacos’ (1996) “Don’t Care” boundaries). For example, take the term “Town Centre”. Common uses (“I’m going to the Town Centre”; “This Town Centre is crowded”) only include points that are definitely “Town Centre” or not – the boundary between these states isn’t of concern. Because of this, there are no specific criteria for a boundary in the common understanding of “Town Centre”. Such terms are the geographical equivalent of open forms in mathematics, though more difficult to capture because they are defined on a point-by-point basis. The natural, though incorrect, assumption when faced with

such a term is that the person using it has some idea of a boundary that they can elucidate upon if questioned. This is not the case, and to use the term as if it was associated with a boundary is a misapplication of the term. We might regard the famous *Sorites Paradox* (see Fisher, 2000) as exactly this kind of misapplication – or, indeed, a tool for spotting when it might arise. The reality is that to use such terms as if they include boundaries requires the terms to be scrapped and redefined anew with extents (for which see (6), below). The fact that finite geographical features may be unbounded rather than simply indeterminately bound is difficult for GIS researchers, who generally solve indeterminacy by assuming there is a crisp border somewhere waiting to be elucidated as a polygon, albeit that it may be impossible to do so (e.g. Clementini and Di Felice, 1996; Cohn and Gotts, 1996; Schneider, 1996; Bennett, 2001; Bittner and Stell, 2003). It is rare for formal definitions of transitional areas between spatial zones to explicitly state that the change is too subtle to include distinct boundaries (for two examples see the initial definitions of the mantle transition zone by Bullen,1940, and the Italian terms discussed by Ferrari, 1996) and these are not currently handled well in GIS. It should be noted that while the solution given in this paper alleviates this problem, it does not in any sense solve it, and we still await a data model that appropriately deals with explicitly unbound areas.

2. **Continuousness:** boundaries can be weakly delimited by people when there is a physical gradient between entities that need dividing, but there are no criteria determining where a boundary might exactly be placed. For example, how do we place a boundary around an area someone describes as “The Town Centre” when retail outlets and businesses are scattered across a city with varying density? This relates to (1) in that the redefinition of an unbound area is most difficult to resolve where the transition between entities is potentially continuous. Formally such an absence of boundary information has been used to justify a belief in *ontic vagueness*, that is, that vagueness is inherent in some objects (see Tye, 1990, for a

typical outline). However, the relationship between continuousness and (3) and (6) below is somewhat moot.

3. **Poor Precision:** people can resort to indistinct boundaries where there is an analytical proof or inductive suggestion of a discrete boundary, but our ability to locate it is limited by our measurement techniques or representational media. In more formal situations the alternative is usually to replace such boundaries with a simple and definite line at some scale of representation (Montello *et al.,* 2003, give the example of a line of latitude; see also, Worboys, 1998, for an alternative method of treatment). This is unfortunate, as these tend to be the boundaries where errors can most readily be quantified and then presented using a diffuse representation. Formally this is a type of *epistemological vagueness*.
4. **Multivariate classification:** people can prefer diffuse boundaries in the situation where a set of continuous or discrete variables that each have a different location in geographical space are binned together in variable space for descriptive convenience. For example, people may merge together socio-economic characteristics and then use this classification to delimit rough "areas" of a town, like a "Working-class area". In more formal situations this tends to lead to discrete boundaries in geographical space based on the average borders of the constituent variables: a classic example is soil types. Prototype-based classifications in which variables at a location are compared to a set of 'text-book' classes and the location binned in the closest class typically produce such distinct boundaries (see, for examples, Kronenfeld, 2003). Formally there is no problem of vagueness here, as a classification could include, for example, the precise category "things that match two criteria from a list of four". However, usually a less distinct boundary is preferable, hence their development in vernacular descriptions. The problems that arise with discrete geographical boundaries formed under this criterion are, again, due to a misapplied term: a term referring to a precise definition formed in one space (variable space) is used in another, inappropriate, space (geographical space). The better representation that a diffuse geographical boundary

can give is ancillary and a result of objects having more properties (in this case, spatial ones) than those they are classified on.

5. **Averaging:** Diffuse boundaries may develop where a more formal, discrete, boundary would be an average of time or scale-varying boundaries associated with a single entity. For example, contrast "the seaside" with the coastal line on a map, or consider what counts as a "hill" under different context scales (see review in Brimicombe, 1997). Cases where there is no strong definition of such boundaries (e.g. Fisher *et al.*, 2004) will, to an extent, additionally fall under (1) and (6).
6. **Definitional disagreement:** diffuse boundaries may arise where areas are represented by linguistic labels which have a different definition to different people. For example, an area felt by 20 people to be a "High Crime" zone may have a diffuse border across which less and less of the people feel their definition of "High Crime" applies. Such definitions are often ostensive (that is, they are defined by reference to examples) and thus this criterion also encompasses problems of relativity. These arise when groups try to compare terms defined using different subsets of all the possible examples: for instance, two people trying to delimit an area based on one's notion of "High Crime" (developed in a Norwegian city) and the other's notion of "High Crime" (developed in a British city). It may, of course, additionally occur that one person has conflicting definitions under different circumstances. Formally this condition is known as *semantic vagueness* under the circumstances where the vagueness comes from the interaction of multiple, possibly ill-defined, definitions. However, it may also be that someone is not able to articulate their definitions or we are not able to determine what all the definitions are. For example, we may have been told "most people think LA is a high crime area" but have no idea how "LA" or "high crime" are being used. Equally we may not be able to tell which out of a set of definitions should apply in a given situation or utilize them properly under our given circumstances. In these cases we are dealing with an *epistemological vagueness*. It should also be noted that this situation is generally different from *ambiguity* in which one term is

attached to two or more distinct and agreed definitions and we are not sure which is being applied (Sainsbury, 1995; Bennett, 2001); under our condition more than one definition is being applied and in addition it may be that the definitions cannot be distinctly and clearly given or used. Ambiguity is more clearly resolvable as there is a definite single definition to be elucidated. Plainly definitional disagreement problems rapidly take over during the resolution of problems associated with (1), as people will have differing ideas about the nature of a newly demanded boundary. Thus, to some extent, a solution to this problem represents a solution to (1). Differing approaches to such geo-semiotic problems are exemplified by Bennett (2001) and Fisher (2000).

(Alternative discussions of general indeterminately bound objects can be found in Leung, 1987; Couclelis, 1996; Fisher, 1996; Hadzilacos, 1996; Molenaar, 1996).

Imagined areas that are casually (rather than scientifically) constructed by human beings tend to fall within all of these conditions. When asked, for example, to outline and justify areas where they think crime levels are high, most people will draw on a slew of continuous and discrete variables at differing scales of detail, historical experiences, urban morphology and mythology, as well as introducing linguistic vagueness. The resultant areas may be bound by prominent landscape features, usually for convenience, but are more often diffuse or unconsidered, and the level at which an area is perceived to belong to a category like “High Crime” often drops off over some distance (for examples and analyses of a range of vernacular areas, see Zelinsky, 1980; Shortridge 1984; 1987; Ferrari, 1996; Campari, 1996).

A second issue with such vernacular entities is that they are internally varied. High crime areas, for example, often have zones of greater or lesser danger. Overlaid on this will be a variation in the familiarity with areas or confidence with which people assign an area to a term.

Generally then, the act of asking people to define areas which they commonly understand by ostensive definition and without recourse to their geographical boundaries is problematic and unlikely to result in a traditionally scientific discrete description. However, the rewards for meeting people half way, delimiting such areas without too much concentration on their boundaries, would appear to be great as it should allow greater comparison with traditional datasets. Treating such areas as having diffuse boundaries would seem to be such a half-way position. In the following sections we present a system based on this treatment, and give an example of its use to show the potential of the data generated.

3.0 Capturing Vernacular Areas

Diffuse boundaries have been utilized in GIS since the development of raster datasets. However, the manipulation of individual diffuse areas is rarer, and generally such areas are generated through the application of either mathematical relationships or expert rulesets to raster or vector data (see Robinson, 2003, for a review). Here we present a series of tools for elucidating the extent of diffuse areas directly. The tools are accessed through three interfaces that span the needs of GIS users, that is:

1. A user input interface: specifically, the user is given a spraycan tool, familiar from many image editing packages, with which they can define diffuse areas of varying density on a map (Figure 1A). Attribute information can then be attached to the area.
2. A querying interface: this allows individuals to pull up attribute information by querying a map generated from the aggregated and averaged responses of all users (Figure 1B). The user querying the system picks a point on the map and all the users' attributes for that point are displayed, ranked/ordered on the basis of how important the point was to each user's definition (a higher density of spraying for that point by a user results in a higher ranking for that user's attribute information).

3. An administrative interface: this features a tool for aggregating/averaging results from chosen multiple users and also displays their areas and attributes for individual-level analysis and editing.

POSITION OF FIGURE 1

The system is comprised of the interfaces and a set of processing and storage components. There are currently three versions of the system: a client-heavy version in which the interface and processing elements are written in Java and the storage components written in Perl; a server side Java version with a thin applet client; and an ESRI ArcMap stand-alone version, written in Visual Basic (not Internet-enabled). The results in this paper were derived from the first of these systems. In the Internet versions, the interfaces run as Applets with a relatively easy setup for those familiar with HTML and Perl scripts. The software, known as "Tagger", can be downloaded from http://www.ccg.leeds.ac.uk/democracy/

A typical area sprayed by a user can be seen in Figure 1A. The spraycan tool will be familiar to most people from elementary graphics packages, and takes little time to master if this is the first time the user has seen one. Graphics packages usually offer one of two types of spraycan (Figure 2). In the first, which we shall denote the "continuous-curve can" (Figure 2A), the whole area within the perimeter of the spray action is filled entirely, though subtly, at the first engagement of the tool. The longer the user holds the mouse button down, the more intense the level of spray becomes, weighted such that the most intense levels are towards the centre of the area. This results in a continuous density surface with a diffuse boundary asymptotic with the background colour from the image. The second type of spraycan (Figure 3B), which we shall denote the "dot-plane can", sprays a set of dots on the image within the perimeter of its action. The longer the user holds the mouse button down the more dots are randomly placed within the perimeter until the area is a solid block of spray of a single level.

POSITION OF FIGURE 2

Usability tests suggest that the dot-plane can gives users better control when delimiting geographical areas (Waters, 2002). Users are happier with their areas when they can spray more or less dots in an area, but spray well defined boundaries if they want to. This matches the notion that psychogeographical areas have diffuse boundaries, but that these can be combined with distinct boundaries at specific landmarks. With the continuous-can, it is a great deal harder to define a clear edge around an area. One might imagine two alternative intermediate spraycans: one in which dots are sprayed with a Gaussian-like distribution around the centre point, and another in which a continuous surface is sprayed without the curved distribution, however these are less familiar from image processing packages and the former seems to combine the worst aspects of the two standard algorithms.

A considerable problem in dealing with density surfaces of any sort over the web is the size of the data that results. While judicious and georeferenced convex-hull clipping can limit the size of the area stored, we are essentially dealing with raster data that includes a considerable amount of variation. The following procedure mitigates the problem:

1. The sprayed areas are separated from the background map.
2. The dot densities are converted into continuous density surfaces using a nine-by-nine averaging kernel (Figure 3A).
3. The images are reduced to a fifth of their size using averaged values (Figure 3B).
4. Standard lossless image compression generates images that can be opened in any standard graphics package. These, and associated attributes, are streamed to the server and stored.

5. The area is also added to a GZip compressed (Deutsch, 1996) data object on the server containing all areas and attributes. This is used for attribute querying and aggregate generation.
6. When the system is queried, the image processing is reversed (Figure 3C/D). The individual or aggregated images are expanded to their original size and inflation artifacts smoothed using a five-by-five kernel. To ease further processing they are left as a continuous density surface rather than re-represented as dots.

Tests suggest a typical compression rate is two orders of magnitude (a data object of 859Kb might typically compress to 67Kb using GZip, and to 14Kb with the addition of the shrinking process). User tests suggest that this compression maximizes the shrinkage and averaging the sprayed areas can be subjected to with the users staying happy that their views / areas are represented. Alternative data treatments based on interpolation (e.g. Laurini and Pariente, 1996) are unlikely to give better results as our criteria for the level of information stored is whether users feel they are being accurately represented or not, and the number of interpolation anchors is therefore likely to be high (for reviews of other field-storage methods see Haklay, 2004; McIntosh and Yuan, 2005).

POSITION OF FIGURE 3

4.0 Example Data Capture

The system was tested in 2002 as part of a study of where inhabitants thought were “High Crime” areas in the city of Leeds in Britain. Crime surveys frequently show that respondents have a higher fear of crime than is justified by actual crime figures (for a review see Hale, 1996). The fear of crime can have a significant impact on peoples’ lives, with 29% of respondents in the 2001/2002 British Crime Survey claiming they didn’t go out alone at night, and 7% of people going out less than once a month because of the fear of crime. 6% of respondents claimed fear of crime had a “great effect” on their quality of life with a further 31% saying it had

a “moderate effect” (Simmons *et al.*, 2002). Concern about crime therefore represents a significant influence on many people’s lives, and influences which areas people travel to at different times. Despite this, current models of fear struggle to predict it accurately (for example, Farrall *et al.*, 2000, who used demographic, psychological and temporal factors only accounted for approximately one third of the fear levels measured). In part, this is because the difficulty in capturing the spatial aspects of fear has led to most models being aspatial. Overall, areas of perceived crime risk are likely to be important to both individuals and policy-makers, and provide a suitable test-case with which to investigate the capture and use of vernacular areas.

Leeds is a city of some 715,402 people residing within 562km^2, with 118,559 reported crimes in 2001/2002 (Kongmuang, 2006). Most of these crimes are concentrated in the city centre (Figure 4A), however this does not necessarily equate with the areas people feel are “High Crime”. To find out where these areas actually are, a web-based GIS was set up using the above tools, and a pilot group of people who lived or worked in the city were asked to spray those areas that they thought were “High Crime”, spraying more in areas that they felt were of highest crime (for example, see Figure 1A). They did not have to have lived in the area, and could spray on the basis of hearsay, media attention, personal experience or any other evidence they cared to bring to bear on the problem.

In addition to defining areas, users could attach comments to the areas. Once they had submitted their areas, users could view a composite map combining all the areas perceived by the community of users, and view people’s comments associated with specific locations (Figure 1B). Plainly in a more widely publicized project such comments would be moderated, though monitoring did not reveal that this was necessary in this particular study.

As the pilot users were gained by advertising the system within the University of Leeds, the group's demographics or knowledge of the city will not necessarily be that of the general populous. Given the general correlation between University workers and broadsheet readership it seems, on the basis of the 2001/2002 British Crime Survey, that the risk overestimation of this group may be somewhat less than the general population (42% of Tabloid readers felt crime rates had increased between 1999 and 2001/2002, compared with 26% of Broadsheet readers, against an actual fall of 14%: Simmons *et al*., 2002). Given this, the results are indicative, but a generalized commentary cannot be derived from them - nor is this the purpose here.

Figure 4B shows those locations the participating community believe have the highest levels of crime. At each pixel the map shows the intensity levels for all users summed and divided by the number of users. That is, the average areas regarded as being "High Crime" zones. The colour scale has been stretched so that white areas are those that were not sprayed and the areas that were sprayed the most are black.

5.0 Example Data Utilization

Having outlined the problematic nature of vernacular geography, and given one method by which it might be collected, it is perhaps worth saying something about how one might utilize it. The analysis of such data is far from simple (see below), however, as a basic example to show the potential for such data to be used in decision making, the data generated by the study above can be compared with absolute crime levels. The same compression and colour rescaling treatment has been applied to the real data as to the user-sprayed areas. Figure 4 shows a simple subtraction of the two datasets which could be used to show broad, quasi-quantitative, estimates of the differences between perceived and actual crime levels for use in a decision making process.

POSITION OF FIGURE 4

This simple processing suggests some of the potential power of vernacular data. We can use comparisons like this to quasi-quantitatively answer questions such as: “where do people have mis-perceptions as to the risk from crime?” or, had the areas matched, “what level of crime do people notice as ‘High’ ?”. However, additional data could further enhance this analysis: spraying the areas users are familiar with might allow us to test whether their knowledge was representative of the broader population, or to normalize areas by level of familiarity. The fear maps could be compared with other human-centred data, like the Galvanic Skin Response maps of Nold (2004). Other information (e.g. demographics of users and where people have lived or experienced crime) would allow us to disaggregate the data, while the comments provided might also reveal the processes they used to determine the areas. Work in these directions is underway (e.g. Cressy, 2004) however it is still plain that this kind of quasi-quantitative approach is yet a long way from a full comparison with scientific data.

As noted above, users were allowed to see the aggregated map and query it for other users’ comments. This allowed the users themselves to gain from reflections such as: “how scared of crime are my neighbours” and “does anyone else feel the same way as me”. While there are obvious ethical, socio-economic and (potentially) libel and policing-related difficulties in presenting such data back to users, both in map form and as comments, in this pilot study it was felt appropriate to gauge use of the system. Initial analysis of feedback suggests the system (both input and querying) was well received by users (Figure 5).

POSITION OF FIGURE 5

6.0 Discussion: Utilizing Vernacular Geography

6.1 Methodologies

A fuller discussion of methods for utilizing such data will be presented elsewhere, however having detailed the problems associated with collecting such data, and provided one possible solution, it is perhaps worth at least outlining some of the methods by which more quantitative analysis could proceed. Research in "Fuzzy" or "Vague" cognitive geographical objects offer at least four methodologies of some promise.

1. **Fuzzy Sets/Logic:** A growing body of work uses fuzzy set theory to define areas where boundaries are gradients (for reviews, see Jacquez *et al.*, 2000; Robinson, 2003). Under this model each point in a perceived area (at some resolution) would have a membership-level for one or more vernacular terms: a coordinate might be 80% "High Crime", for example. The use is clearest when there are contrasting classes, even if these are implicit: 80% "High Crime" also suggests a 20% membership of a "Low Crime" class. Fuzzy Logic then allows us to build and quantify policy statements like "*if* CRIME *is* HIGH, INVEST = MORE". Complaints against Fuzzy methods include the somewhat arbitrary nature of Fuzzy Logic and that it does not cope well with multiple, non-conflicting classes; contrast, for example "High crime areas based on media reports" and "High crime areas based on personal experience". Generally, however, Fuzzy Logic's ability to embracing and aggregate data of different definitions is usually seen as its strength. Fuzzy methods are often (unfairly) taken to imply an ontic vagueness which some theorists object to (c.f. (4), below).
2. **Statistical and probabilistic approaches:** If we represent perceived areas as surfaces across which a membership level varies we can make simple statistical comparisons with other data surfaces. The great advantage of standard statistical techniques like regression is that they compare datasets that we might imagine incomparable ("Rainfall" and "Air Pressure", for example). Entropy / Confusion measures allow

us to assess the rationality of comparing a subjective classification with real data (e.g. Brown, 1998; Kronenfeld, 2003) while links between the datasets and evidence/action can be built through Bayesian reasoning.

3. **Mereotopological calculi:** Mereological Algebras, those that deal with parts and wholes, have developed to cope with three-part logics – that is, logical problems dealing with true, false and indeterminate questions. These have been extended to cope with mereotopological geographical entities (e.g., Clementini and Di Felice, 1996; Cohn and Gotts, 1996; Cohn *et al.*, 1997; Casati *et al.*, 1998) and positioned within Rough Set theory by Bittner and Stell (2003). The classic representation used is the "fried egg" model, in which a given geographical entity has an inner core which definitely matches a description (the yolk), a surrounding region in which there is no certainty this description applies (the white), and an external area where it definitely doesn't apply (the pan). Representation is usually using vector polygons, and while the more sophisticated calculi can deal with multiple regions of certainty and suggest methods should be available for dealing with higher order uncertainty i.e. in the location of the polygon lines, there are still many real problems in these areas. Most mereotopology work centres on the spatial relationships between such entities ("do they overlap?" "is one inside another?") rather than their comparison with other datasets, however, they are the basis of much of the work in Supervaluation Semantics (below) and can be used in Fuzzy Logic-like decision making (e.g. Kulik, 2003).
4. **Supervaluation semantics:** It is still a moot point as to whether all vagueness in world can be attributed to semantic vagueness – that is, because we can only use words to understand the world, all problems of vagueness are due to the vagueness of the terms we use. It would certainly seem the case in many situations. Supervaluation theory holds that multiple people hold multiple spatial and aspatial definitions of an entity, and the overlapping of these multiplicities generates semantic vagueness (e.g. Bennett, 2001). Although it isn't a necessary component of the theory, it is usually assumed these descriptions are clearly

spatially defined, that is, for example, a single person asked for the area they consider to be “Downtown” could give a well defined and distinct boundary. Given this, multiple definitions of an area can be overlapped to construct a mereotopological entity in which the yolk is areas everyone agrees are “Downtown”, the white is where there is some agreement, and the pan areas where everyone agrees no definition holds (e.g. Montello *et al.*, 2003). As such, supervaluation attempts to extend classical logic to vague entities – it allows one to say where something is, is not, or is definitely impossible to talk about consistently (c.f. Fuzzy Logic where an entity matches a description to some degree). Thus, supervaluation theory is clearly centred on the comparability of contrasting definitions (Bennett, 2001).

In addition, a set of less well developed techniques might be applied to such entities. For example, entities could be characterize as a set of Beliefs, avoiding some of the problems of not having clear definitions, and allowing analysis using a geographically enhanced calculi based on Doxastic Logic (i.e. one based on belief) or Dempster-Shafer (Evidence) Theory. The latter relates the proportion of evidence backing up a claim to Belief and Plausibility (Rocha, 1999; Comber *et al.*, 2004). Techniques tying Beliefs and Behaviour together would avoid many of the problems of understanding the individual drivers behind these two issues. It is not the intention of this paper to look in detail at these possibilities, however, it should be clear that there are a range of techniques applicable to such data.

5.3 Problems

Whichever technique is used, there are a variety of data-centred problems that would need addressing before we might rely on such data. Briefly, these are twofold:

1) Fitness for purpose: both with reference to an individual (can data collected from a person for one purpose be used for another?) and the group (are two people’s data comparable?). As with standard datasets the appropriateness of utilization will be controlled by the level of generalization/specificity in

both the datasets, the uses in question, and the constraints applied. For example, we might envisage a continuum between "the area you mean when you say you are '"*going to the shops*' " and "locations you go to shop, sprayed with an intensity directly proportional to the number of times you think you go in a week". It seems clear that as the definition becomes more detailed the specificity of the problems we can deal with increases. In addition, it may be that we can move from quasi-qualitative analyses ("this comparison uses areas people produced on the basis of these other questions, and we believe the answers are co-applicable here") to more quantitative comparability (cp. Guarino, 1998's *reference* and *sharable* ontologies). For an extensive discussion of the possibilities here see Freksa and Barkowsky (1996) on a GIS to collect data specifically appropriate for a particular concept-based question area.

2) Precision and Accuracy: We need to quantify the precision during data collection and the appropriateness of our data as we move into higher or lower granularity fields. This is little different from standard datasets (see Worboys, 1998, for methods for coping with resolution changes in both standard and vague data). More problematically, how can we quantify the accuracy and confidence we have in our data? In standard data the precision of the collection instrument gives us accuracy metrics that can be used to build a confidence metric or even a measure of the plausibility that such data is useful in inference (see, for suggestions, Wilson, 2004). The same is certainly true of perceived areas, however, as the instrument used in delimiting the area includes the mind of the perceiver, we may also need to collect from them the confidence they have in their areas, either as a statistic for the whole area, or individual locations. The fact that relatively mature methodologies such as Multi-Criteria Evaluation, expert-driven analysis, and the generation of fuzzy linguistic areas from standard datasets all suffer from the same problem gives us some hope.

It is clear that the problems above relate to issues implicit in formal datasets, however our experience in dealing with them in this new context is limited to the point that we are largely willing to accept our paralysis rather than attempt what would undoubtedly be an initially naïve set of analyses. The considerable advances in solving these issues in traditional datasets should give us more confidence in dealing with perceived areas. Despite this, even in combination with the techniques from Fuzzy/Vague research noted above, it is far from clear that the sum total of problems is resolvable to produce a fully quantitative science. For this we would ideally construct a single algebra for manipulating perceived areas, possibly based on the notions of Belief, Plausibility and Action mentioned above. The degree to which this is possible or practically useful is unclear. Quasi-quantitative analyses, such as the "High Crime" example above, suggest it is a problem at least worth investigating. It should also be noted that such analyses, while problematic, are a considerable advance in representation on the aspatial aggregated statistics on opinions currently used in policy making.

6 Conclusions

We all live in a geographical world, even those without the excellent fortune to be professional geographers. This may seem obvious, but how often do we take this fact on board when describing this world? There are some obvious reasons why, as professional researchers, we work on the datasets we do: they are relatively simple to collect, have nice clear qualities, and we have reached a mutual agreement with decision makers that they are important. However, we tend to ignore the larger point: that this is not how almost all of the billions of people on the earth experience, utilize, and are driven by geography on a day to day basis. Standard geographical datasets, with their crisp boundaries, standardized metrics and precise definitions are plainly useful – astonishingly so, in fact, when you consider how few people even know of their existence, but to gain a real insight into humanity's use, understanding, and interaction with this world, we need to see it as the majority of human beings see it: not thin and anemic, but rich and inventive.

With this mind, this paper details some of the problems associated with collecting vernacular geography, presents a system that overcomes some of these problems, and discusses how such data may be analysed across a range of quantitative depths. An example outlining “High Crime” areas has been given. While attribute information was collected for this (albeit very generally: the way people felt about an area), there is no reason why the system should not be used simply to delimit one type of area (“where is your community?”; “what areas do you know most about?”). Equally, such attribute information can provide the input into a more formal cognitive/semantic/ontological GIS (for reviews see Mennis, 2003; Cross and Firat, 2000; Agarwal, 2005).

The drawing together of professional and popular understanding of the world is to their mutual advantage. For the professional, capturing the popular worldview should bring enhanced appreciation of the driving forces behind people’s spatial actions, and allow us to better match policy to needs. For the population at large it will allow them to use their own voice to communicate the things of importance to them, rather than the ‘Latin tongue’ of the professionals. As such, the capturing of vernacular geography promises to enhance both understanding and democratic policy making, and to give a louder voice to a geography that is both rich and significant.

Acknowledgements

The suggestion that data could be assessed against familiarity surfaces was made by Dave Martin at GISRUK 2003. Many thanks to him and the many others at that conference for their useful comments.

References

AGARWAL, P., 2005, Ontological considerations in GIScience. *International Journal of Geographical Information Science*, **19**, 5, 501-536.

BENNETT, B., 2001, Application of Supervaluation Semantics to Vaguely Defined Spatial Concepts. In *Spatial Information Theory: Foundations of Geographic Information Science; Proceedings of COSIT'01*, D.R. Montello (Ed.), LNCS, **2205** , 108–123, (Morro Bay: Springer).

BITTNER, T. and STELL, J.G., 2003, Stratified Rough Set and Vagueness , In W. KUHN, M. WORBOYS, and S. TIMPF, (Eds), Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. International Conference COSIT'03, (Springer), pp.286–303.

BULLEN, K.E., 1940, The problem of the Earth's density variation, *Seismological Society of America Bulletin*, **30**, 235–250.

BRIMICOMBE, A., 1997, A universal translator of linguistic hedges for the handling of uncertainty and fitness-for-use in GIS. *Innovations in GIS 4*, Z. Kemp (Ed.) (London: Taylor and Francis), pp. 115–126.

BROWN, D.G., 1998, Classification and Boundary Vagueness in Mapping Presettlement Forest Types. *International Journal of Geographical Information Science*, **12**, 2, 105-129.

CAMPARI, I., 1996, Uncertain boundaries in Urban Space. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 57–69.

CASATI, R., SMITH, B., and VARZI, A., 1998, Ontological Tools for Geographic Representation. In *Formal Ontology in Information Systems*, N. Guarino (Ed.) (Amsterdam: IOS Press) 77–85.

CLEMENTINI, E. and DI FELICE, P. 1996, An Algebraic Model for Spatial Objects with Indeterminate Boundaries. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 155–169.

COHN, A.G. and GOTTS, N.M., 1996, The 'Egg–Yolk' Representation of Regions with Indeterminate Boundaries. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 171–188.

COHN, A.G., BENNETT, B., GOODAY, J., GOTTS, N.M., 1997, Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, **1**, 275–316.

COMBER, A.J., LAW, A.N.R. and LISHMAN, J.R., 2004, A comparison of Bayes', Dempster–Shafer and Endorsement theories for managing knowledge uncertainty in the context of land cover monitoring. *Computers, Environment and Urban Systems*, **28**, 4, 311–327.

COUCLELIS, H., 1996, A Typology of Geographic Entities with Ill-defined boundaries. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 41–55.

CRESSY, K.M., 2004, Investigation Of Geographical Perceptions Of Crime Using A Public Participation GIS. MSc thesis, City University, London. pp.128.

CROSS, V. and FIRAT, A., 2000, Fuzzy Objects for Geographical Information Systems. *Fuzzy Sets and Systems*, **113**, 19–36.

DEUTSCH, P., 1996, GZIP file format specification version 4.3. Request for Comments 1952 (Network Working Group). Available online at: www.isi.edu/in-notes/rfc1952.txt (accessed 9 July 2003).

FARRALL, S., BANNISTER, J., DITTON, J., and GILSCHRIST, E., 2000, Social Psychology and the Fear of Crime: Re-examining a Speculative Model. *British Journal of Criminology*, **40**, 399–413.

FERRARI, G., 1996, Boundaries, Concepts, Language. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 99–108.

FISHER, P., 1996, Boolean and Fuzzy Regions. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 29–40.

FISHER, P., 2000, Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, **113**, 1,7–18.

FISHER, P., WOOD, J., and CHENG, T., 2004, Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, **29**, 1, 106–128.

FRANK, A.U., 1996, The Prevalence of Objects with Sharp Boundaries in GIS. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 29–40.

FREKSA, C. and BARKOWSKY, T. 1996, Relations between Spatial Concepts and Geographic Objects. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 109–121.

GUARINO, N., 1998, Formal Ontology in Information Systems In *Formal Ontology in Information Systems*, N. Guarino (Ed.) (Amsterdam: IOS Press) 3-15.

HADZILACOS, T., 1996, On Layer-based Systems for Undetermined Boundaries. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 237–255.

HAKLAY, M., 2004, Map Calculus in GIS: a proposal and demonstration. *International Journal of Geographical Information Science*, **18**, 2, 107-125.

HALE, C., 1996, Fear of crime: A review of the literature. *International Review of Victimology*, **4**, 2, 79–150.

JACQUEZ, G.M., MARUCA, S., and FORTIN, M.-J., 2000, From fields to objects: A review of geographic boundary analysis. *Journal of Geographical Systems*, **2**, 3, 221–241.

KULIK, L., 2003, Spatial vagueness and second-order vagueness. *Spatial Cognition and Computation*, **3**, 2&3, 157–183.

KONGMUANG, C., 2006, Modelling Crime: a spatial microsimulation approach. PhD thesis, University of Leeds.

KRONENFELD, B.J., 2003, Implications of a data reduction framework to assignment of fuzzy membership values in continuous classes. *Spatial Cognition and Computation*, **3**, B. Bennett and M. Cristani (Eds.) Special Issue on "*Spatial vagueness, uncertainty, granularity*", 221–237.

LAURINI, R., and PARIENTE, D., 1996, Towards a Field-oriented Language: First Specifications. In *Geographic Objects with Indeterminate Boundaries* Burrough and Frank (Eds.) (Taylor and Francis), pp. 225–236.

LEUNG, Y., 1987, On the Imprecision of boundaries. *Geographical Analysis*, **19**, 2, 125–151.

MCINTOSH, J. and YUAN, M., 2005, A framework to enhance semantic flexibility for analysis of distributed phenomena. *International Journal of Geographical Information Science*, **19**, 10, 999–1018.

MENNIS, J.L., 2003, Derivation and implementation of a semantic GIS data model informed by principles of cognition. *Computers, Environment and Urban Systems*, **27**, 5, 455–479.

MOLENAAR, M., 1996, A Syntactic Approach for Handling the Semantics of Fuzzy Spatial Objects. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 207–224.

MONTELLO, D.R., GOODCHILD, M.F., GOTTSEGEN, J., and FOHL, P., 2003, Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, **3**, B. Bennett and M. Cristani (Eds.) Special Issue on "*Spatial vagueness, uncertainty, granularity*", 185–204.

NOLD, C., 2004, BioMapping project. Available online at: www.biomapping.net (accessed 9 July 2003).

ROBINSON, V.B., 2000, Individual and Multipersonal Fuzzy Spatial Relations acquired using Human-Machine Interaction. *Fuzzy Sets and Systems*, **113**, 133–145.

ROBINSON, V.B., 2003, IA perspective on the Fundamentals of Fuzzy Sets and their Use in Geographic Information Systems. *Transactions in GIS*, **7**, 3, 3–30.

ROCHA, L.M., 1999, Evidence sets : modelling subjective categories. *International Journal of General Systems*, **27**, 457–494.

SAINSBURY, R.M., 1995, *Paradoxes*. (Cambridge: Cambridge University Press) pp.165.

SHORTRIDGE, J.R., 1984, The Emergence of the “Middle West” as an American Regional Label. *Annals of the Association of American Geographers*, **74**, 2, 209–220.

SHORTRIDGE, J.R., 1987, Changing Usage of Four American Regional Labels. *Annals of the Association of American Geographers*, **77**, 3, 325–336.

SIMMONS, J., and Colleagues, 2002, *Crime in England and Wales 2001/2002*. (London: HMSO). Available online at: www.homeoffice.gov.uk/rds/pdfs2/hosb702.pdf (accessed 9 July 2003).

SCHNEIDER, C., 1996, Modelling Spatial Objects with Indeterminate Boundaries using the Realm/ROSE Approach. *In Geographic Objects with Indeterminate Boundaries*, P.A.BURROUGH and A.U.FRANK (Eds.) (London: Taylor and Francis) 141–152.

TYE, M., 1990, Vague Objects. *Mind, New Series*, **99**, 396, 535-557.

WATERS, T., 2002, A Java Public Participation GIS Using a Spray Can Tool for an Investigation on the Perception of Crime in Leeds. MSc thesis, University of Leeds. Available online at: www.ccg.leeds.ac.uk/software/tagger/docs/Waters2002.pdf (accessed 9 July 2003)

WILSON, N., 2004, The beginnings of a logical semantics framework for the integration of thematic map data. *International Journal of Geographical Information Science*, **18**, 4, 389–415.

WORBOYS, M., 1998, Imprecision in Finite Resolution Spatial Data. *GeoInformatica*, **2**, 3, 257–279.

ZELINSKY, W., 1980, North America's Vernacular Regions. *Annals of the Association of American Geographers*, **70**, 1, 1–16.

FIGURES

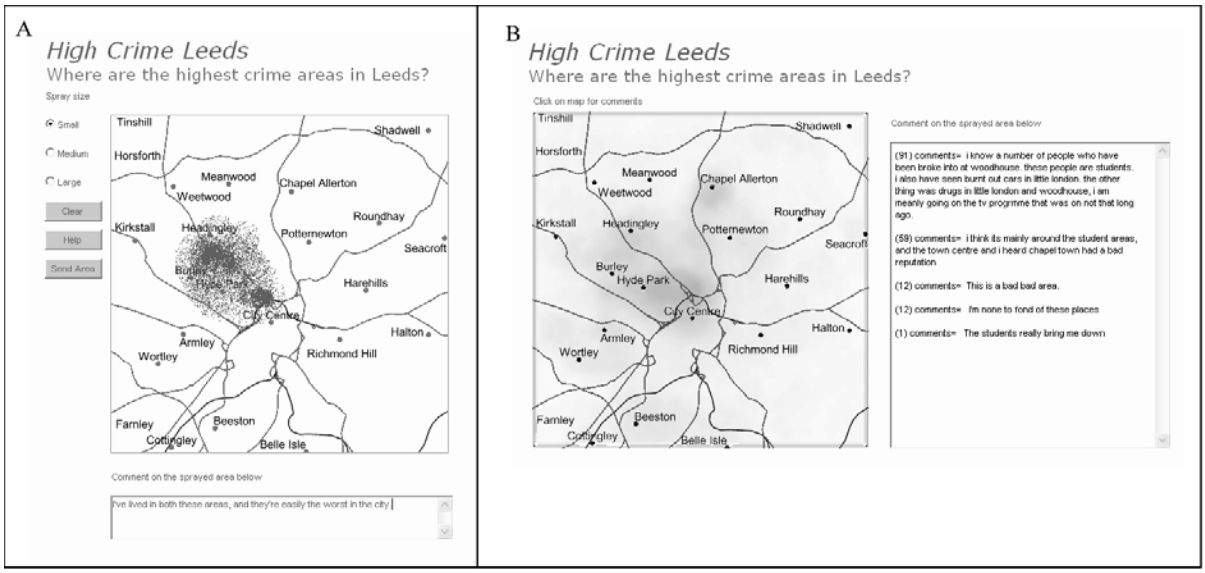

Figure 1. A: a user inputted area of perceived “High crime”. In this case the attribute data attached is simple comments about why specific areas were chosen by the users. B: Output showing all user areas averaged and ranked comments for one area.

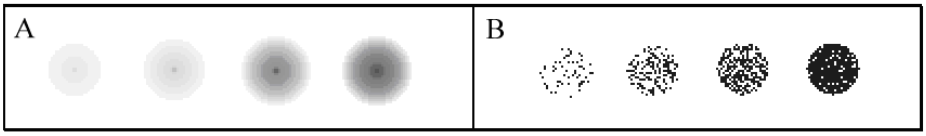

Figure 2. Two spraycan examples. A: continuous-curve can example from Corel® PaintShop Pro®. B: dot-plane can example from Microsoft® Paint®.

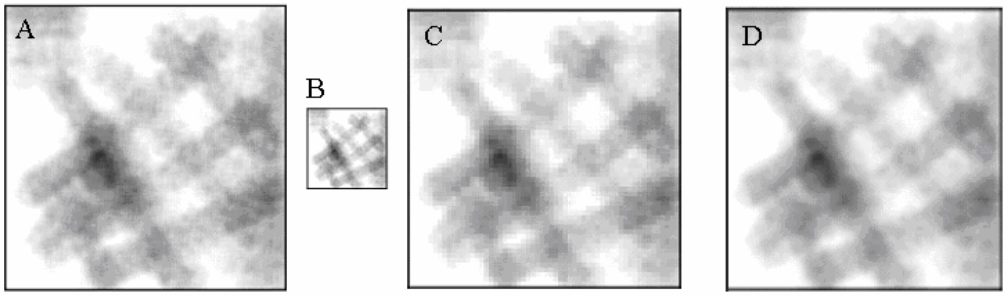

Figure 3. Compressed and inflated figures. Processing proceeds from left to right. A: density surface generated by kernel-averaging sprayed dots. B: shrunk image. C: inflated image. D: kernel-averaged inflated image.

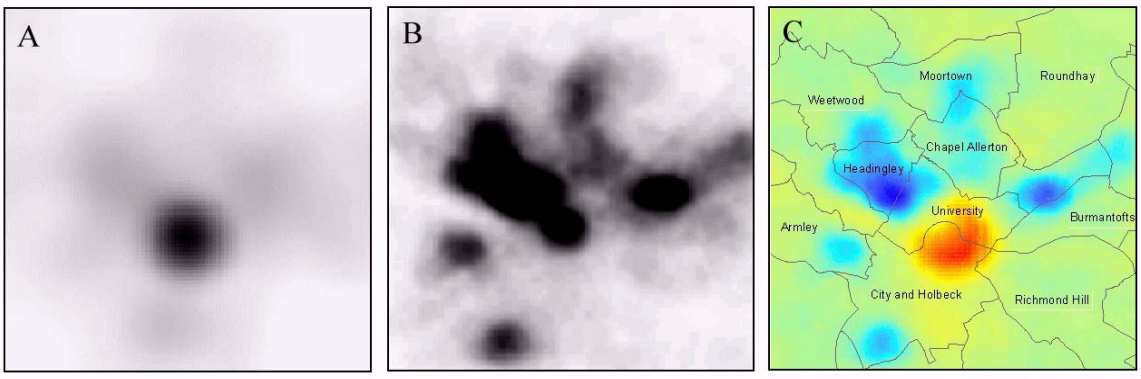

Figure 4. A: Total crime densities for Leeds for all crimes recorded in 2002. Darker areas are higher in crimes (see text for compression and colour scaling details). The circular high is real and largely reflects the position of the inner ring-road around the city. B: Areas selected as “high crime” areas by users cumulated from August to September 2002. Darker areas are thought higher in crime. C: Difference in perceived and real crimes, generated after stretching the highest perceived crime area levels to the highest real crime levels and the lowest perceived crime levels to the lowest crime levels. Red areas may have higher crime than expected, blue areas lower. UK Census Wards are shown for reference (© Crown Copyright/database right 2007. An Ordnance Survey/EDINA supplied service).

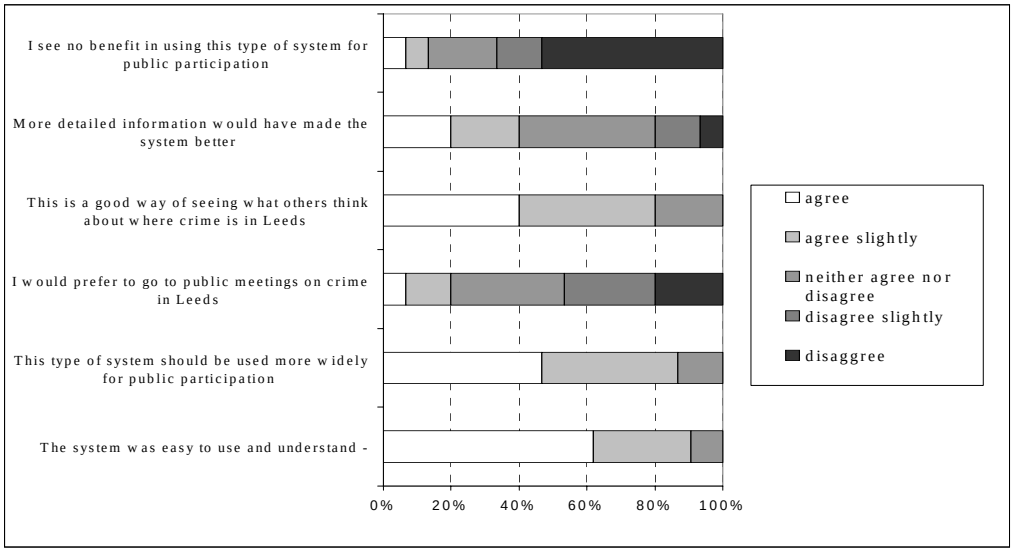

Figure 5: User feedback garnered by questionnaire at the end of system use.