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

R. Frew\textsuperscript{1}, G. Higgs\textsuperscript{1}, M. Langford\textsuperscript{1}, J. Harding\textsuperscript{2}

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

April 6, 2015

Summary
This paper addresses the comparative dearth of research on spatial data usability by employing sensitivity analyses to the findings from applying GIS-based accessibility models. Comparisons were made using approaches based on Euclidean distances and more sophisticated accessibility measures that utilise network travel distances: the latter incorporating measures of supply and demand by using innovative extensions to the Two-Step Floating Catchment Area method (2SFCA). To illustrate the sensitivity of findings from applying such models with a range of data sources, geographic accessibility to secondary schools was calculated for Output Areas in South Wales using a 2SFCA plug-in to ArcGIS\textsuperscript{TM}. By using different permutations of spatial data, for both the supply- and demand-side parameters in such models, differences in walking distances and 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 therefore an increased scrutiny as to the quality and usability of such data and the respective advantages and limitations of both proprietary and crowdsourced data.

Although the highest quality data often remains expensive to obtain for some users (for example high resolution LiDAR data, or Ordnance Survey MasterMap products), other data sets are becoming available without the need for expensive capital or revenue outlay. Recent reports (for example, Avery and Gittings, 2014) on the use of unmanned aerial vehicles to produce a variety of remotely-sensed data as well as the availability of various software solutions, both at low costs relative to traditionally-sourced equivalents are enabling new data-producers to emerge. Such trends are paralleled by 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, reinforcing earlier debates surrounding the use of VGI (volunteer geographic information) and previous 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) did investigate the impacts of different approaches to measuring accessibility to green space using comparable sources of data. Few studies to date incorporate sensitivity analysis that involve the use of different sources of spatial data applied to different stages of a ‘typical’ GIS project (although see Jones (2010) for 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 relatively routine 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 decision making processes. 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 include 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 sports facilities in Wales (Higgs et al, 2015).

2. Study approach

2.1 Study area

Two areas in South Wales were chosen for study: the city and county of Cardiff; and the neighbouring Vale of Glamorgan County. Cardiff is the largest city in Wales, although within its county boundary are villages located in 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 proximity to major transport links (the M4 motorway) and to large towns and cities (such as Bridgend and Cardiff).

2.2 Geographic data

The spatial data products chosen were typical of those commonly used in UK-based accessibility analysis studies, including Ordnance Survey MasterMap Integrated Transport Network™ (ITN) Layer and the additional Ordnance Survey Urban Path layer. OpenStreetMap (OSM) network data for South Wales was obtained from a third-party provider, as an example of the crowdsourced/VGI 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 from Ordnance Survey OpenData, was built into a network using standard, readily-available GIS tools (using Arc GIS). At the time there was no free-to-use network dataset using Ordnance Survey data, though in March 2015 Ordnance Survey launched the Open Road network dataset, also under their OpenData programme, and this will be subject to later analysis.

The accessibility assessment tasks were also conducted using Euclidean (straight line) distances.

The relative accessibility of different locations within the study areas were assessed using different methods, with the processes subjected to sensitivity analysis in order to identify areas of similarity and difference. Several iterations of each analytical task were therefore performed, using permutations of the different datasets.

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 features were secondary schools (with travel-to-school journeys being the focus of many studies relating to active travel and children’s health, child road safety, catchment areas, parental choice, etc). The location of such facilities is also 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 often occupy large sites, they are ideal for use in comparing different methods of representing polygons by points. The Ordnance Survey Points of Interest dataset was used as the initial 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 were compared: centroid (the geometric centroid of the entire site); access point (one or more way in to the school site for pedestrians); and boundary (any point on the perimeter of the site). Figure 1 illustrates how this choice may impact on travel distances, using a typical school site 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.
2.4 Location of demand

Various demand (population) representations are available to researchers. In this study, the method of locating the population included the use of 2011 UK Census Output Area (OA) population-weighted centroids. Other methods are available, both more detailed (at post code or address level, for example) or more generalised (for example, either of the census Super Output Area layers, both of which are aggregations of OAs). The method chosen uses readily-available and free-to-use data that is sufficiently detailed to allow differences to be identified between smaller areas while avoiding the increased computational loads and visualisation challenges resulting from the use of more detailed representations.

2.5 Methodology

This study calculated Euclidean and network distances with different permutations surrounding supply-side options. Comparisons were then made between the results of the various iterations of the accessibility models, both visually and statistically. One indication of similarity was achieved through comparison of Destination Overlap, where the identities of the supply facility nearest to each demand centre was compared for each network option (see Table 1).

Euclidean distance ignores the actual travel route taken, and as with network distances takes no account 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 developed by researchers at the University of South Wales was used in ArcGIS™ (Figure 3 shows the user interface of the 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 published 2011 census data. Although there was no convenient category of “secondary school age” in the census, there was information on 12 to 16 year olds, and an estimate was made of the numbers of students at school outside these age categories.

Figure 3 Illustration of the E2SFCA plug-in tool first screen. Further screens offer the options of incorporating levels of supply and demand.
3. Preliminary Findings

Patterns of “worst” and “best” accessibility were broadly similar, but with differences stemming from the method of locating the supply feature (as in Figure 4) and the choice of network dataset.

Figure 4  Examples of variations in distances from OAs to their nearest facility, depending on the method used to locate that facility (in this case, secondary schools in the Vale of Glamorgan).

(a) OS Sites dataset, OA to polygon centroid, Euclidean distance;
(b) OS Sites dataset, OA to nearest access point, Euclidean distance;
(c) OS Sites dataset, OA to nearest point on site boundary, Euclidean distance.
Differences in findings when applying alternative methods of measurement (i.e. between distance measures and the results of E2SFCA calculations, examples of which are shown in Figure 5) were evident, highlighting the impacts of using different approaches on the results from GIS-based models. Figure 5(b) illustrates the effect of supply and demand on accessibility, and the influence of the size of catchment area used in E2SFCA calculations. Preliminary findings also indicate urban/rural differences which also merit further investigation, an example of such differences is shown in the Distance Overlap figures of Table 1.

In contrast, other findings suggest that the use of different datasets or different network products may make no statistical difference to outcomes either in terms of distance or E2SFCA scores. All distance results and all E2SFCA results were significantly correlated (using Spearman’s rank correlation) and Table 2 shows Wilcoxon Z scores for E2SFCA results from the Vale. Several paired comparisons (18%) have no significant differences between their scores at the 5% level, though the majority of comparisons (76%) have differences that are significant at the < .001 level.

(a) Distance measures for OAs to school centroids, ITN network.
In rural areas such as the Vale, with relatively few facilities, a large number of OAs (150 out of 412) are in the “worst” distance category.

(b) 2SFCA for OAs to school centroids, ITN network.
Direct comparison with Distance measures are difficult, due to the different scales used. Results here use rounded quintile splits, with lowest E2SFCA score equating to “worst” accessibility.

Figure 5 Example of differences in accessibility visualisations obtained when using ITN, comparing distance and E2SFCA measures: (a) Distance; (b) E2SFCA.
As part of this PhD research, further analysis of the results will be conducted with the intention of isolating the factors within the underlying data that contributed to the differences in accessibility scores within the same output areas. However, the extent of differences identified in this preliminary analysis leads us to suggest that researchers need to be made more aware of the implications of using different sources of data in ‘typical’ GIS tasks.

Practical issues with different datasets (for example cost, ease of download, availability, etc) along with their currency and update patterns, are also worthy of further study. Identifying methods whereby the usability of these spatial data sources can be made more transparent to researchers prior to their implementation in GIS analytical tasks is one of the ultimate aims of this research.

### Table 1 Destination overlap (using distances), comparing Cardiff and Vale Secondary Schools.
Out of 105 comparisons, Vale had higher figures in 86 (82%). The 19 comparisons in which Cardiff figures were higher are highlighted for ease of identification.

<table>
<thead>
<tr>
<th></th>
<th>Euclidean</th>
<th>ITN</th>
<th>UP</th>
<th>OSM</th>
<th>VMD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cents AccPts Cents AccPts Cents AccPts Cents AccPts Cents AccPts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cents</td>
<td>87.6 86.4 81.6 84.2 84.7 81.1 89.0 84.0 84.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AccPts</td>
<td>86.9 84.0 83.0 92.0 90.2 93.2 93.5 86.5 85.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perim</td>
<td>84.9 83.7 83.7 87.4 84.2 83.3 86.6 86.7 86.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITN</td>
<td>82.0 84.0 84.0 87.4 87.4 83.3 83.3 82.5 82.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AccPts</td>
<td>76.5 78.4 78.4 81.2 80.7 81.2 82.5 82.5 82.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perim</td>
<td>72.8 78.9 78.9 79.7 77.2 78.9 81.1 81.1 81.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UP</td>
<td>89.2 84.2 84.2 89.0 89.0 84.2 85.0 85.0 85.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cents</td>
<td>88.6 81.5 81.5 89.3 89.3 84.2 85.5 85.5 85.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AccPts</td>
<td>89.0 81.5 81.5 89.3 89.3 84.2 85.5 85.5 85.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perim</td>
<td>81.1 74.9 74.9 81.5 81.5 84.2 82.2 82.2 82.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMD</td>
<td>57.0 67.7 67.7 66.7 66.7 66.7 66.7 66.7 66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cents</td>
<td>62.7 67.7 67.7 66.7 66.7 66.7 66.7 66.7 66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AccPts</td>
<td>60.2 67.7 67.7 66.7 66.7 66.7 66.7 66.7 66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perim</td>
<td>62.7 67.7 67.7 66.7 66.7 66.7 66.7 66.7 66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2 Differences in E2SFCA scores between networks and location method for Vale Secondary Schools. Wilcoxon Z scores are shown below diagonal, statistical significance above diagonal (black = sig at < .001 level; green = sig at .01 level; amber = sig at .05 level; red = not significant at .05 level).
4. Acknowledgements

The study reported here forms part of an Ordnance Survey-sponsored PhD research programme. The plug-in to ArcGIS™ was developed during Phase 1 of the WISERD project funded by the ESRC and HEFCW (ESRC Grant Reference: RES-576-25-0021). However any views expressed herein do not necessarily represent those of these organisations.

5. Biography

Robin Frew is a final year PhD student at the University of South Wales.

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


