Tools for web-based GIS mapping of a “fuzzy” vernacular geography

T. Waters and A.J. Evans

Centre for Computational Geography, School of Geography,
University of Leeds,
Leeds LS2 9JT, UK.
Telephone: +44 (0) 113 34 33342
FAX: +44 (0) 113 34 33308
Email: a.evans@geog.leeds.ac.uk

Abstract

The vast majority of people don’t use a scientific geographical vocabulary, nevertheless most use a wide variety of geographical terms on a day to day basis. Identifiers like “Downtown” or “The grim area around the docks” are part of a vernacular geographical terminology which is vastly more used than the coordinate systems and scientifically defined variables so beloved of professional geographers. These terms not only identify areas, but give members of our socio-linguistic group information about them, building up a jointly-defined cultural world-view within which we all act on a daily basis. Despite its importance for policy making and quality of life, attention is rarely paid to this vernacular geography because it is so hard to capture and use. This paper presents a new set of tools for capturing these “fuzzy” psychogeographical areas and their associated attributes, through a web based mapping system. The system contains a spraycan tool that allows users to tag information onto diffuse areas of varying density. An example of their use to define areas people consider are “high crime” within a UK city is also presented, along with users’ responses to the system. Such a system aims to pull together professional and popular geographical understanding, to the advantage of both.

1. 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 down 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 metaphors like “the East End” 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, and they place us in a connected socio-linguistic community 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 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 decisions or policy.

There has been a growing body of work in the last few years using fuzzy logic to define ambiguous geographical data (for a review, see Jacquez *et al.*, 2000). Ambiguous or fuzzy geographical boundaries can be used between areas when one or more of the following criteria are present…

1. Continuousness: when boundaries are difficult to define because the measurements of an entity produce a gradient (the start of a mountain, for example).

2. Aggregation in the categorization of variables: where discrete boundaries actually represent the average location of a geographically varying set of continuous or discrete variables that are binned together for descriptive convenience (soil types, for example).

3. Averaging: where discrete boundaries are actually an average of time or scale varying geographical boundaries (the definition of a col, the boundary of a river, etc.).

4. Ambiguity: where boundaries are tied to linguistic factors (for example, “high” crime areas).

Imagined areas that are casually (rather than scientifically) constructed by human beings tend to display all four of these criteria. 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 ambiguities. The resultant areas may be bound by prominent landscape features, usually for convenience, but are more often diffuse. The level at which an area is perceived to belong to a category like “high crime” often drops off over some distance, and the actual areas themselves internally often have more or less “high crime” zones.

2 Recording and Utilizing Vernacular Geographical Entities

2.1 Introduction to the tools

Here we present a Geographical Information System (GIS) aimed at recording and manipulating such fuzzy vernacular or perceived areas. Its use centres around three interfaces, each with a separate tool for dealing with the data. As an example of a data collection exercise using the system we also present the locations considered to be “high crime areas” in the city of Leeds (UK), gathered using the tools during a web-based public participation study.

The interfaces and tools presented span the needs of a GIS system user, from input to output. They are…

1) A user input interface: specifically, the user is given a spraycan tool, familiar from many

image editing packages, with which they can define fuzzy areas of varying density on a map (Figure 1A). Attribute information can then be attached to the fuzzy area.

2) 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.

3) A querying interface: this is centred around a tool that allows individuals to query an aggregated “average” map generated from multiple user responses, displaying the attribute information back to them (Figure 1B). The user querying the system picks a point on the aggregate map and all the attributes for that point are displayed, ranked/ordered on the basis of how important the point was to each user (a higher density of spraying results in a higher ranking for the attribute information).

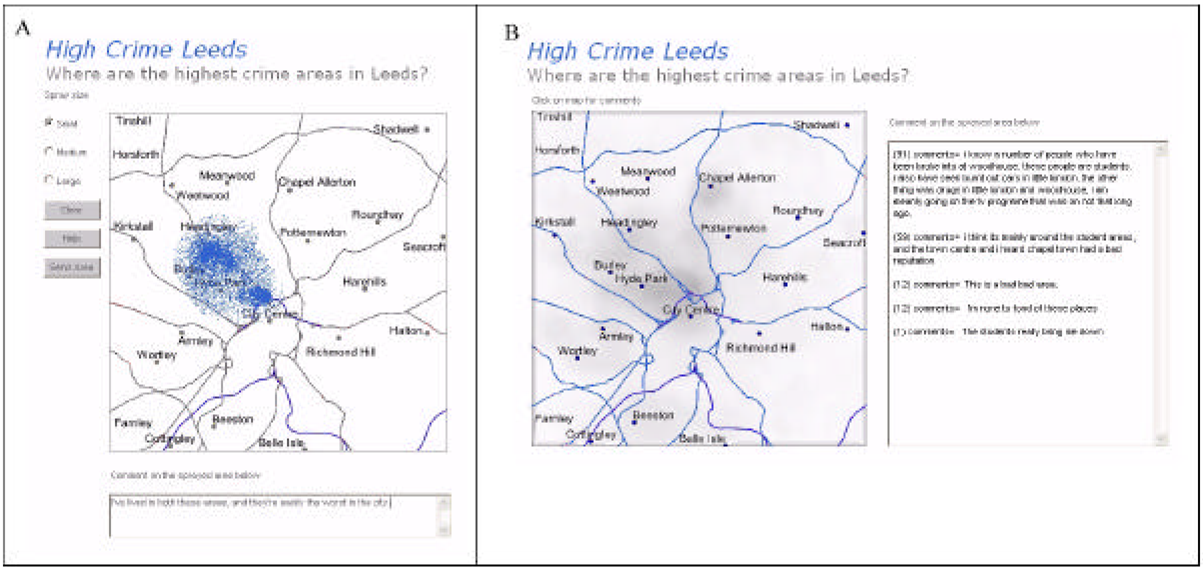

Figure 1. A: a user inputted area of perceived “high crime”. B: Output showing all user areas aggregated/averaged and ranked comments for one area.

2.2 Details of the system

The system is based around the recording and combining of user areas defined using a spraycan. Broadly, each sprayed area entered by the user is processed into a standard image format and compressed. The image and associated attribute information is stored and can be viewed, either as an individual entity or combined into an aggregate “average” map based on all the users’ responses. The three interfaces give access to the same system for different users. The system is comprised of the interfaces, a set of processing components and a set of storage components. The sequence diagrams for the three use cases are given in Figure 2. The interface and processing elements are currently written in Java, and the storage components written in Perl. The interfaces run as parameterised Applets with the option for passing through hidden-style CGI variables. This combination enables a relatively easy setup for those familiar with Applets and CGI scripts, and no understanding of either programming language is necessary to run or use the system.

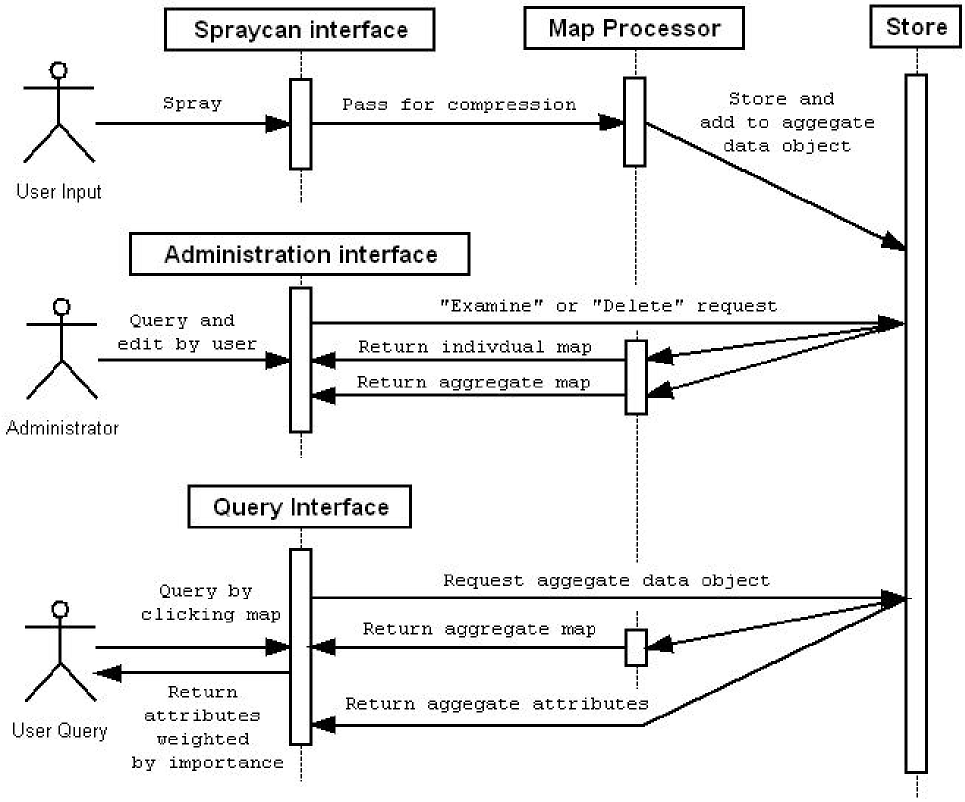

Figure 2. Unified Modelling Language sequence diagrams for use cases with the following actors: a user wanting to add one or more areas to a map; an administrator wanting to look at individual records and a combined map; a user wanting to see an aggregated average map and query it for other people's attributes.

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, but takes little time to master if this is the first time the user has seen one (the interface has a "clear" facility to allow for experimentation). Graphics packages usually offer one of two types of spraycan (Figure 3). In the first, which we shall denote the "continuous can" (Figure 3A), 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 of the image. The second type of spraycan (Figure 3B), which we shall denote the "dot 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.

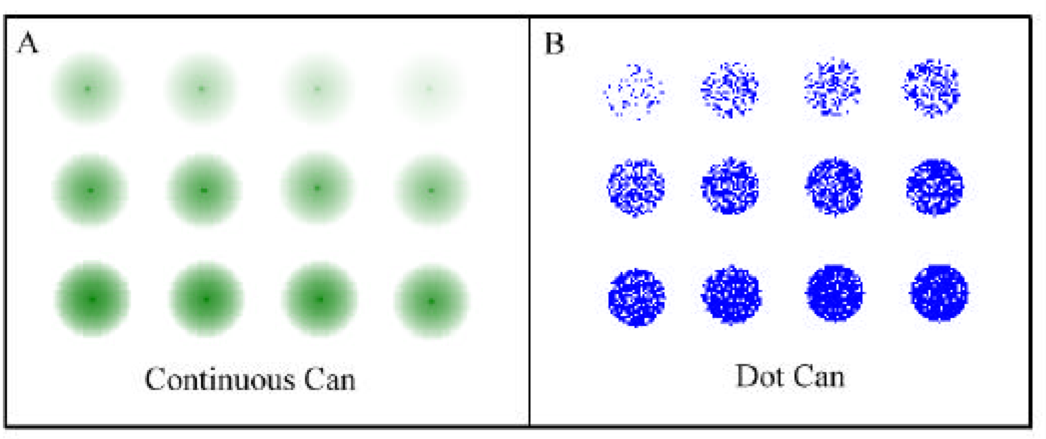

Figure 3. Two spraycan examples. A: continuous can example from PaintShop Pro. B: dot can example from Microsoft Paint.

Usability tests with both types of spraycan suggest that the dot can gives users better control when delimiting geographical areas. 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. There is a hybrid version in which dots are sprayed with a Gaussian distribution around the centre point, however this seemed to combine the worst aspects of the two pure algorithms.

A considerable problem in dealing with density surfaces of any sort is the size of the data that results. While judicious convex-hull clipping and georeferencing can limit the size of the area stored, we are, essentially, dealing with raster data that includes a considerable amount of variation. As the tools were designed to collate multiple user entries over the web, this raises significant problems. To reduce the data storage, transfer and processing overheads the sprayed areas undergo considerable processing in the system.

1) The sprayed areas are converted into images separate from the background map.
2) The dot densities are converted into continuous density surfaces using a nine-by-nine averaging kernel on each pixel and its neighbours (Figure 4A).
3) The images are reduced to a fifth of their size (Figure 4B).
4) Standard lossless image compression techniques are used to convert the uncompressed data into images in a format that can be opened in any standard graphics package. The images are then streamed to the server code and stored.
5) A data object combining the compressed images and attributes is compressed using the GZip algorithm (Deutsch, 1996) and stored on the server. This stored object also contains all other user responses.
6) When the system is queried, the image processing is reversed (Figure 4C/D). The individual or aggregated images are expanded to their original size and kernel-averaged using a five-by-five kernel to smooth inflation artefacts. However, they are left as a continuous density surface rather than re-represented as dots, as this makes further processing simpler.

Tests suggest a typical compression rate is two orders of magnitude. For example, a combined image and data object of 859Kb was compressed to 67Kb just using the GZip algorithm, and further compressed to 14Kb with the addition of the shrinking process. User tests suggest that this compression methodology includes the maximum amount of shrinkage and averaging the sprayed areas can be subjected to with the users staying happy their views / areas are represented as they sprayed them. There are, however, alternative data treatments: for example, it is likely that the field-orientated approach outlined by Laurini and Pariente (1996) could further improve compression. In their approach, continuous surfaces are defined by boundary polygons and sample points that enable the re-interpolation of the continuous field, with additional vectors controlling discontinuities or barriers. However, such compression is unlikely to give much better results in this case, 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.

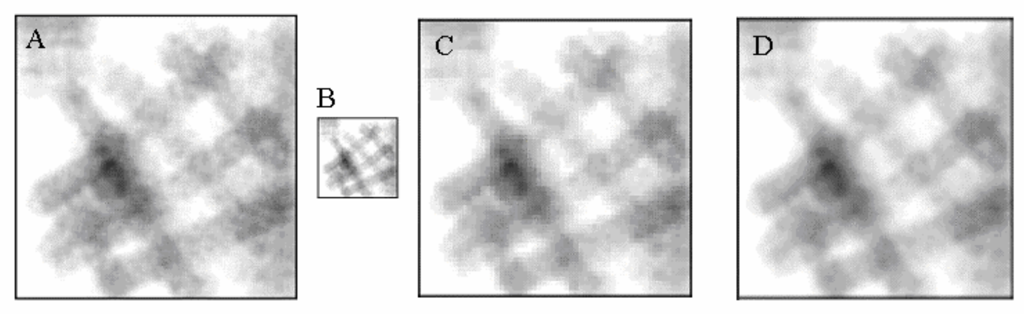

Figure 4. Compressed and inflated figures. Processing proceeds from left to right. A: density surface generated by kernel-averaging sprayed dots. B: shrunk image. C: inflated image. D: kernel-averaged inflated image.

3. Example data collection: "High Crime" areas in a UK city

3.1 The study

The system was tested in a live context as part of a study of where inhabitants thought were "High Crime" areas in the city of Leeds (UK). In Britain generally, people have a higher fear of crimes occurring to them than the actual rates of victimhood suggest is likely (Simmons *et al.*, 2003). 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.*, 2003). Concern about crime therefore represents a significant influence on many peoples' lives, and influences which areas people travel to at different times.

Leeds is a city of some 715,402 people residing within 562km^2, with 118,559 reported crimes in 2001/2002 (Kongmuang, in prep). Most of these crimes are concentrated in the city centre (Figure 5A), 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 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.

As the pilot users were gained by advertising the system within the authors’ University, the group’s demographics or knowledge of the city will not necessarily be that of the general populous. Given the correlation between University workers and broadsheet readership (CACI, 2003) 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 increase between 1999 and 2001/2002, compared with 26% of broadsheet readers, against an actual fall of 14%: Simmons *et al*., 2003).

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 whole community of users, and view people’s comments associated with specific locations (Figure 1B).

3.2 Responses

The study system captures those locations individuals, and the participating community, believe have the highest levels of crime. These areas are shown in Figure 5B. At each pixel the map shows the intensity levels for all users summed and divided by the number of users. That is, it shows 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.

As a simple example analysis these areas can be compared with absolute crime levels. These are shown in Figure 5A. The same averaging and colour rescaling treatment has been applied to this data as to the user-sprayed areas. Comparison of these two datasets allows us answer questions such as: “where do people have mis-perceptions as to the level of risk from crime?” (Figure 5B) or (if the areas had matched) “what level of crime do people notice as ‘high’?”. Additional data sprayed by users could further enhance this analysis: spraying the areas they are familiar with might allow us to test whether their knowledge was representative of the broader population, or to compare or normalize levels of crime concern with the familiarity of areas. Knowing other information, such as the demographics of users and where people have lived or experienced crime might also allow us to disaggregate their results. The comments provided by users give an insight into the processes they used to determine the areas, though a full analysis of this data is beyond the scope of this paper.

Users were also 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 justice related difficulties in presenting such data back to users, both in map form and the 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 6).

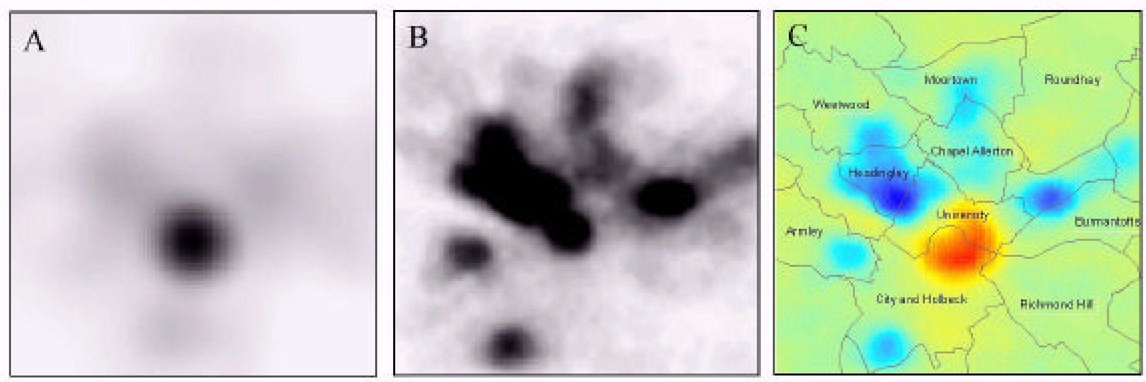

Figure 5. A: Total crime densities for Leeds for all crimes recorded in 2002. Darker areas are higher in crimes. The circular high is real and largely reflects the position of the inner ring road. B: Areas selected as "high crime" areas by users cumulated from August to September 2002. Darker areas are 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 have higher crime than expected, blue areas lower. Wards are shown for reference.

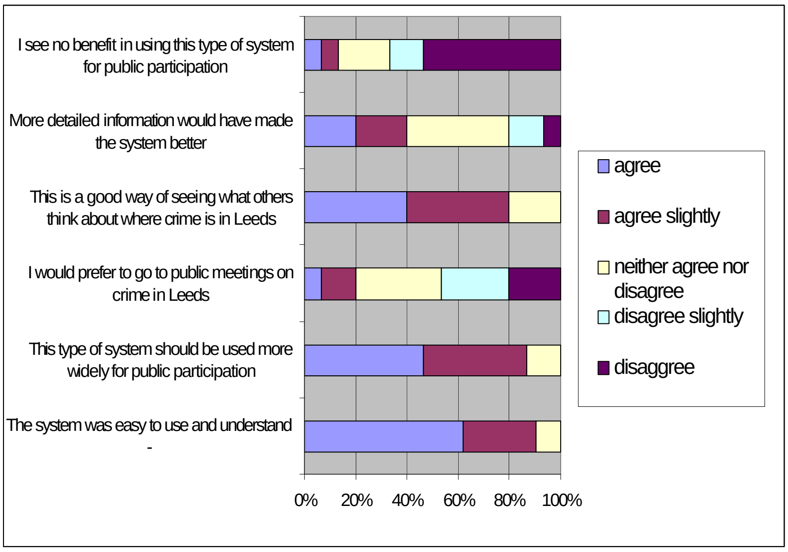

Figure 6: User feedback garnered by questionnaire at the end of system use.

4. Conclusions

We all live in a geographical world, even those without the dubious good fortune to be geographers. This may seem an obvious statement, 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're relatively simple to collect, have nice clear qualities, and we've reached a mutual agreement with the decision makers that they're 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, utilise, 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 humanities' use, understanding, and interaction with this world, we need to see it as the majority of human beings see it: not thin and anaemic, but rich and inventive.

With this mind, this paper presents a set of tools collected in a Geographical Information System for the elucidation of vernacular geography. The system allows the collection of diffuse and varying geographical areas, and the tagging of them with attribute information. While the attribute information collected in our example was very general (the way people felt about an area), there is no reason why the system should not be used without this option simply to delimit one type of geographical area ("where is the 'East End'?", "what areas do you know most about?"). Equally, such attribute information can provide the input into a more formal cognitive/semantic GIS (for an example and review, see Mennis, 2003).

The system is presented with the aim of making it easier to draw together the professional and popular ways of understanding of the world, to their mutual advantage. For the professional, capturing the popular worldview should bring enhanced understanding 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, rather than the Latin tongue. 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.

Further information and source code can be found via http://www.ccg.leeds.ac.uk/democracy/

5. Acknowledgements

The suggestion that the crime data could be made more general using these tools by assessing where users were familiar with was made by Dave Martin at GISRUK 2003. Many thanks to him and the many others at that conference for their useful comments.

6. References

CACI, 2003, Acorn: the complete consumer classification. (CACI Limited).

DEUTSCH, P., 1996, GZIP file format specification version 4.3. Request for Comments 1952, Network Working Group. Online: http://www.isi.edu/in-notes/rfc1952.txt Accessed 9th Jul 2003.

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.

KONGMUANG, C., In prep., Modelling Crime: a spatial microsimulation approach. Unpublished Ph.D. thesis, University of Leeds.

LAURINI, R., and PARIENTE, D., 1996, Towards a Field-oriented Language: First Specifications. In *Geographic Objects with Indeterminate Boundaries* Edited by

Burrough and Frank, (Taylor and Francis), pp. 225-236.

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

SIMMONS, J., and Colleagues, 2002, *Crime in England and Wales 2001/2002*. (HMSO) Online: http://www.homeoffice.gov.uk/rds/pdfs2/hosb702.pdf Accessed 9th Jul 2003.