Data usability from an end-user perspective: assessing contextual quality through geospatial analysis

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

[1]Faculty of Computing, Engineering and Science, University of South Wales, Treforest, Pontypridd CF37 1DL
Presenter's details:
Tel: (044) 01443 483268
robin.frew@southwales.ac.uk
http://gis.research.southwales.ac.uk

[2]Ordnance Survey, Explorer House, Adanac Drive, Southampton SO16 0AS

1. Introduction

The availability of geographical data has increased hugely in recent years, partly due to web-based developments (such as Google Earth) and crowdsourced mapping products (such as OpenStreetMap). Although there has been considerable research in the field of GIS-based usability, the majority of studies to date approach usability from the perspective of software development in areas such as computer interface design and testing, visualisation and cognition, and in aspects of device design (Hunter et al, 2003). Much of the literature on the usability of spatial data has been concerned with conceptual or theoretical frameworks in relation to concepts such as 'fitness for purpose' (Josselin, 2003 and Wachowicz and Hunter, 2003). There has been some recent research concerned with applying usability concepts to real-life applications (e.g. Brown et al, 2012) and with examining the implications of data quality in relation to the application of crowd sourced data (Haklay, 2010). However, very few studies to date have been concerned with evaluating the use and quality of different sources of spatially referenced data in relation to specific GIS-based tasks. With the ongoing trend towards the use of open source GIS, and with the increasing amounts of freely available data through initiatives such as data.gov.uk, there is an urgent need to examine the usability of data sources in different contexts which draw on their application in a range of GIS-based analytical tasks.

This presentation will explore the type of factors that impact on the appropriateness and suitability of spatially referenced datasets in 'typical' GIS-based tasks. In particular the focus here is on their application in accessibility studies (using the example of access to primary schools) as a precursor to developing usability metrics which can be used to gauge the usefulness of spatial data in different contexts.

2. Study Approach and Findings

Three contrasting areas in south Wales were chosen study areas: the city and county of Cardiff; the town of Pontypridd in the South Wales Valleys; and a rural sample area in the Vale of Glamorgan (Figure 1). The study draws on five sources of spatial data which have been used to represent network based presentations in UK-based studies that have focused on accessibility analysis, namely:

Ordnance Survey's OS MasterMap® Integrated Transport Network™ (ITN) Layer and Urban Path layer; OS VectorMap® Local; OS VectorMap® District; and OpenStreetMap (OSM). The ITN products were designed by Ordnance Survey (OS) specifically for use as network products. Standard VectorMap products were obtained and subsequently built into networks for the purpose of this study, using standard Arc network-building tools, with no attempt to amend or quality-check prior to their use in measuring different facets of accessibility. The OSM data for South Wales was obtained from a third-party provider, Metro Extracts (http://metro.teczno.com/ (accessed 12 November 2013)). The network was built in ArcGIS through a free, open-source add-on for ArcGIS[1].

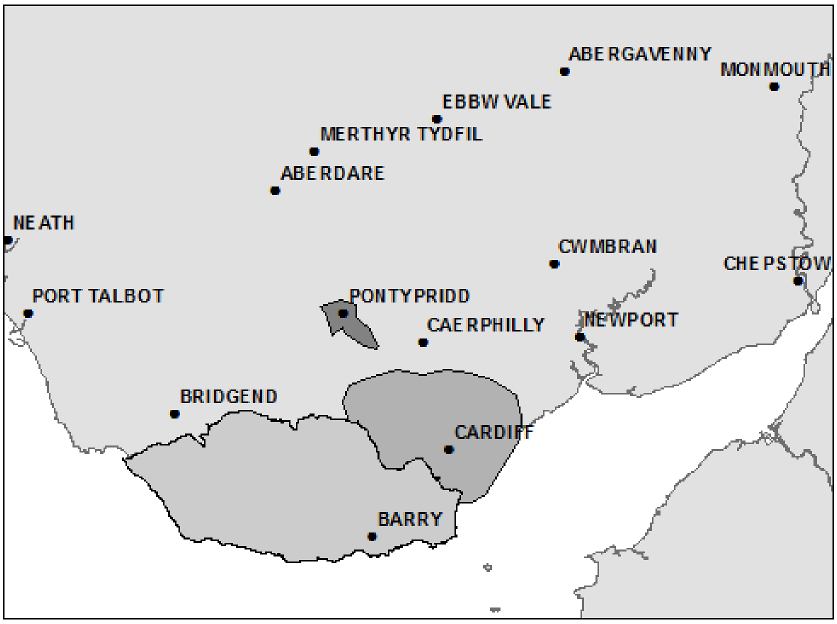

Figure 1. South Wales, showing the three study areas.

For all exercises, 2011 Census Output Area (OA) population-weighted centroids were used to represent the location of population demand in each sample area, and destination locations were taken from Point of Interest (PoI) data, provided by PointX.

Analysis was carried out in Esri ArcMap 10.2 using the Network Analyst extension.

Methods used to measure completeness drew on the work of Haklay (2010) and Zielstra and Zipf (2010). The total network of roads within each sample area was measured, and when assessed against that of ITN with Urban Paths (see Figure 2) was found to differ by up to 34% in Cardiff (VectorMap District returning the lowest figure), 30% in the Vale of Glamorgan (VectorMap District being lowest, by a small margin), and by 27% in Pontypridd (with OSM marginally lower than the VectorMap District figure).

UK Government figures (Dept of Transport, 2012) stated the average trip length to a primary school was 1.8 miles (2.9km) in 2012; and that only 9% of households did not have a primary school within a 15 minute walk. Figures from the Welsh Government suggest that 3 miles (5km) was considered as a reasonable distance for a child to cycle to primary school (Statistics for Wales, 2012). These averages were compared with the situation in Pontypridd, Cardiff, and the Vale of Glamorgan using these different sources of spatial data. Various methods were used to assess travel distance and travel time to the nearest primary schools to examine whether policies promoting walking or cycling for pupils were starting from a

[1] http://esriosmeditor.codeplex.com/wikipage?title=Create%20a%20Network%20Dataset%20from%20OSM%20Data

disadvantageous position, on the basis that areas with better than average accessibility had a greater opportunity to reduce car use and increase walking and cycling, whereas those with worse than average accessibility may require identification in order to target further initiatives or resources.

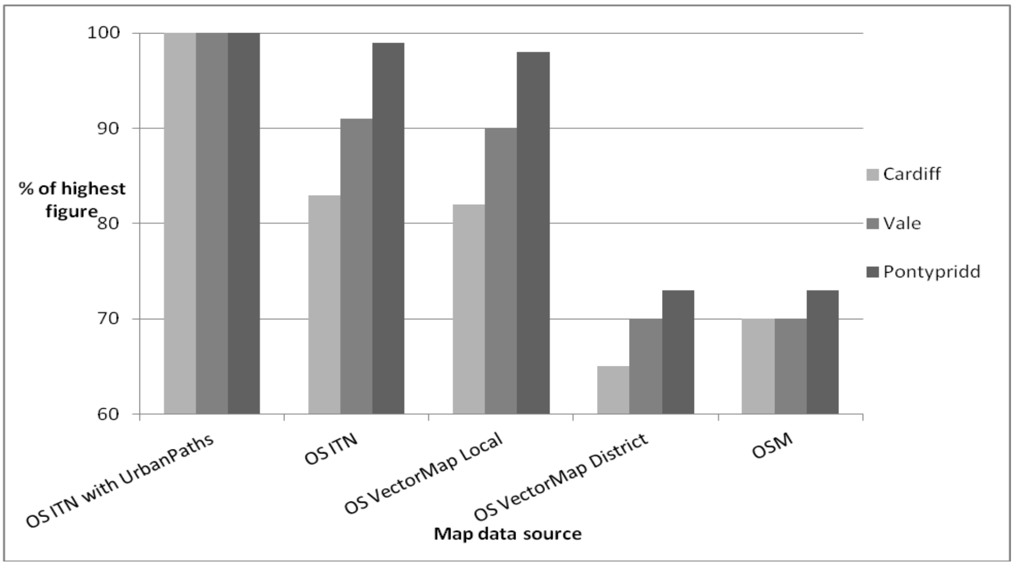

Figure 2. Comparison of total network lengths.

ArcMap Network Analyst OD Cost Matrix maps and Service Area maps were produced for the primary schools within the three study areas. In the absence of school roll data the assumption here is that parents send their children to the nearest primary school, regardless of religious denomination or language (in Wales there are three different state school options: standard, Welsh language, or church).

Figure 3 shows an OD Cost Matrix output for Pontypridd with the shortest walking distance to each destination (the nearest primary school) represented by the black lines from each origin (OA centroid) with the results calculated as network distances or travel times.

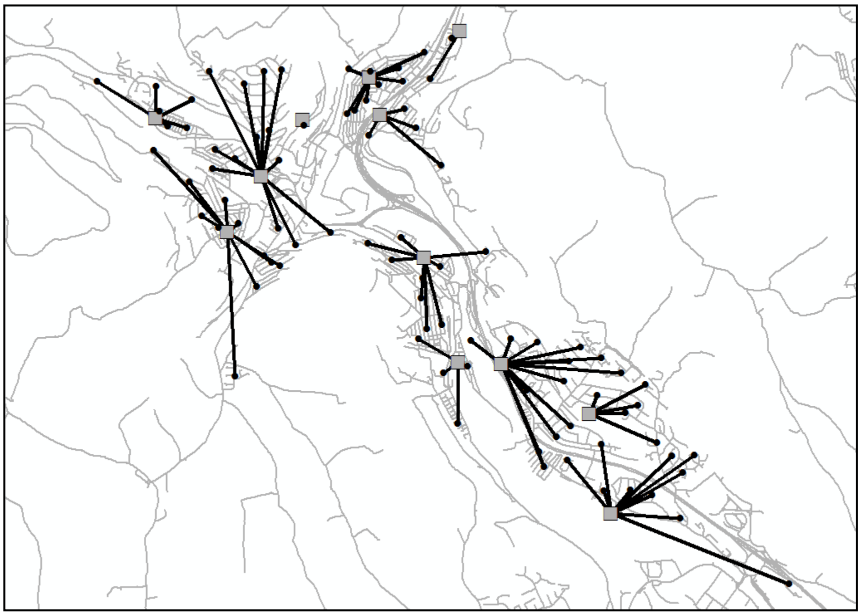

Figure 3. Typical OD Cost Matrix output showing Pontypridd OA centroids (black dots) and their nearest primary schools (grey squares).

The average distance to the nearest primary school ranged from 565m to 690m in Pontypridd (depending on network data set), from 627m to 740m in Cardiff, and from 838m to 970m in the Vale of Glamorgan (against a national average of 2.9km).

Figure 4 illustrates the result for Pontypridd, as an example, showing the mean distance to the nearest primary school for each of the five datasets. The data in each set of results was not normally distributed (following Shapiro-Wilk tests), and the Friedman test indicated there significant differences were present, depending on the map data used. Wilcoxon tests found that ITN with Urban Paths was significantly shorter (at the 1% significance level) than ITN, OSM and VectorMap Local and, at the 5% significance level, shorter than VectorMap District. The difference between VectorMap District and VectorMap Local was also found to be significant at the 1% level, while OSM was found to differ from VectorMap Local at the 5% significance level. The differences between the other possible combinations of these datasets were found to be not statistically significant.

Further statistical tests will be carried out on the equivalent data for urban and rural contexts, using data from Cardiff and the Vale of Glamorgan, respectively.

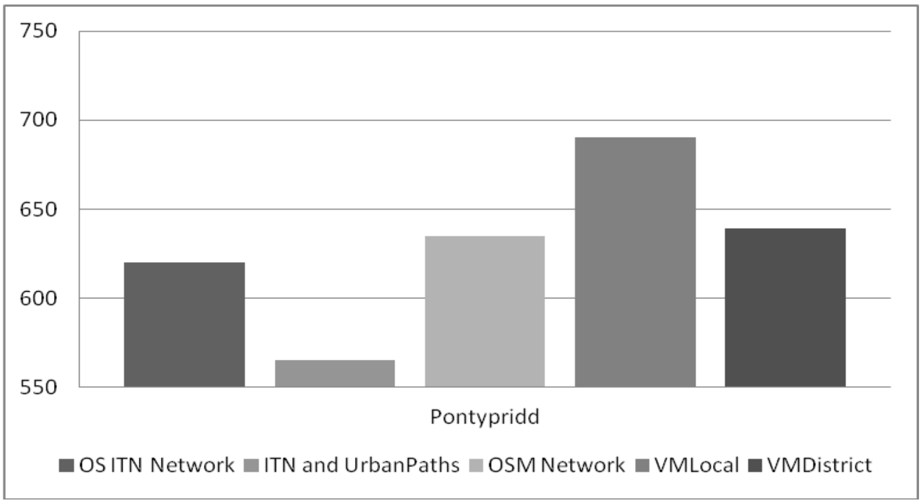

Figure 4. Mean distance (m) from population-weighted OA centroids to primary schools in Pontypridd.

The choice of network dataset not only influences the magnitude of the journey, but also the feature designated as the 'nearest' school (as Figure 5 illustrates). Around 19 to 23% of origins in Pontypridd had alternative choice of designations, depending on the network dataset used. Variations in Cardiff ranged from 15% to 20%, while those in the Vale of Glamorgan varied from 12 to 17% (as shown in Figure 6). A wider disparity may have been expected between the results for urban and rural location, but was not apparent from these results.

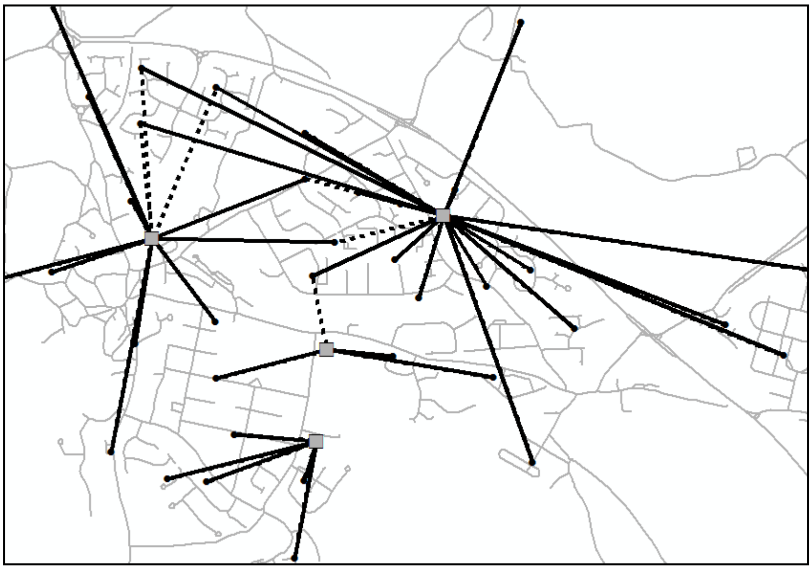

Figure 5. Example from Vale of Glamorgan of different destinations chosen depending on network used (in this case comparing the OD lines produced from ITN and ITN with Urban Paths, overlaid on the ITN network).

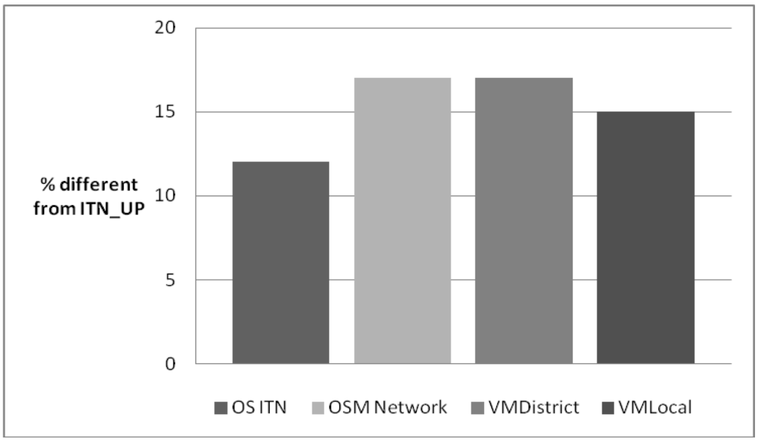

Figure 6. Variation in destination – Vale of Glamorgan. The pattern of results is similar to that of the other study areas.

Service Areas were produced, using Arc Network Analyst, to calculate walking times to primary schools, once again using census OA population-weighted centroids to represent population. The results for Pontypridd (Figure 7) show a typical output, with the largest polygon representing the 30-minute walking time from primary schools, and including all OA centroids from this study area within its range.

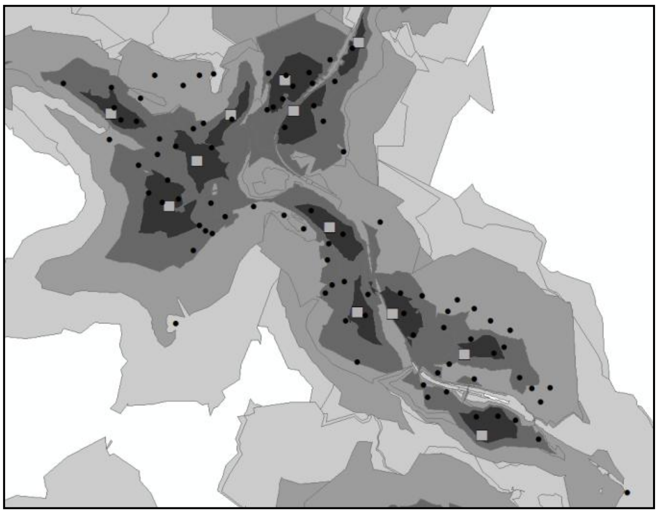

Figure 7. Service Area polygons for Pontypridd, with schools (grey squares) and OAs (black dots), showing walking times, up to 30 minutes, on ITN.

Service Area polygons were used to identify the population within a 15-minute walk of their nearest primary school in each of the three areas being studied. Note that if a constant walking time is assumed, then walking time can equate to distance, providing topographical factors (grade, etc) and mobility factors (age, disability, etc) are ignored. Figure 8 shows the results.

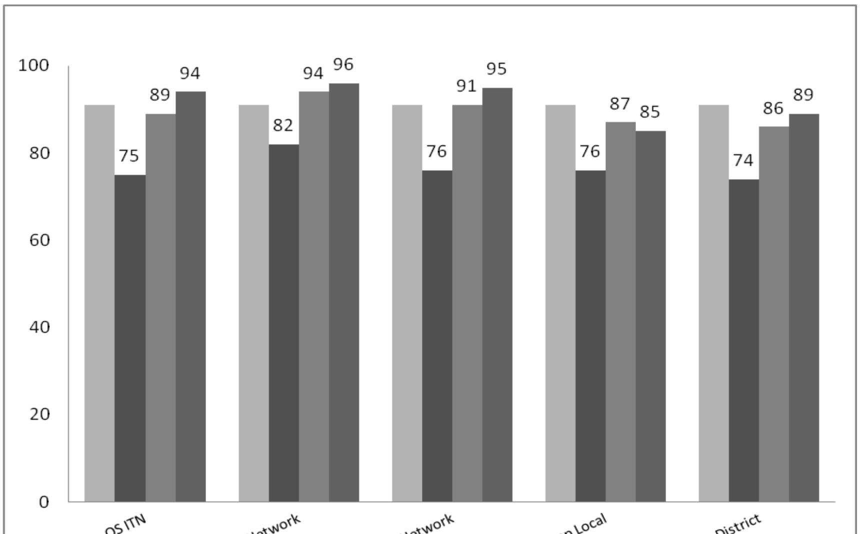

Figure 8. Percentage of population in the study areas within a 15-minute walk of their nearest primary school, compared to the national average.

When walking times are compared, the similarities are striking, though there are differences, an example of which is shown in Figure 9. Note that the OSM data is hidden by the ITN_UP line.

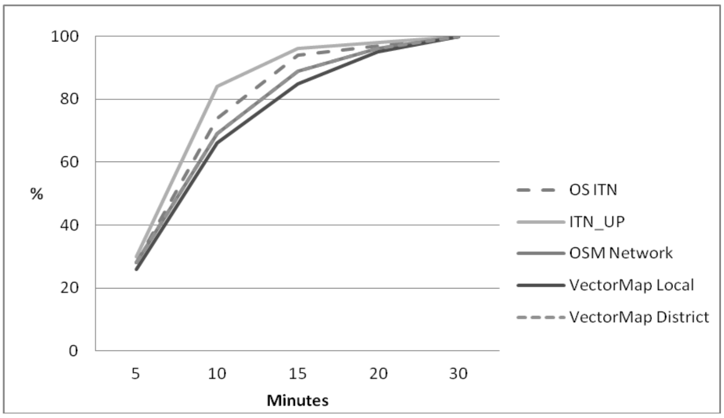

Figure 9. Walking times to nearest primary school – Pontypridd.

3. Discussion

The use of alternative network data sets demonstrate considerable differences in findings from the GIS-based accessibility conducted to date, which will be expanded on in the presentation and which will impact on the usability of these data sets. The aim here has been to apply these datasets in three different geographical settings, and these differences will be explored in the presentation. The two OS VectorMap products (one currently open, one available for purchase), produce similar results to OSM despite not being specifically designed for network analysis. This, similarly, applies to OSM, though networking for travel directions and full network ability seems to be seen as one of its aims OpenStreetMap, 2014).

Both Haklay (2010) and Zielstra and Zipf (2010) noted how OpenStreetMap coverage drops in terms of quality and quantity outside major urban areas (such as the ‘big 5' cities of the UK, or around the larger cities of Germany), with many other areas simply "not covered very well,” (Haklay, 2010, Section 4.3). With Cardiff being the largest city in Wales it may have been thought that coverage would be higher than the 70% calculated here. Haklay found OSM’s road length was 69% of OS Meridian. As Meridian is a generalised data set, perhaps OSM having 70% of the more detailed network of ITN with Urban Path shows a relatively high coverage, though research into OSM progress over the UK is lacking. Further, more detailed, findings will be presented during the course of the presentation including a statistical comparison of variations for the three study areas.

4. Future work

This study forms part of an Ordnance Survey-sponsored PhD research project which is examining the factors that impact on the appropriateness and suitability of various spatially referenced data for a range of typical GIS-based tasks. A comparison of spatial datasets is provided for a range of Ordnance Survey data products plus some broadly-equivalent third party datasets (including a crowd-sourced dataset). Processes involve a range of typical spatial-analytical operations such as the computation of straight-line and network distances, the evaluation of spatial intersection and containment and adjacency between features. These processes are applied at a variety of spatial scales from amongst the input datasets, and stored in various elemental object types (points, lines, polygons, grids, etc.) to enable contrasts and comparisons to be made. To date the project has compared the use of these spatial data sets in accessibility-based analysis using a range of supply-side features in various geographic contexts.

In future work, features and analytical methods will be chosen in order to compare the 'performance' of different databases in a variety of GIS-based tasks. In addition, different representations of population will also be assessed, with the computational load of OA centroids being considered against less detailed representations (such as Lower or Super OAs). On the other hand, the advantages of using more locally-precise representations will also be considered against the disadvantage of greater processing times.

Comparison of the data when used with commercial GIS such as ArcGIS, with open source GIS such as QGIS, could also produce interesting results, perhaps indicating the usefulness of data depends not only on what is done with it in which context, but also on the tools used to analyse them.

5. References

BROWN, M., SHARPLES, S. and HARDING, J., 2013, Introducing PEGI: a usability process for the practical evaluation of geographic information. *International Journal of Human-Computer Studies.* **71**(6), pp. 668-678.

DEPARTMENT OF TRANSPORT, 2012, National Travel Survey. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/243957/nts2012-01.pdf (Accessed 6 December 2013).

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**, pp. 682-703.

HUNTER, G., WACHOWICZ, M., and BREGT, A., 2003, Understanding spatial data usability. *Data Science Journal,* **2**, pp. 79-89.

JOSSELIN, D., 2003, Spatial data exploratory analysis and usability. *Data Science Journal (Spatial Data Usability Special Section),* **2**, pp. 100-116.

OPENSTREETMAP, 2014, Mapping projects wiki page. Available at: http://wiki.openstreetmap.org/wiki/Mapping_projects (Accessed 14 March 2014).

STATISTICS FOR WALES, 2012, Monitoring "A Walking & Cycling Action Plan for Wales 2009 – 2013": Interim update. *Statistical Bulletin SB 107/2012*. Available at: http://wales.gov.uk/docs/statistics/2012/121114sb1072012en.pdf (Accessed 11 March 2014)

WACHOWICZ, M. and HUNTER, G., 2003, Spatial Data Usability. *Data Science Journal, preface to Spatial Data Usability Special Section*, **2**, (26).

WELSH ASSEMBLY GOVERNMENT, 2009, National Travel Survey 2008. Available at: http://wales.gov.uk/docs/statistics/2008/080522sb272008en.pdf (Accessed 6 December 13).

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).

6. 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.

7. Biographical notes

Robin Frew is a postgraduate student at the University of South Wales, and has recently completed the first year of his PhD research project on 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.