Assessing geographic data usability in analytical contexts by using sensitivity analyses of geospatial processes.

Robin Frew

University of South Wales

The number and variety of sources of spatial data continues to expand, as do the debates regarding the quality and usability of such data, particularly those which are Free and Open Source (FOS) or free-to-use. The highest quality data is often expensive to obtain and the option of cost-free data sets is tempting for many users.

With the existence of the huge hinterland of data *quality* research acknowledged, and a great number of studies investigating the usability of devices and interfaces, little attention has been paid to the *usability* of data, and even less into the usability of geographical data in typical GIS research situations. There has been some research into the use of volunteered geographic information (VGI) in the field of data quality theory and assessment (see for example Haklay, 2010; Zielstra and Zipf, 2010), but relatively few studies have incorporated sensitivity analysis involving the application of different sources of spatial data to a range of GIS tasks. Jones's (2010) study into the use of open data in presenting and visualising public health information is one notable exception, with another being that of Higgs et al's (2012) examination of the impact of alternative approaches to measuring accessibility to green space.

This study set out to address cross-cutting themes that are topical in GIS and geographical analysis given trends towards the use of open source data, namely:

- Do different methods of representing real-world features have an effect on the findings from GIS analyses?
- To what extent does choice of data sources affect network analysis?
- In considering network accessibility, are results affected by the representation of supply and demand considerations?

 In: A. Comber, B. Bucher, S. Ivanovic (eds.): Proceedings of the 3rd AGILE PhD School, Champs sur Marne, France, 15-17-September-2015, published at http://ceur-ws.org

There is little evidence to date on which to quantify the effects of these issues on final results. This research is intended to take a step in redressing gaps that exist in the knowledge, understanding and perception of such data.

This study argues that even the best quality data may not be appropriate in certain contexts. To highlight the type of scenarios where this may indeed be the case several commercial and free-to-use data sources were used in sensitivity analyses of the application of well-established GIS network analysis tasks. The aim is to assess whether findings vary according to the application of alternate data sets used to represent the same features within such models.

The research took the form of various case studies, all tied around a similar theme, that of accessibility. Some of the studies assessed accessibility to features that have been the subject of much research in the past (such as GP surgeries), while some looked at less commonly assessed features (such as primary schools, secondary schools and sports facilities). All were linked by an interest in various health and fitness initiatives and investigations that have taken place in South Wales (UK), such as those looking at active travel to schools, equitable access to health care and reasonable geographical access to sport and leisure facilities.

The part of the study relating to accessibility to primary schools will be used as an example.

The supply feature (primary schools) were represented in four different ways by the two datasets examined: a Point of Interest[1] point (nominally the centroid of the main school building); the pedestrian access points of each school; the geometric centroid of the entire school site (including play areas, sports fields and car parks); and the site perimeter. The Ordnance Survey Sites dataset[2], by providing the footprint of each school as well as the access points, offered more detail and precision to measurements of access, raising another interesting question as to whether any increase in precision automatically resulted in an increased accuracy of results.

The places of origin for journeys to the schools were kept constant, and were UK census Output Area population-weighted centroids (the smallest unit of published UK census data).

Distances from each population centroid were measured to the nearest school, looking at each representation in turn, using the various network datasets. The network datasets included commercial data (Ordnance Survey ITN and ITN with Urban Paths[3]), free-to-use data (Ordnance Survey OpenRoads[4]) and FOS data

[1] https://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html

[2]https://www.ordnancesurvey.co.uk/business-and-government/products/topography-layer.html

[3]https://www.ordnancesurvey.co.uk/business-and-government/products/itn-layer.html

(OpenStreetMap[5]). Sensitivity analysis was conducted through repetitions of the distance calculations, ensuring every combination of network (plus Euclidean measurement) was used for every feature representation. The process was then repeated in its entirety using a Two-Step Floating Catchment Area (2SFCA) measurement. As described by Luo and Wang (2003), 2SFCA incorporates levels of supply and demand by calculating population-to-provider ratios for each supply centre within a defined threshold distance, then identifying all those supply centres within the same threshold distance of each demand centre, and summing all their ratios for each population. Supply was represented by the student capacity of each school (the school 'roll', from figures published by the local authority). Demand was represented by the number of primary school-age children in each census area (as extracted from published 2011 census data).

The large number of results generated were cross-compared. The comparison revealed that for primary schools the vast majority of results (over 80% of all comparisons) were statistically significantly different from the others at the < .001 level, for both distance and 2SFCA measures. This indicated that the different datasets used were not interchangeable and therefore not equally usable in this type of study.

At this early stage of analysis initial indications were that differences between the network datasets had the greater effect on results. Differences due to method of demand- or supply-side feature representation were less important.

Initial findings suggest that more attention needs to be given to the nature of data sets used to represent such features in GIS-based analytical tasks. The exact context in which such data sets are applied may determine how usable different sources of data are in relation to common GIS spatial analytical tasks and a useful addition to GIS-based analysis going forward could be the derivation of a typology of circumstances in which adopting alternative sources of open data are more appropriate.

Acknowledgments

This research was funded by Ordnance Survey (OS) but any interpretations of findings are those of the student and do not necessarily reflect the opinions of OS.

---

[4]https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-roads.html

[5]http://www.openstreetmap.org/

References

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.

Higgs, G., Fry, R. and 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, pp. 326-343.

Jones, S. (2010) ‘Open geographical data, visualisation and dissemination in public health information’, *AGI Geocommunity 2010* [Online]. Available at: http://www.agi.org.uk/storage/geocommunity/presentations/SamuelJones.pdf

(Accessed: 5 February 2014).

Luo, W., Wang, F., (2003) 'Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region',

Environment and Planning B: Planning and Design, 30, pp. 865–884.

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. [Online]. Available at http://agile2010.dsi.uminho.pt/pen/ shortpapers_pdf/142_doc.pdf (Accessed: 18 April 2013).