Assessing geographic data usability in analytical contexts: Undertaking sensitivity analysis of geospatial processes

R. Frew[1], G. Higgs[1], M. Langford[1], J. Harding[2]

[1] GIS Research Centre, WISERD, Faculty of Computing, Engineering and Science, University of South Wales, Treforest, Pontypridd CF37 1DL
[2]Ordnance Survey, Explorer House, Adanac Drive, Southampton SO16 0AS

November 7, 2014

Summary

This paper addresses the continuing dearth of research on spatial data usability by applying sensitivity analysis to GIS-based accessibility models. Comparisons were made using approaches based on Euclidean distances and more sophisticated accessibility measures that utilise travel distances and times: the latter incorporating measures of supply and demand by using innovative extensions to the Enhanced Two-Step Floating Catchment Area method (E2SFCA). To illustrate the sensitivity of findings from applying such models with a range of data sources, accessibility to secondary schools was calculated for Output Areas in South Wales using an E2SFCA plug-in to ArcGIS™. By using different permutations of spatial data, for both the supply- and demand-side parameters in such models, differences in FCA scores were sought in order to comment on the usability of such data sources. Preliminary conclusions are made on the appropriateness of such data sets in relation to different types of network-based accessibility modelling tasks.

KEYWORDS: Usability; GIS-based accessibility models; spatial data; sensitivity analysis; E2SFCA.

1. Introduction

Sources of spatial data continue to expand with inevitable debates surrounding the provenance of such data and their usability for GIS-based tasks. There is thus an increased scrutiny on the quality of such data and the respective advantages and limitations of both proprietary versus crowd-sourced data.

Although the highest quality data remains expensive to obtain (for example high resolution LiDAR data, or Ordnance Survey MasterMap products), other data is becoming available without the need for expensive capital or revenue outlay. Recent reports (Avery and Gittings, 2014) of the use of unmanned aerial vehicles to produce a variety of remotely-sensed data and the availability of various software solutions, both at low costs (relative to traditionally-sourced equivalents) are enabling new data-producers to emerge, or even provides the opportunity for data users to generate their own data for their own purposes. At the same time, the quality of such data is being questioned in some quarters, building on earlier debates surrounding the use of VGI and earlier work in the field of data quality theory and assessment (Haklay, 2010; Zielstra and Zipf, 2010). However, there is still very little research into the usability of such data in relation to different types of GIS-based tasks, although Higgs et al (2012) investigated the impacts of different approaches to measuring accessibility to green space. Few studies to date incorporate sensitivity analysis which includes the use of different sources of spatial data within different stages of a 'typical' GIS project, with Jones (2010) a notable exception.

This paper will report on the usability of a range of geographical data in one such application area: namely their use in network-based accessibility modelling. Based on these findings preliminary assessments will be made on their usefulness in such tasks, using both basic and more sophisticated methods of measuring accessibility.

Accessibility studies using GIS have become a well-established component of geographical studies concerned with measuring potential inequalities in provision of both public and private services and

are beginning to be used by policy makers to inform their decisions. Related fields include studies of the spatial distribution and optimisation of services, in areas such as public health, welfare provision and environmental justice. Recent examples of such research includes those concerned with examining the geographical distribution of alcohol outlets in Glasgow in relation to deprivation (Ellaway et al, 2010) and disparities in locations of watershed restoration projects depending on socio-demographic profile (Dernoga et al 2015).

2. Study approach

2.1 Study area

Two areas in South Wales were chosen as study areas: the city and county of Cardiff; and the neighbouring local authority area of the Vale of Glamorgan. Cardiff is the largest city in Wales, although its boundary also contains outlying villages lying within the green belt that separates Cardiff from Newport (to the east) and the densely-populated Rhondda valleys to the north. The Vale of Glamorgan has several smaller population centres, with much of the area having rural characteristics, despite close proximity to major transport links (the M4 motorway) and to large towns and cities (such as Cardiff and Swansea).

2.2 Geographic data

Spatial data products used were typical of those commonly used in UK-based accessibility analysis studies, including Ordnance Survey MasterMap Integrated Transport Network™ (ITN) Layer and, additionally, ITN and Urban Path layer. OpenStreetMap (OSM) network data for South Wales was obtained from a third-party provider (due to download restrictions on the OSM website), as an example of crowdsourced/VGI (volunteer geographic information) data that is now routinely available to GIS researchers.

One further dataset was used to examine whether a product not designed for use as a network could approximate the results of the specifically-designed datasets. VectorMap District, available free-to-use under Ordnance Survey OpenData, was built into a network using standard, readily-available GIS tools (using Arc GIS).

The accessibility assessment tasks were also conducted using Euclidean (straight line) distances, again to be used in comparison with the other datasets.

The relative accessibility of different locations within the study areas was assessed using different methods, with the processes subjected to sensitivity analysis in order to identify areas of similarity and difference. In the context of this research, interest was focussed on the areas of difference, in order to identify what factors (topographic or data-related, for example) contributed to those differences. Several variations of the analytical tasks were therefore performed, using the different datasets available.

2.3 Location of supply features

Accessibility studies use various methods to assess the accessibility (or inaccessibility) of demand to supply. In this study, the supply feature is represented by secondary schools (with travel-to-school journeys being the focus of various other studies relating to active travel and children's health, child road safety, catchment areas, parental choice, etc). The location of such facilities is subject to a degree of choice by the researcher. Accordingly, as part of the sensitivity analysis, different methods of locating these features were compared. Many studies use points to represent locations, and as secondary schools may occupy large sites, they are ideal for use to compare different methods. Ordnance Survey Points of Interest were used as point locations of the schools, and Ordnance Survey Sites dataset was used to extract the "footprint" of entire school sites, including playing fields, etc.

From the Sites dataset, three different location methods could be used: centroids (the geometric centroid of the entire site); access point (one or more way in to the school site); and boundary (any point on the perimeter of the site). Figure 1 illustrates how Euclidean distances may vary, using a school site in the Vale of Glamorgan as an example, and Figure 2 shows how network distances may vary, using the same variety of location methods.

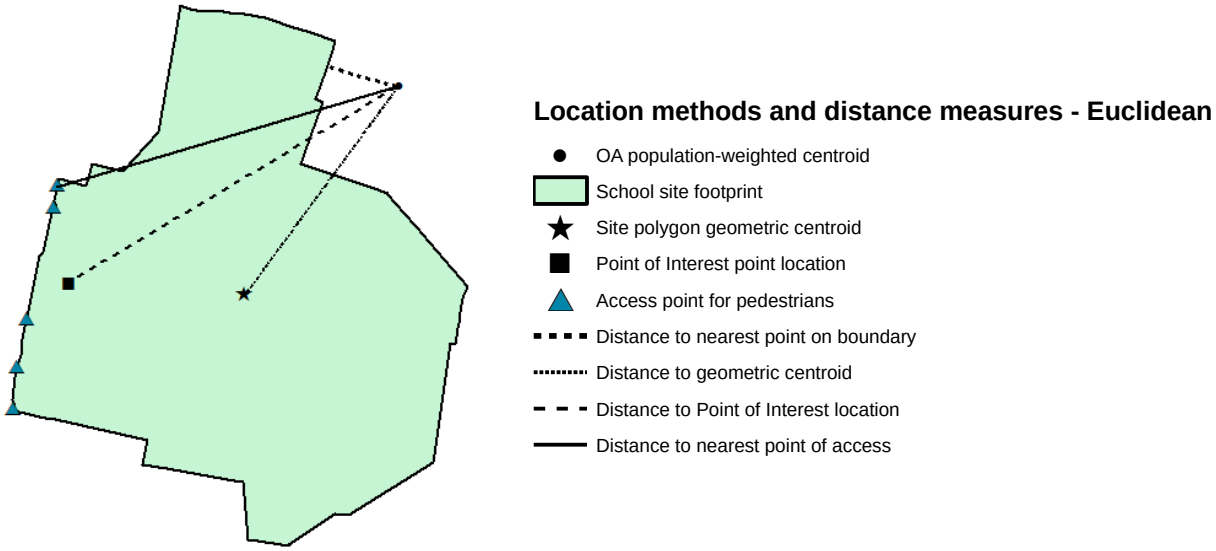

Figure 1 Four different approaches to measuring Euclidean distance from a location to a facility

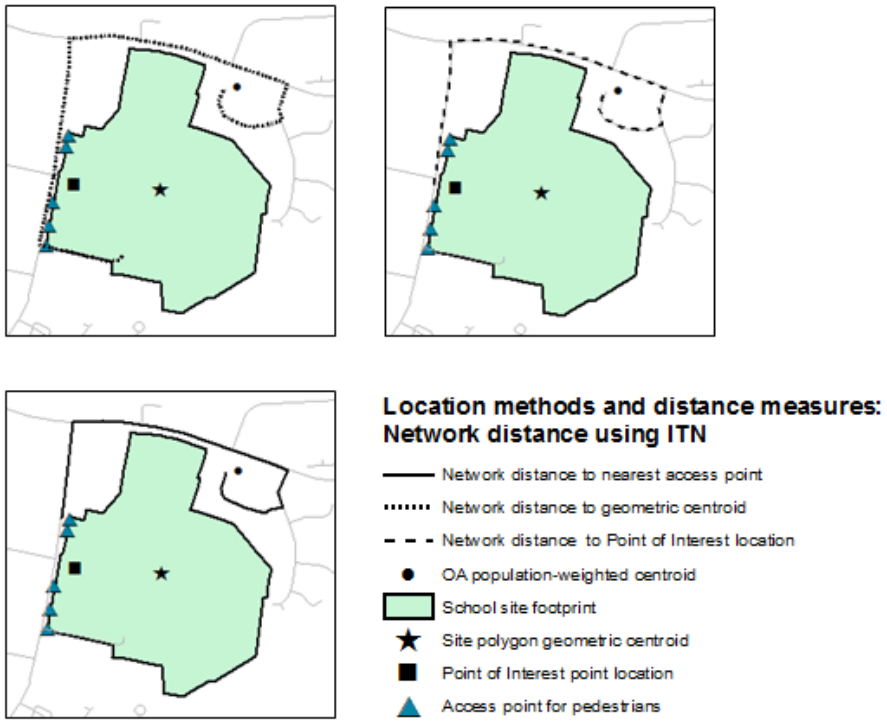

Figure 2 Examples of network distance variations, from a point location to a local facility

A further Ordnance Survey product, AddressBase Premium, was also used to locate the sites at the postal address level, represented by a point at the postal delivery location. This was used in the comparison process.

2.4 Location of demand

Demand was represented by population, and various such representations are available to researchers. In this study, the main method of locating the population was by the use of UK Census Output Area population-weighted centroids. Other methods area available, both more detailed or more generalised, though the method chosen uses readily-available and free-to-use UK data that is sufficiently detailed to allow differences to be identified within larger areas while avoiding the increased computational loads imposed by using a more detailed dataset (through, for example, representing population at the post-code or household level).

2.5 Methodology

This study initially used one of the most basic measures (Euclidean distance) before considering the differences observed when network distances were used. At the end of each permutation, the "Worst Ten" areas were identified, that is, the ten output areas which were the furthest from the facility, and the results compared. The process was then repeated using each of the network datasets, and the network distances compared, before the models were run again, this time using various permutations offered by the different supply-side options. Comparisons of differences between the various iterations of the models enabled identification of output areas to be examined in detail, isolating factors within the underlying data that contributed to these differences. Results of this stage of analysis will be presented at conference.

Using Euclidean distances is perhaps a simplistic approach of measuring accessibility, as no account is taken of the capacity of the supply facility (in this case the number of school places available), nor the level of demand (in this case the secondary school-age population). Accordingly, a more sophisticated measure of accessibility was used based on the enhanced, two-step floating catchment area method (E2SFCA). A tool created by researchers within the Wales Institute of Social and Economic Research, Data and Methods (WISERD) research institute was used in ArcGIS™, and "classic" E2SFCA utilised, that is without a distance-decay function. Figure 3 shows the user interface of the WISERD plug-in. In order for the tool to be used effectively, data on pupil numbers/school roll was obtained from information published by the relevant local authority, and an estimate of the school-age population of each OA made from age categories contained in the census returns (though there is no convenient category of "secondary school age" in the census, there is information on 12 to 16 year olds, and an estimate was made of other pupil numbers).

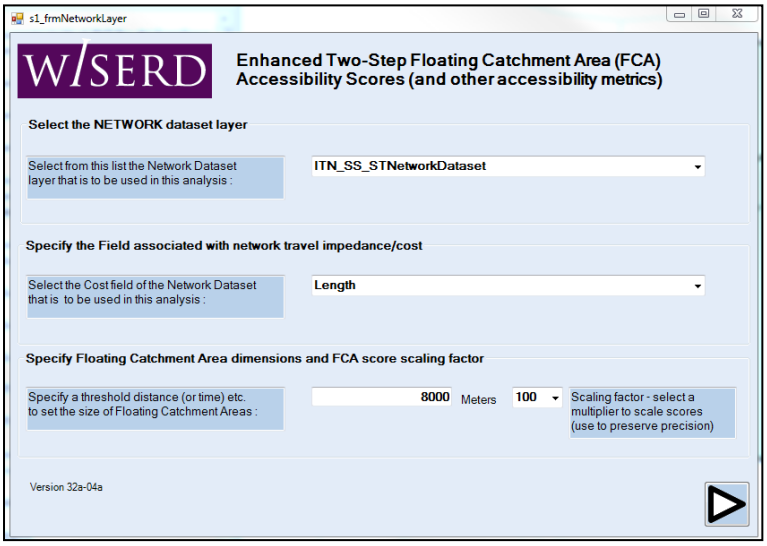

Figure 3 Screen grab of the WISERD E2SFCA plug-in tool first screen. Further screens offer the options of incorporating levels of supply and demand

3. Preliminary Findings

In order to demonstrate the implications of sensitivity analysis, preliminary research involved identifying those locations (at the Output Area level) that had the poorest accessibility measures. Although early results indicate that areas have general similarity in accessibility despite using differing data, there were also areas where results differed significantly. Figure 4 gives an example of early results in which the method of locating the supply facility influenced those OAs found to be the least accessible through mapping the ten locations with the worst (ie greatest) distance measurements to the particular location.

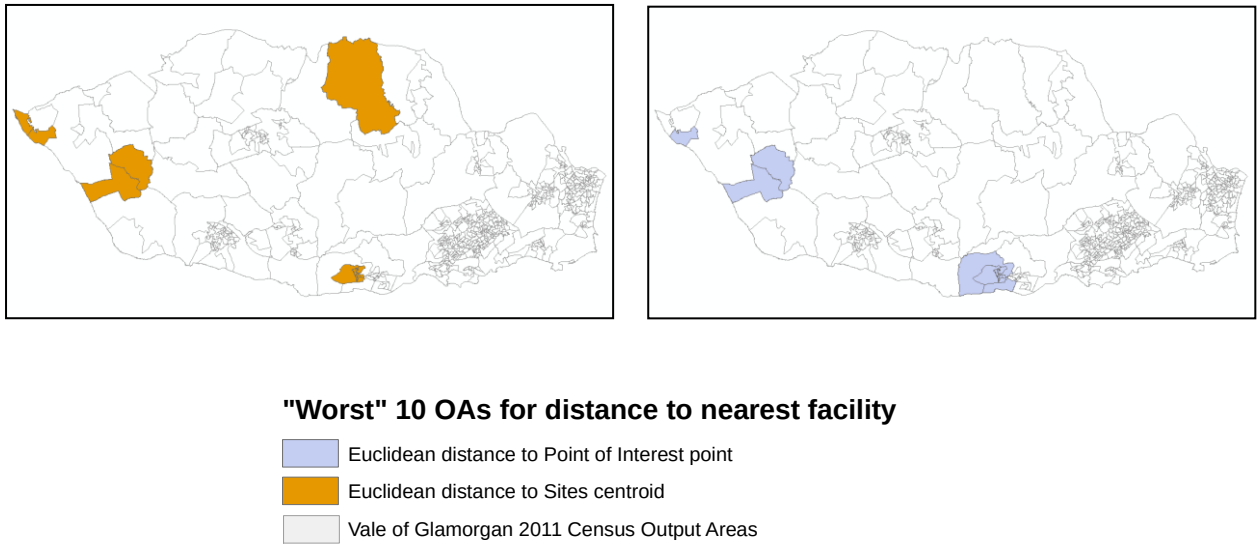

Figure 4 Distances from OAs to their nearest facility can vary depending on the method used to locate that facility (in this case, secondary schools). This example illustrates how the choice of method can affect the result

Although the patterns of 'worst' accessibility are broadly similar according to each method of measurement, there are subtle differences stemming from the choice of dataset. Differences in findings when applying alternative methods of measurement (i.e. between the basic Euclidean-distance measures and the results of E2SFCA measures) are considerable, again raising questions as to the implications of using different approaches on the results from GIS-based models. Preliminary findings also indicate urban/rural differences which also merit investigation, implying that what is usable in an urban context may not be ideal for rural research applications.

Further methods of understanding the implications of using a range of spatial datasets within these network-based models are being developed and will be presented at the conference.

Practical issues with the different datasets, along with their currency and update patterns, are also worthy of further study, suggesting that researchers need to be made more aware of the implications of using different sources of data in 'typical' GIS tasks. Preliminary methods whereby the usability of spatial data sources can be made more transparent to researchers in relation to the nature of such tasks will be included in the conference presentation.

4. Acknowledgements

The study reported here forms part of an Ordnance Survey-sponsored PhD research programme. However any views expressed herein do not necessarily represent those of Ordnance Survey.

5. Biography

Robin Frew is a postgraduate student at the University of South Wales and is in the second year of a PhD investigating spatial data usability.

Professor Gary Higgs is currently Director of the GIS Research Centre in the Faculty of Computing, Engineering and Science, University of South Wales and a co-Director of the Wales Institute of Social and Economic Research, Data and Methods (WISERD). Over-arching research interests are in the application of GIS in social and environmental studies, most recently in the areas of health geography and emergency planning.

Dr Mitch Langford is a Principal Lecturer in the Faculty of Computing, Engineering and Science, University of South Wales. His current research interests include dasymetric mapping, population modelling, and geospatial analysis within the fields of healthcare, social equality and environmental justice.

Dr Jenny Harding is a Principal Research Scientist at Ordnance Survey (GB) with particular interests in user focused research, geography and geographic data usability. Her role includes leading research in these areas both internally within Ordnance Survey and externally in collaborative projects with universities.

References

Avery S and Gittings B (2014). Do it yourself drones – experimenting with low cost UAVs. *GIS Professional*, October, 12-14.

Dernoga M, Wilson S, Jiang C, Tutman F (2015). Environmental justice disparities in Maryland's watershed restoration programs. *Environmental Science & Policy*, 45, 67-78.

Ellaway A, Macdonald L, Forsyth A, Macintyre S (2010). The socio-spatial distribution of alcohol outlets in Glasgow city. Health & Place, 16(1), 167-172.

Haklay M (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37, 682-703.

Higgs G, Fry R, Langford M (2012). Investigating the implications of using alternative GIS-based techniques to measure accessibility to green space. *Environment and Planning B: Planning and Design*, 39, 326-343.

Jones S (2010) Open geographical data, visualisation and dissemination in public health information. *AGI Geocommunity '10*. Available at:
http://www.agi.org.uk/storage/geocommunity/presentations/SamuelJones.pdf
(Accessed 5 February 2014).

Zielstra D and Zipf A (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal.
Available at http://koenigstuhl.geog.uni-heidelberg.de/publications/2010/Zielstra/AGILE2010_Zielstra_Zipf_final5.pdf
(Accessed 18 April 2013).