Modelling Spatial Extents of Village Territories from Postal Address Records

Firdos Almadani[1], Peter Fisher[2] and Claire Jarvis[3]

[1]Department of Geography
University of Leicester
Leicester LE2 3TD
fma7@le.ac.uk
[2]Department of Geography
University of Leicester
pff1@le.ac.uk,
[3]Department of Geography
University of Leicester
chj2@le.ac.uk

KEYWORDS: Fuzzy objects, modeling vague regions, kernel density estimation, postal address records

Abstract:

This paper describes a method to derive approximate boundaries for imprecise regions recorded as a set of points such as rural settlements from postal address records. We discuss how such regions are typically not represented as formal administrative boundaries and are often considered to be vague. We then propose methods to approximate the extent of the villages as fuzzy objects based on kernel density estimation. The methods introduced are evaluated to judge the potential of the approach.

1. Introduction:

There are many situations in Geographical Information Science (GISc) where it would be useful to be able to consider poorly or vaguely defined geographical places. Some regions, like urban neighbourhoods and village territories, are geographically vague, both in definition and extent, as they often do not have officially defined boundaries (e.g. downtown) or have formal or informal names connected to past events or landmarks. The spatial footprints of such regions are thus a matter of individual perception and may depend on social factors as much as physical environmental differences. With these issues in mind, several attempts have been made to develop modelling techniques that adequately handle vagueness and imprecision in GISc.

The pages that follow propose a method to model the spatial extent of and generate approximate boundaries for vague places. Looking at rural settlements in this case, the paper goes on to point out how the notion of vague places can be used intelligently within a GIS. This paper is structured as follows: in the next section, we review some related work on the representation of vague regions; in Section 3, we introduce our method for constructing fuzzy footprints of villages; next, in Section 4, we evaluate the performance of our approximation in the context of GIS; and finally, Section 5 presents some concluding remarks and directions for future work.

2. Background:

2.1 Vagueness in Geography:

Vagueness is a particular type of imprecision where it is difficult to decide, in borderline cases, whether a concept or situation applies or not (Worboys, 2001). In other words, a concept is vague if locations exist that cannot be classified either to the concept or to its complement. It might be hard to decide whether a certain location belongs to one particular vegetation class or to another when mapping vegetation, for example. The transition between these classes may be gradual, as between forest and grassland, sometimes called an ecotone or more correctly an ecocline. The 'sorites paradox' can be applied to explain the notion of vagueness. This can be illustrated by considering whether a few sand grains make a heap (Fisher, 2000). Other examples include what is a mountain is and where it begins, how far the Australian outback extends, or where to mark the borders of Central London.

It is now generally recognised that difficulties arise when an attempt is made to identify the category and extent of many geographical features. There are some words or expressions used in the literature to describe such features; they are said to be hard to define, ill-defined, indeterminate, or vague. As Hollenstein (2008) argues, there are two broad categories of vague geographic entities. The first type relates to the majority of natural geographic phenomena that are spatially ill-defined, such as vegetation zones or soil types. The second type expresses human conceptions about vague places and their extents, for example, aboriginal territories or urban neighbourhoods. This view is supported by Erwig and Schneider (1997), who consider fuzziness as an intrinsic feature of an object itself; each object may either lack a precisely definable border (first type) or by its nature lack the ability to be precisely defined (second type).

2.2 Modelling Vague Regions:

Some work has already been done modelling villages which inherently have indeterminate boundaries. Crawford (2002) for example develops and evaluates a novel approach for linking population and environmental data (by transforming discrete village points, with associated demographic and economic attribute data) across a thematic domain, and spatially representing functional regions that partition the landscape into village territories. Likewise, Gray (2008) generates fuzzy models of a settlement's spatial extent, using settlements' place names gained from address point data for postcodes.

Alternatively, Montello *et al.* (2003) conducted an empirical study that investigates people's perception about the extent of downtown Santa Barbara. Similarly, Mansbridge (2005) explored where people consider the vague area of Sheffield City Centre to be; and on a smaller scale the Midlands. Moreover, Waters and Evans (2003 & 2008) use a web based mapping system for capturing fuzzy areas and their associated attributes. Their system contains a "spraycan" tool that allows users to tag information onto diffuse areas of varying density to present the locations considered to be "high crime areas" in Leeds. Recently, a similar study suggested by Rosser and Morley (2010) offers a social networking application for eliciting people's understanding of places, in which users are encouraged to rate and mark the extent of places on a map according to whether they "love", "like", "dislike" or "hate" that area.

Furthermore, there is a large volume of published studies that employ the web as a source to model the extent of vague places (Goodchild *et al.* 1998; Arampatzis *et al.* 2006; Schockaert and de Cock 2007; Jones *et al.* 2008; Twaroch *et al.* 2008; Hall 2010; Hollenstein and Purves, 2010). This body of work focuses on obtaining sample data to recognise vague or vernacular place names that users commonly employ when dealing with photographic websites (e.g. Geograph, and Flickr), search engines (e.g. SPIRIT or Google Local) or social websites (e.g. Gumtree or Facebook). Most of these studies are based on methods that provide a density surface as a representation of the vague region while also facilitating generation of crisp approximations of the boundary of the region at different levels of confidence if they are required (Jones *et al.* 2008). In the current paper, we pursue the same density surface modelling approach in a different context and in greater depth (see Section 3 onwards). Here the methods are used to model human settlements (villages or towns; Crawford *et al*. 2002; Gray 2008) from point data of postal addresses.

3. Modelling the fuzzy spatial extent of a settlement:

3.1 Study area and data acquisition:

The study area of this research covers the rural villages in Hinckley and Bosworth District in Leicestershire County (Figure 1). Address point data for this area were obtained from the Ordnance Survey (OS Address Layer 2). These data consist of postal address information where for every house a village name can be identified and extracted and used to approximate an extent of any village. The work also incorporates additional sets of data needed for a better analytical model on the problem, such as parish boundaries and the OS MasterMap. These further data were downloaded under the academic license from UKBorders and EDINA Digimap.

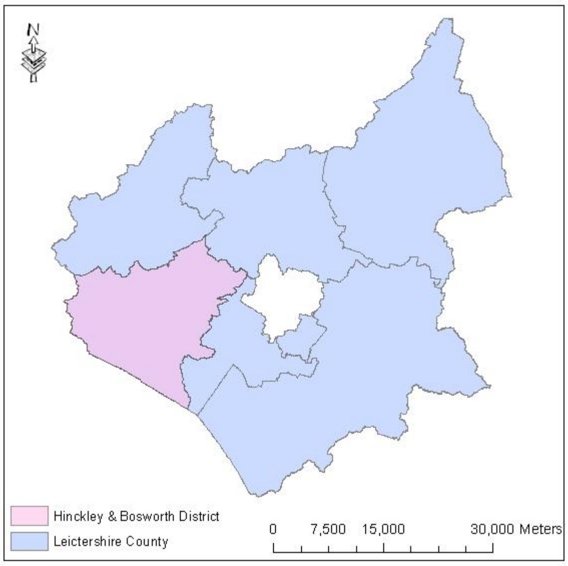

Figure 1: Map showing the entire study area "Hinckley & Bosworth District".
© Crown Copyright/database right 20(yy). An Ordnance Survey/EDINA supplied service.

3.2 Methods and techniques:

Most of the work available on fuzzy spatial modelling, as reviewed above in Section 2.2, is based on either empirical research focused on how people think and communicate with vague regions (Montello *et al.* 2003; Mansbridge, 2005) or methods that focus on obtaining sample data from the web (Arampatzis *et al.* 2006;

Jones *et al.* 2008), although some studies belong to both categories (Waters and Evans, 2003; Evans and Waters, 2008; Rosser and Morley, 2010;). Since our main interest here is not people's perceptions of vague regions but on deriving fuzzy regions from postal addresses, we follow and extend Gray's (2008) approach. He developed fuzzy models using settlements' place names gained from address point data rather than using those from the web (tag points) or human subjects (questionnaires). No research to date, however, has been found that adequately covers the characteristics of these vague places (attributes of the fuzzy objects) and even the operations derived from them have not been identified. This research, essentially, seeks to remedy these omissions by taking further steps to exploring fuzzy regions. Here we only consider that aspect of our methodology to model the fuzzy spatial extent of villages.

A common approach applied to generating approximate boundaries for vague places is density estimation (Hollenstein, 2008) also known as *Kernel Density Estimation (KDE)* (Jones *et al.* 2008; Twaroch *et al.* 2008; de Berg *et al.* 2011). KDE estimates the proportion of total incidents (addresses) that can be expected to occur at any given map location. It works by first overlaying an area of interest with a rectangular 2-dimensional grid. It then calculates an estimate of the density of incidents in each grid cell which is based on a weight function, the kernel (a function of specify shapes and bandwidths or search radius). This reflects the fact that there is a greater probability of an incident occurring in a given location the closer it is to the location of a known incident.

It should be noted that the choice of the surface resolution (grid cell size) and the bandwidth (kernel radius) in this function both influence the shape of the surface in terms of its smoothness or peakness (Bowman and Azzalini, 1997; Silverman, 1986). According to Jones *et al.* (2008), caution must be applied in the selection of the surface resolution and the kernel radius. The surface resolution must be sufficient to resolve the boundaries of the region and will vary according to region. The kernel radius should ideally be small enough to represent local variation within the region at a scale commensurate with the size of the region and large enough to capture multiple point locations within the kernel radius. In this research, therefore, the KDE for each village was computed independently based on postal address point data using the *standard distances.* Similarly to the way a standard deviation measures the distribution of data values around the statistical mean, *standard distance* measures the degree to which features are concentrated or dispersed around the geometric mean centre (Silk, 1979; de Smith *et al.* 2007). It may therefore provide a suitable bandwidths for all the villages (Hollenstein and Purves, 2010). This is evaluated based on recognition of the core area of the village, an area that has a high density of addresses.

4. Results and Discussions:

4.1 Extracting villages from postal records:

A number of rural villages were extracted from the OS AddressPoint data based on two types of addresses – a British Standards Institute format (*BS7666*) and the traditional *Postal* Addresses. Uncertainty was noted in the imprecision related to village names and settlement spatial extents. Although some of these settlements have precisely defined parish boundaries, the actual extent for the addresses is sometimes different (Figure 2). In addition, parishes for the most part comprise two or more rural settlements, which often have overlapping addresses as can be seen in Figure 3.

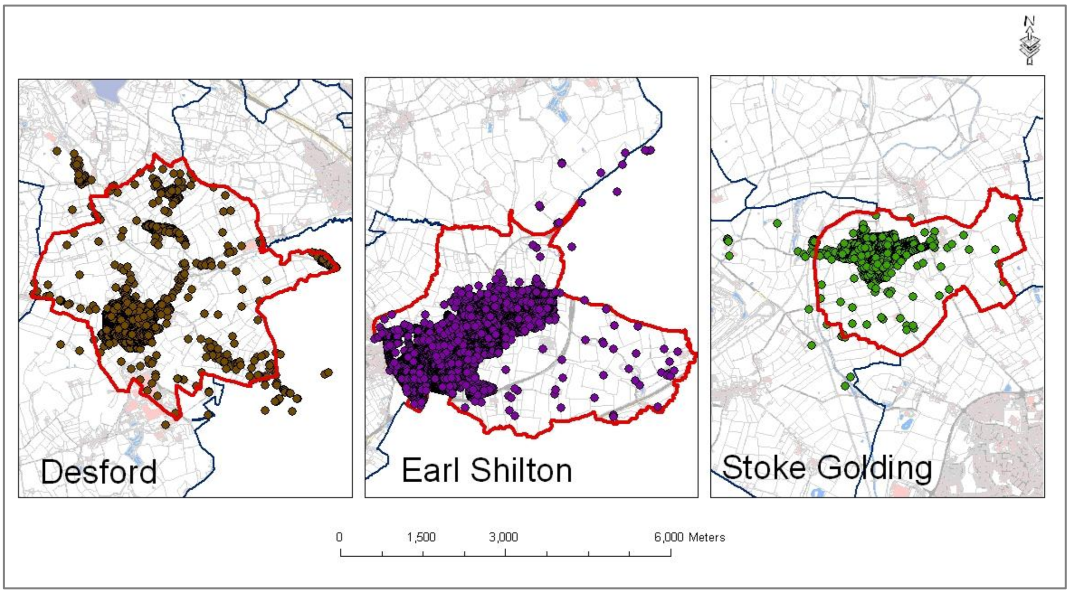

Figure 2: Examples of village names whose extent does not coincide with the parish boundaries.

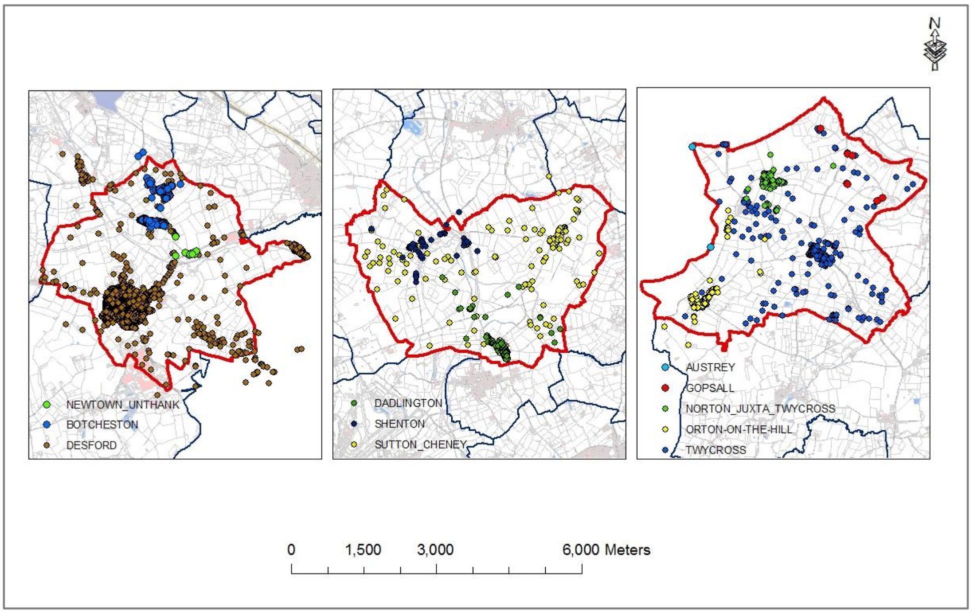

Figure 3: Examples of parishes that comprise more than two rural settlements.

Another interesting observation is the disparity in the village names recorded in both address types which would be expressed to be the same. For instance, using a hyphen or space to separate between words in village names as the cases in “Higham on the Hill” (or “Higham-on-the-Hill”) and “Stanton under Bardon”; Figure 4 below illustrates these differences. There are several possible explanations for these variations. It was decided therefore to consider these variations in names as different villages.

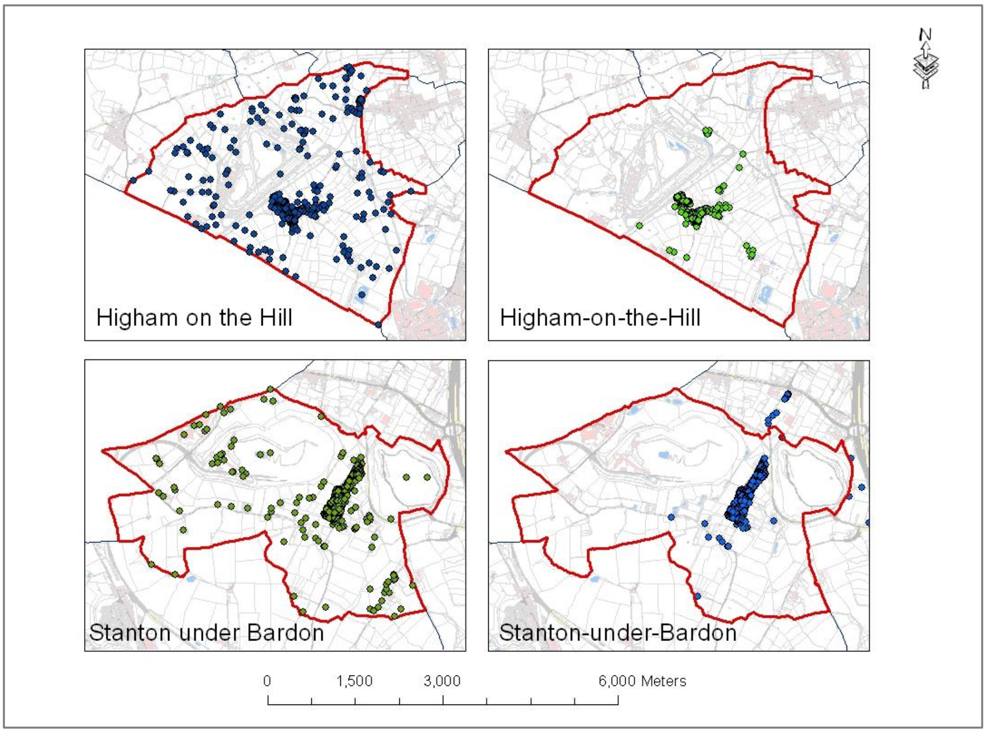

Figure 4: Examples of the disparity in the village names.

4.2 Generating fuzzy spatial extents:

In this section we evaluate various aspect of using KDE to generate approximate boundaries for rural settlements. As prior studies have noted the importance of the choice of bandwidth parameter (Jones *et al.* 2008; Twaroch *et al.* 2008;), the kernel radius used in this paper is based upon a measure of standard distance (Silk, 1979; de Smith *et al.* 2007). From the results presented in Figure 5, it is apparent that one standard distance provides a sensible option to explore how address points are distributed across the spread of villages. The present method is useful at defining the territory in at least two respects: first, it has important implications for analyzing spatial patterns and distribution of phenomena. It can be seen from the maps in Figure 5 that the village patterns are reflected in the density estimation, whether the villages are nucleated like Burfton and Market Bosworth, or dispersed (unfocussed), or multi-focal, Sutton Cheney, with three clear focii. Second, it is also useful in both determining the overlapping between rural settlements and comparing the approximated vague region with its equivalent formal parish boundaries, as indicated in Figure 5.

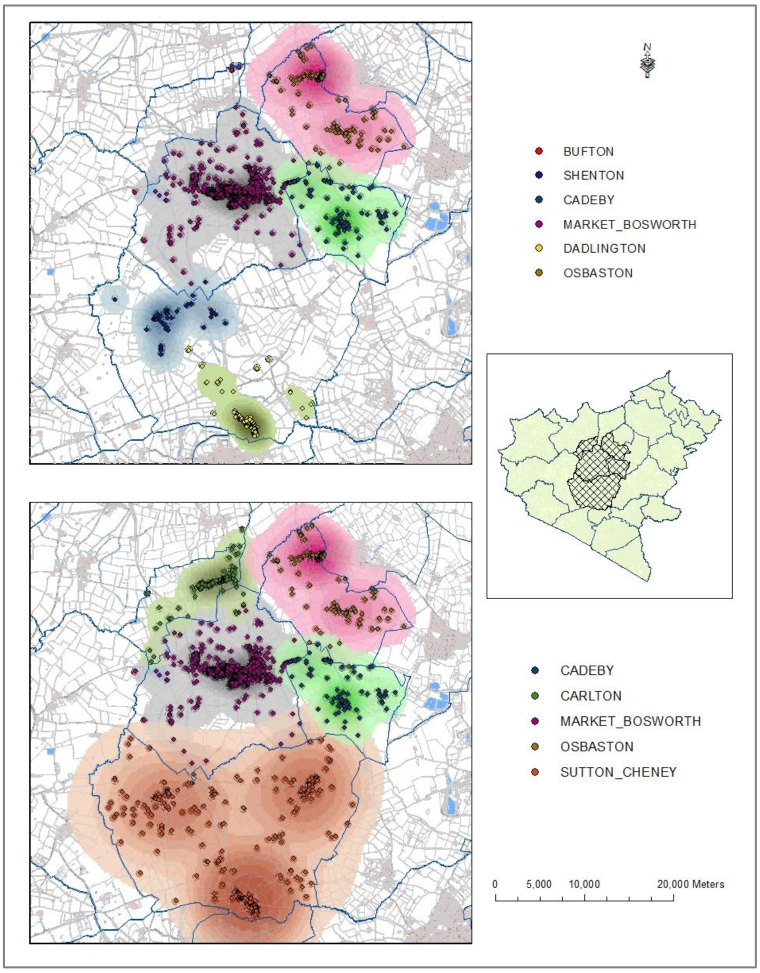

Figure 5: Some examples of the resultant KDE surfaces.

5. Conclusions and Future Work

The proposed method seems capable of representing the spatial extent of village territories from address point data. Further, we have shown that the resulting representations can differ relatively from the equivalently named settlements' parish boundaries. Not discussed in this paper, but also an important part of our research, are methods for defining fuzzy regions based on other techniques including convex hulls and α-shapes. However, these results are not very encouraging when considering further spatial analysis once fuzzy regions have been modelled. A priority for more research thus is needed to better understand when we have the right approximation of settlements' spatial extents as fuzzy objects. Future work will also have to deal with exploring the fuzzy distance between them.

6. Acknowledgments:

We would like to thank Ordnance Survey for providing us with the Address Point data to use in our research.

7. References:

ARAMPATZIS, A., VAN KREVELD, M., REINBACHER, I., JONES, C.B., VAID, S., CLOUGH, P., JOHO, H. and SANDERSON, M., (2006). Web-based Delineation of Imprecise Regions. *Computers, Environment and Urban Systems*, **30**(4), pp. 436-

459.

BOWMAN, A.W. and AZZALINI, A., (1997). *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations.* Oxford: Clarendon Press.

CRAWFORD, T.W., (2002). Spatial modelling of village functional territories to support population-environment linkages. In: S.J. WALSH and K.A. CREWS-MEYER, eds, *Linking People, Place, and Policy: A GIScience Approach.* Boston: Kluwer Academic Publishers, pp. 91-111.

DE SMITH, M.J., GOODCHILD, M.F. and LONGLEY, P., (2007). *Geospatial analysis: a comprehensive guide to principles, techniques and software tools.* 3ed edn. Troubador Pub.

ERWIG, M. and SCHNEIDER, M., (1997). Vague Regions, M. SCHOLL and AND A. VOISARD, eds. In: *Advances in Spatial Databases, 5th International Symposium, SSD '97 Conference Proceedings*, July 15-18, (1997), Berlin: Springer, pp. 298-320.

EVANS, A.J. and WATERS, T., (2008). Mapping Vernacular Geography: Web-baed GIS Tools for Capturing "Fuzzy" or "Vague" Entities. *International Journal of Technology, Policy and Management,* **7**(2), pp. 1468-4322.

FISHER, P., (2000). Sorites Paradox and Vague Geographies. *Fuzzy Sets and Systems*, **113**, pp. 7-18.

GOODCHILD, M.F., MONTELLO, D.R., FOHL, P. and GOTTSEGEN, J., (1998). Fuzzy Spatial Queries in Digital Spatial Data Libraries, *Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference*, 4-9 May 1998.

GRAY, T., (2008). *Modelling a settlement's spatial extent from placename records using kernel density, convex hull and viewshed analysis.* MSc edn. University of Leicester.

HALL, M., (2010). *Modelling and Reasoning with Quantitative Representations of Vague Spatial Language used in Photographic Image Captions.* PhD Thesis edn. Cardiff University.

HOLLENSTEIN, L., (2008). *Capturing Vernacular Geography from Georeferenced Tags*, University of Zurich.

HOLLENSTEIN, L. and PURVES, R.S., (2010). Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science,* **1**, pp. 21-48.

JONES, C.B., PURVES, R.S., CLOUGH, P.D. and JOHO, H., (2008). Modelling Vague Places with Knowledge from the Web. *International Journal of Geographical Information Science,* **22**(10), pp. 1045-1065.

MANSBRIDGE, L., (2005). *Preceptions of Imprecise in Relation to Geographical Infrormation Retrieval*, MSc thesis, University of Sheffield.

MONTELLO, D.R., GOODCHILD, M.F., GOTTSEGEN, J. and FOHL, P., (2003). Where's Downtown? Behavioral Methods for Determinig Referents of Vague Spatial Queries. *Spatial Cognition & Computation,* **3**(2 & 3), pp. 185-204.

ROSSER, J. and MORLEY, J., (2010). Rate my place: a social network application for crowd-sourcing vernacular geographic areas, M. HAKLAY, J. MORLEY and H. RAHEMTULLA, eds. In: *Proceedings of the GIS Research UK 18th Annual Conference*, 14th – 16th April (2010), University College London, pp. 155-160.

SCHOCKAERT, S. and DE COCK, M., (2007). Neighborhood Restrictions in Geographic IR, *The 30th Annual International ACM SIGIR Conference*, 23-27 July 2007 (2007), pp. 167-174.

SILK, J., (1979). *Statistical concepts in geographay.* London: Allen & Unwin.

SILVERMAN, B.W., (1986). *Density Estimation for Statistics and Data Analysis.* New York, USA: Chapman and Hall.

TWAROCH, F.A., JONES, C.B. and ABDELMOTY, A.I., (2008). Acquisition of A Vernacular Gazetteer from Web Sources, *Proceedings of the first international workshop on Location and the web*, 21-25 Apri 2008 (2008), ACM New York, USA, pp. 61-64.

WALSH, S.J. and CREWS-MEYER, K.A., eds, (2002). *Linking People, Place, and Policy: A GIScience Approach.* Boston; London: Kluwer Academic Publishers.

WATERS, T. and EVANS, A.J., (2003). Tools for Web-based GIS Mapping of A "Fuzzy" Vernacular Geography, *Proceeding of 7th International Conference on GeoComputation, Southampton*, 8-10 September (2003).

WORBOYS, M.F., (2001). Nearness Relations in Environmental Space. *International Journal of Geographical Information Science,* **15**(7), pp. 651.

8. Biography:

Firdos Almadani is a postgraduate student at the University of Leicester, and a Teaching Assistant in the Geography Department at King Abdul-Aziz University, Jeddah, Saudi Arabia. Work reported here relates to her PhD research.

Peter Fisher is Professor of Geographical Information at the University of Leicester. He has researched many aspects of uncertainty in geographical information.

Claire Jarvis is a Senior Lecturer in GIScience at the University of Leicester. Her current research is mainly concerned with visualization and virtual reality.