Barriers to reproducible research in GIS; a practical example using a national index of multiple deprivation

Nicholas Page*[1], Mitchel Langford†[1], and Gary Higgs‡[1]

[1]Wales Institute of Social and Economic Research, Data and Methods (WISERD) and GIS Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL

January 12, 2018

Summary

There are growing calls for more reproducible research within GIS science. This is principally due to concerns that many studies lack methodological transparency owing to unclear computational methods, the use of closed-source analytical software, and because the provenance and availability of data are often unclear. Applying a practical example, this paper highlights the problems faced when attempting to reproduce the geographical access to services domain of a national index of multiple deprivation to examine the effects of substituting the original accessibility metric with a more spatially advanced approach based on 'floating catchment area' (FCA) techniques.

KEYWORDS: Reproducible Research; GIS; Accessibility; Two-Step Floating Catchment Area (2SFCA); Indices of Multiple Deprivation

1. Introduction

Ensuring that research is wholly reproducible, where possible, has been suggested as a "reasonable goal" for geocomputational and spatial analytical studies (Brunsdon and Singleton, 2015, p.254). In short, the primary aim of reproducible research is to make possible the exact replication of results by providing explicit detail of both data and method – generally, this could include, for example, the use of open-source software, providing detailed transcripts of computational methods, and storing all metadata in a freely accessible data repository. In this paper, we examine the feasibility of replicating, from published sources, the geographical access to services domain of the 2014 Welsh Index of Multiple Deprivation (WIMD-2014), as part of a broader study investigating the potential implications of substituting the existing approach to accessibility measurement used within WIMD-2014 with a more sophisticated approach based on floating catchment area (FCA) methods.

2. Background

WIMD-2014 measures relative levels of disadvantage between Welsh lower layer super output areas (LSOAs), and is so named because it captures multiple aspects of both material and social deprivation. As its name implies, the geographical access to services domain (hereafter 'access domain') measures a community's ability to access a subset of essential services based upon a measure of geographical proximity between service and user. The existing accessibility metric is network-based and created using service-weighted average travel times from each demand point (1.3 million Welsh residential dwellings) to the nearest available supply point by public and private transport (ibid). A limitation of this approach is that only physical barriers to access are considered within the accessibility calculation, whilst other potential mediatory factors, such as levels of service demand, for example, are neglected. A derivative of the gravity model, approaches based on FCA techniques, in contrast, account for

* nicholas.page@southwales.ac.uk

† mitchel.langford@southwales.ac.uk

‡ gary.higgs@southwales.ac.uk

interactions between service supply and potential demand and are becoming increasingly adopted in health-related studies (Langford et al., 2016, Bauer et al., 2017, Higgs et al., 2017). The FCA score represents the relative share that a person has of the total service capacity available to them within reasonable proximity of their demand centre - usually a residential address or a population-weighted centroid (Luo and Wang, 2003; Luo and Qu, 2009).

3. An enhanced two-step FCA approach (E2SFCA)

The E2SFCA specification used in the current study consists of two computational steps:

In step one, a supply-to-demand ratio is calculated using supply volume, S_j and the sum of all demand centre populations, P_k that are contained within a defined time (or distance) threshold, d_0. W_{kj} is a linear weighting function based on the distance between supply and demand (equations 1 and 2).

$$R_j = \frac{S_j}{\sum_{k \in \left(d_{kj} \leq d_0\right)} P_k W_{kj}} \tag{1}$$

$$W_{kj} = \frac{(d_0 - d_{kj})}{d_0} \quad \mathit{if} \quad d_{kj} \leq d_0 \tag{2}$$

$$W_{kj} = 0 \qquad \mathit{otherwise}$$

In step two, for each demand centre, k, an E2SFCA score is calculated by summing all supply-to-demand ratios located within the defined time (or distance) threshold, d_0, which is again subject to the weighting function, W_{kj} (equation 3).

$$A_k = \sum_{j \in \left(d_{kj} \leq d_0\right)} R_j \, W_{kj} \tag{3}$$

4. Data and methods

As suggested, for research to be considered truly reproducible any metadata must be open-source or be fully replicable from published sources. In the case of WIMD-2014, the supply-side dataset used to calculate levels of access was neither freely accessible nor available upon request. This necessitated the creation of a replicate dataset of ‘best-fit’ service locations using data from available sources and drawing on vague variable descriptions reported within the WIMD-2014 technical guidance – for example, a ‘food shop’ was simply described as any store where break and milk is purchaseable. Although data sources and service counts were reported, enabling a comparison between the original metadata and our replica data, the lack of data transparency meant discrepancies were not only expected but wholly unavoidable (Table 1). Some minor disparity would be expected between the original supply-side dataset and our replicate dataset owing to differences in data collection periods, as well as the inclusion within the former (but not the latter) of services located on the English side of the Wales-England border. However, despite the best efforts of the research team, no explanation can be provided for the substantial variation in GP surgeries. Population counts used to adjudge for potential demand volume within the E2SFCA calculation were downloaded freely from Nomis (http://www.nomisweb.co.uk) and population-weighted centroids were obtained from the Office for National Statistics (ONS) Open Geography Portal (http://geoportal.statistics.gov.uk/).

Service	WIMD-2014 dataset *	Replicate dataset **
Food shop	2,656[a]	3,444[a]
GP surgery	983[b]	453[d]
Leisure centre	319[c]	195[c]
Petrol station	860[a]	540[a]
Pharmacy	819[bc]	714[e]
Post office	1,000[a]	807[a]
Primary school	1,522[c]	1,093[d]
Public library	297[c]	207[a]
Secondary school	258[ad]	205[d]

Table 1 Comparison of data points reported in WIMD-2014 Technical Report and replicate dataset. Source: [a] OS Points of Interest; [b] NHS Wales Directory; [c] Local Authorities; [d] Welsh Government; [e] Extracted from NHS sources. * data collected 2013-14; ** data collected 2015-17

A summation of the GIS computational methods of the depth required to accurately reproduce the access domain of WIMD-2014 was also not available; again, nor was it available upon request. Principally, this is because the analysis was externally commissioned by Welsh Government. Whilst some basic information on the GIS parameters used in WIMD-2014 to estimate both public and private transportation times are available, such as assumed travel speeds, for example (Welsh Government 2014, p52/53), this information alone was not sufficient enough for the authors to be confident of an exact replication of the WIMD-2014 network model. Service-specific E2SFCA scores were therefore computed using a road network model based on Ordnance Survey ITN layer data (Ordnance Survey, 2015) and the Network Analyst extension in ArcGIS (ESRI, 2015). Once estimated, all service-specific E2SFCA scores were weighted based on published weighting factors (Welsh Government, 2014) and combined into a single composite access measure with one score per LSOA.

5. Results and discussion

This study attempted to reproduce the access domain of WIMD-2014 in order to explore prospective implications of incorporating a more spatially sophisticated, supply-to-demand based approach to accessibility measurement into an existing index of multiple deprivation. Comparisons are shown in Figure 1 (representing existing patterns based on proximity) and Figure 2 (representing revised patterns based on E2SFCA scores) respectively. In summary, existing patterns of access-deprivation in Wales reflect clear urban-rural gradients, with greater levels of access identified in and around major urban areas due to the closer proximity of services to demand centres. In contrast, accounting for levels of potential demand and distance-decay effects in an E2SFCA approach generated a more diverse picture of levels of access-deprivation, with many urban areas obtaining lower access scores due to higher potential service demand in these areas.

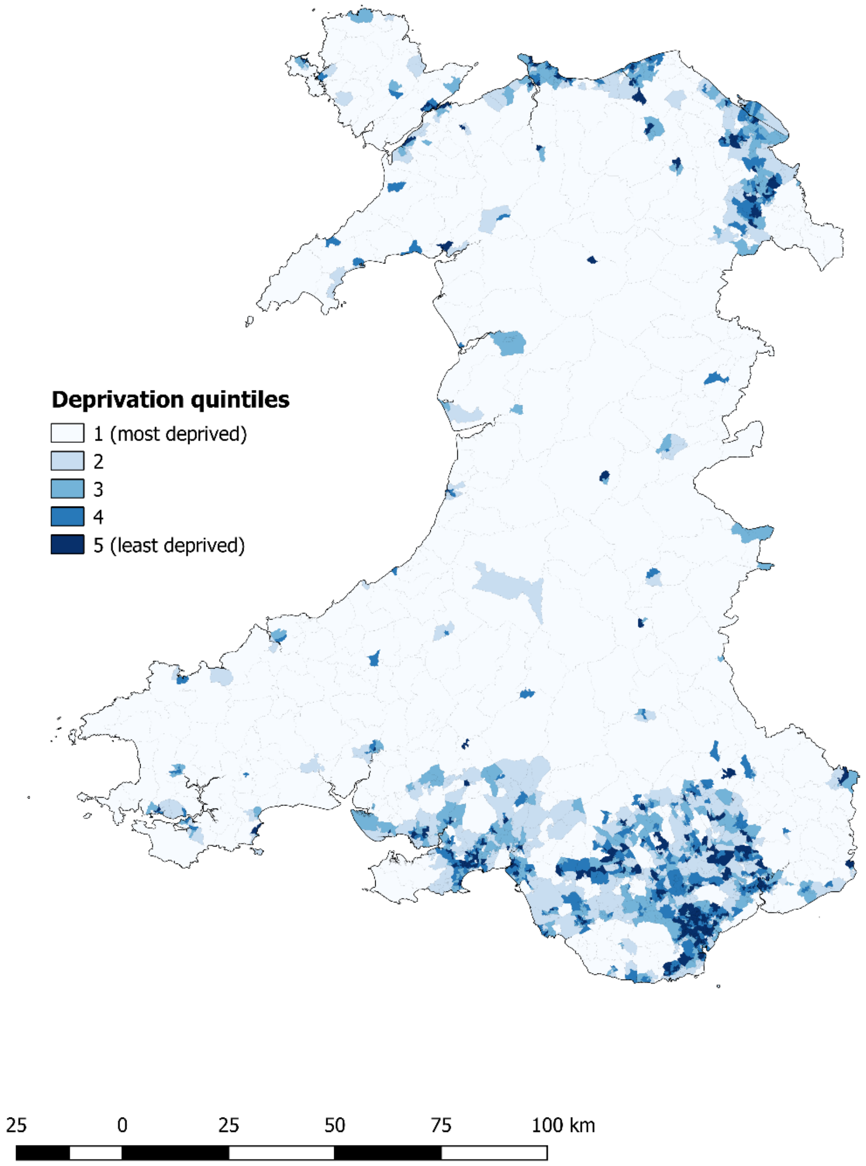

Figure 1 Access-deprivation quintiles based on existing WIMD-2014 scores

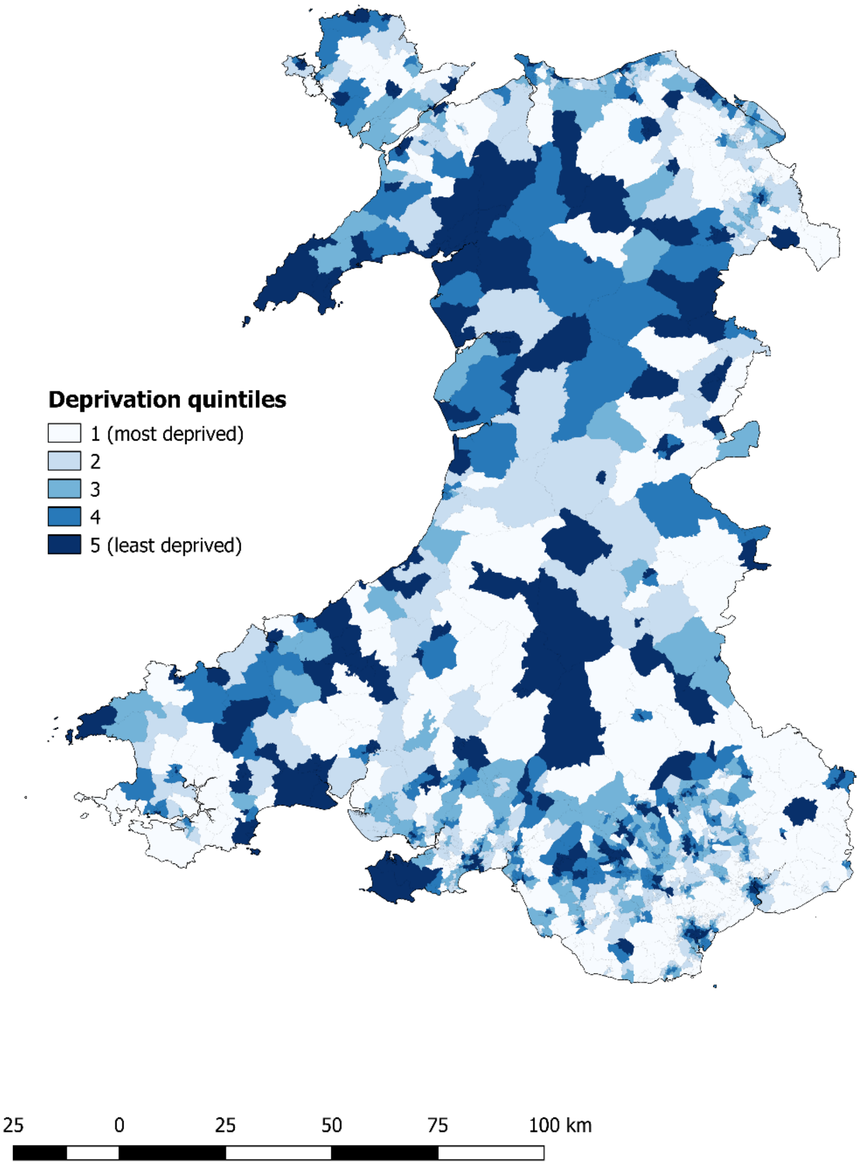

Figure 2 Access-deprivation quintiles based on estimated E2SFCA scores using a 5-minute travel time threshold

Overall, we are unable to comprehensively state whether variations in levels of access-deprivation between the existing access domain and our revised E2SFCA version are due to differences in data or methods. Principally, this is because we were unable to apply E2SFCA methods to the original service supply dataset, which would have been preferable, or as an alternative, apply the original WIMD-2014 methodology to the replicate service-supply dataset. This study therefore provides a practical example of the types of issues raised by Brunsdon and Singleton (2015), where a lack of published details concerning aspects of the meta-data and the computational methods adopted, worsened by a reliance on closed-source GIS software, meant that the approach taken to constructing the access domain of WIMD-2014 can best be described as ‘black-box’ and therefore is, in essence, largely irreproducible. In contrast, to encourage the application of the methods used in this study, we have made the ArcGIS plug-in used to calculate E2SFCA scores freely accessible (Langford et al., 2015). Work is on-going to replicate this tool for use with the open-source GIS software package, QGIS.

6. Acknowledgements

The research presented here is supported by the Wales Institute of Social and Economic Research, Data and Methods (WISERD) – a collaborative venture between the Universities of Aberystwyth, Bangor, Cardiff, South Wales, and Swansea – funded by the Economic and Social Research Council (ESRC: grant number ES/L009099/1).

7. Biography

Nicholas Page is a Senior Research Assistant at the Wales Institute of Social and Economic Research, Data and Methods (WISERD) based in the Faculty of Computing, Engineering and Science, University of South Wales. His recent research interests include the application of GIS in social and environmental studies.

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

Gary Higgs is a Professor of GIS 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). His recent research interests include the application of GIS in areas of health geography and emergency planning.

8. References

Bauer, J., Müller, R., Dörthe, B., Groneberg, D., 2017. Spatial accessibility of primary care in England: a cross-sectional study using a floating catchment area method. *Health Services Research*. DOI:10.1111/1475-6773.12731.

Brunsdon, C., and Singleton, A. (2015). Reproducible research: concepts, techniques, and issues, in Brunsdon, C., and Singleton, A. (eds.) *Geocomputation: A Practical Primer*. London: SAGE Publications, pp. 254-263.

Environmental Systems Research Institute [ESRI]. (2015). ArcGIS Desktop: Release 10.4. Redlands, CA: Environmental Systems Research Institute.

Higgs, G., Zahnow, R., Corcoran, J., Langford, M., Fry, R., 2017. Modelling spatial access to General Practitioner surgeries: does public transport availability matter? *Journal of Transport & Health*, 6, 143-154.

Langford, M., Higgs, G., Fry, R. (2015). “USW-FCA2: An ArcGIS add-In tool to compute Enhanced Two-Step Floating Catchment Area accessibility scores” Software Package – ArcMap add-in. https://www.researchgate.net/publication/287198887_USW-FCA2_An_ArcGIS_add-In_tool_to_compute_Enhanced_Two-Step_Floating_Catchment_Area_accessibility_scores. DOI:10.13140/RG.2.1.3178.8884.

Langford, M., Higgs, G., Fry, R., 2016. Multi-modal two-step floating catchment area analysis of primary health care accessibility. *Health & Place*, 38, 70-81.

Luo, W., and 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*, 30, 865-884.

Luo, W., and Qi, Y. (2009). An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & Place*, 15, 1100-1107.

Ordnance Survey. (2015). OS MasterMap Integrated Transport Network Layer. http://www.ordnancesurvey.co.uk/business-and-government/products/itn-layer.html. Accessed 21 May 2015.

Welsh Government. (2014). Welsh Index of Multiple Deprivation 2014 (WIMD 2014) Technical Report. http://gov.wales/docs/statistics/2014/141218-wimd-2014-technical-en.pdf. Accessed 2 May 2017