Geographies of gazetteers in Great Britain

Stefano De Sabbata[1]* and Elise Acheson[2†]

*Department of Geography, University of Leicester
†Department of Geography, University of Zurich

March 9, 2016

Summary

Gazetteers are playing a central role in the current data revolution as key tools to link content to geographical space. However, their geographies and idiosyncrasies are poorly understood despite their potential to worsen application outcomes and information inequalities. In this study, we analyze two open gazetteers, GeoNames and the Getty Thesaurus of Geographic Names, in terms of the quantity and spatial distribution of features in Great Britain, illustrating how they provide a different and inconsistent picture of the region and are still far less detailed than the institutional gazetteers curated by Ordnance Survey.

KEYWORDS: Gazetteers, GeoNames, Getty Thesaurus of Geographic Names, Geographic Information Retrieval, Data Science.

1. Introduction

The rapid expansion of information technology to most aspects of everyday life is leading to a widespread adoption of 'big data' analysis. Gazetteers play a fundamental role in GIScience and neighbouring disciplines that apply computational methods such as natural language processing to texts in order to relate information to space (e.g. Jones et al., 2001; Yoshioka and Kando, 2013).

Even when the challenges of ambiguity (e.g., Leidner and Lieberman, 2011) and vagueness (e.g. Jones et al., 2008) are laid aside, the practical success of most applications is strongly related to the quality of underlying gazetteers. Put simply, any application using gazetteers to link content to geography can only reflect the content of the gazetteer itself. However, recent analyses showed how global gazetteer coverage can vary widely in space (Graham and De Sabbata, 2015).

While further study is necessary to understand the nature, origin, and practical implications of these global inequalities, it is also illuminating to analyse gazetteers locally, since global trends might mask local issues. The Ordnance Survey (OS) 1:50,000 Scale[3] gazetteer (OS50k) and its replacement, the recently released OpenNames[4] gazetteer, provide excellent points of reference for external quality testing. For instance, Smart et al. (2010) provide a comparison of the number of toponyms provided by different gazetteers for Great Britain, including OS50K and GeoNames[5].This paper thus presents a detailed study of the geographies of two widely-used and open gazetteers in their coverage of Great Britain: GeoNames and the Getty Thesaurus of Geographic Names[6] (TGN).

[1] s.desabbata@le.ac.uk
[2] elise.acheson@geo.uzh.ch
[3] ordnancesurvey.co.uk/business-and-government/products/50k-gazetteer.html, last acc. on Dec. 15th, 2015.
[4] ordnancesurvey.co.uk/business-and-government/products/os-open-names.html, last acc. on Dec. 15th, 2015.
[5] geonames.org, last accessed on December 15th, 2015.
[6] getty.edu/research/tools/vocabularies/tgn, last accessed on December 15th, 2015.

2. Materials and methods

In order to compare GeoNames and TGN with the two OS gazetteers, we limited our analysis to the extent common to all four datasets: that of Great Britain[7]. We compare the full GeoNames and TGN gazetteers with OS50k, as all three include natural features, whereas OpenNames currently provides natural features only via the API service. As OS50k has recently been deprecated, we also compare its replacement, OpenNames, to GeoNames and TGN, but do so based on populated places only. Using the entire downloaded OpenNames dataset would be unsound since transportation network features and postcodes account for 34% and 64% respectively of its almost 2.6 million features.

Table 1 Number of features selected for the analyses

Gazetteer	*Features (Great Britain)*	*Populated places (Great Britain)*
GeoNames	54,701	16,475
TGN	24,003	16,816
OS50k	248,626	
OpenNames		41,490

Our analysis aims to show how GeoNames and TGN compare to OS50k and OpenNames, as well as to each other, in terms of quantity and spatial distribution of features. The analysis is based on feature counts aggregated by Royal Mail postcode area (N = 120) for each gazetteer, due to the use of these areas in statistical data production and analysis in the UK. We compare these feature counts between gazetteers through a series of linear regression models. If the gazetteers have similar spatial distributions, the models should show high adjusted R^2 values (close to 1), and if they have a similar quantity of features, the models should show regression line slope coefficients close to 1.

For a visual overview of the spatial distribution of features in each dataset, we map the density of features per square kilometer using the ArcGIS Point Density tool. As a compromise between the coarser postcode areas and the 1km precision limit of the OS50k gazetteer (National Grid square dimensions), we chose 5km cells with 10k rectangular neighborhoods for the density maps.

3. Results

The maps in Figure 1 depict the density of all features in GeoNames and TGN, using the quantiles calculated from the GeoNames dataset, and thus illustrate how GeoNames contains more data than TGN across Great Britain – about twice as much, according to the postcode area-based linear model presented in Table 2. In GeoNames, coverage increases sharply around urban centres, whereas in TGN some of the most densely covered areas are in the Orkney and Shetland islands, perhaps due to TGN's focus on places of historical interest. Populated places in both are broadly similar in quantity and correlate highly across postcode areas (*Adj*. $R^2 = 0.90$).

The results of the regression analyses comparing GeoNames and TGN to OS gazetteers are reported in Table 3 and Figures 2 and 3. All linear models satisfy the assumptions of normal distribution and heteroskedasticity of the residuals, and independence of errors. Looking at all features, GeoNames and TGN seem to consistently account for about 10% (0.11 and 0.08 respectively) of the amount of content in OS50k throughout Great Britain. Looking at populated places only, the differences are less marked, as GeoNames and TGN seem to consistently account for about 40% (0.41 and 0.43 respectively) of the amount of populated places in OpenNames. Figure 4 presents point density maps for populated places, showing that despite fairly high correlation values across postcode areas, coverage varies importantly across the datasets, such as TGN's poor coverage of Greater London.

[7] Gazetteer data for Northern Ireland are provided by the Ordnance Survey of Northern Ireland, which only released their gazetteers as open data on November 26th, 2015 (see nidirect.gov.uk/news-nov15-free-online-access-to-public-sector-data, last accessed on December 15th, 2015)

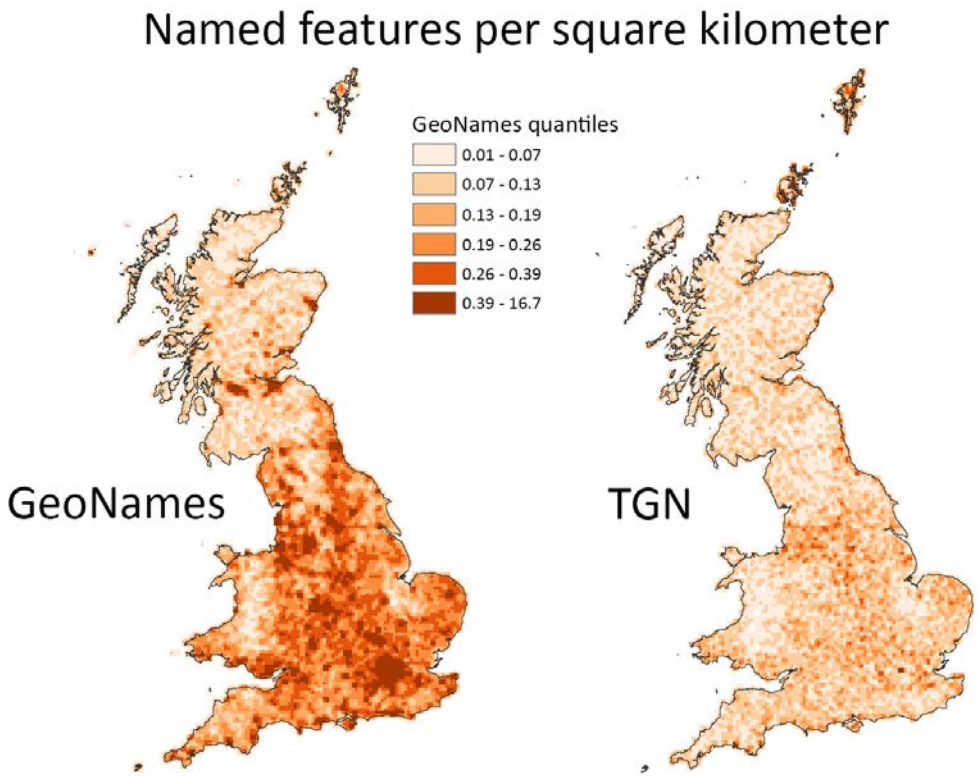

Figure 1 Number of named features per square kilometer in GeoNames and TGN

Table 2 Linear models of postcode area feature counts for GeoNames and TGN

Model	***Adj. R^2***	***Coefficient***	***Std. Error***	***P (sign.)***
1 TGN (all)	0.77			
Constant		-42.24	14.742	<.01
GeoNames (all)		**0.53**	0.026	<.001
2 TGN (Pop. places)	0.90			
Constant		-3.04	5.433	>.05
GeoNames (pop. places)		**1.02**	0.031	<.001

Table 3 Linear models of postcode area feature counts for GeoNames and TGN vs OS50k (all features) and vs OpenNames (populated places).

Model	***Adj. R^2***	***Coefficient***	***Std. Error***	***P (sign.)***
1 GeoNames (all)	0.68			
Constant		218.31	22.220	<.001
OS50k (all)		**0.11**	0.007	<.001
2 TGN (all)	0.86			
Constant		38.54	8.974	<.001
OS50k (all)		**0.08**	0.003	<.001
3 GeoNames (pop. places)	0.80			
Constant		-3.14	7.739	>.05
OpenNames (pop. places)		**0.41**	0.018	<.001
4 TGN (Pop. places)	0.77			
Constant		-10.57	9.034	>.05
OpenNames (pop. places)		**0.43**	0.021	<.001

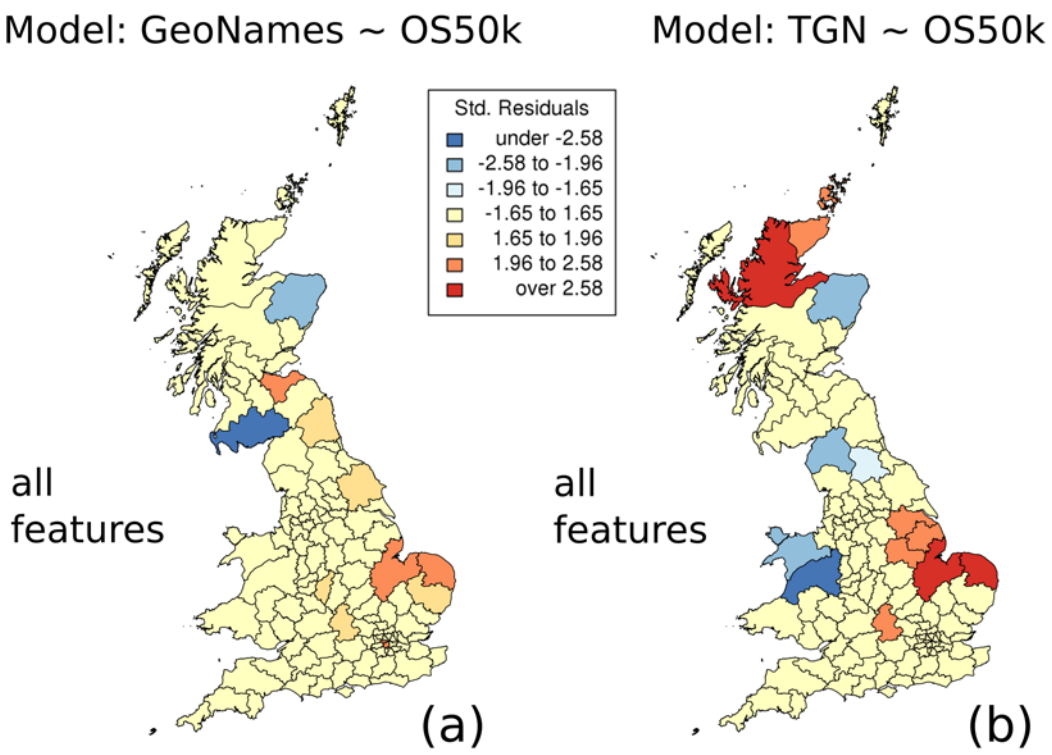

Figure 2 Standard residuals of the linear models 1 (a) and 2 (b) from Table 3

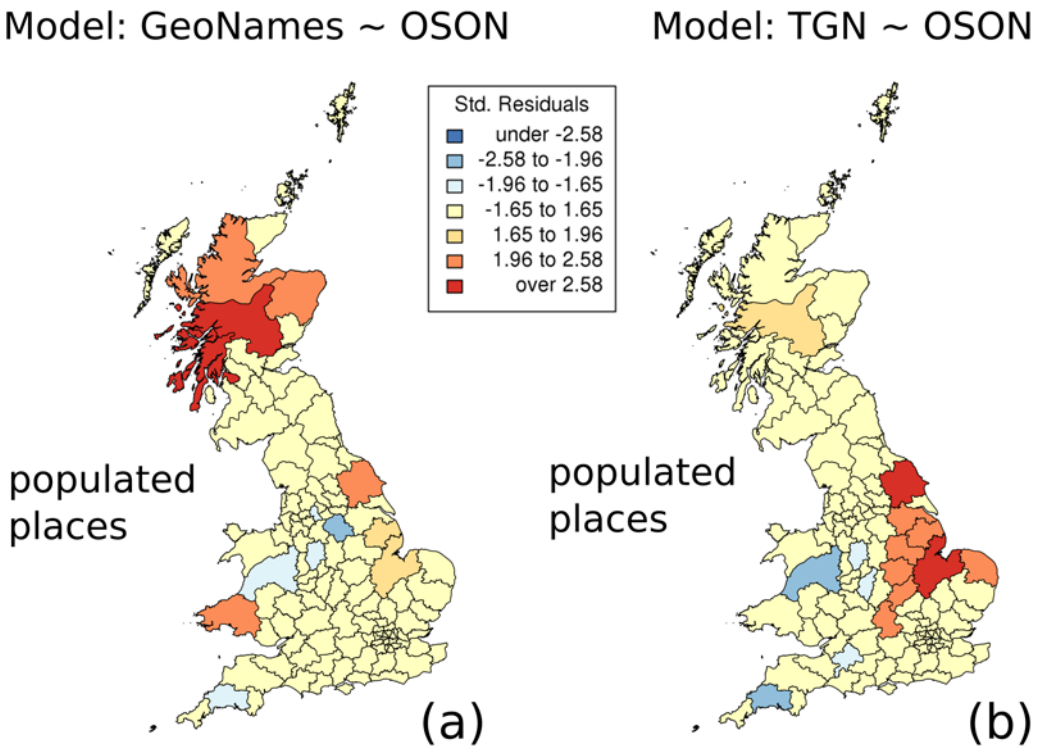

Figure 3 Standard residuals of linear models 3 (a) and 4 (b) from Table 3

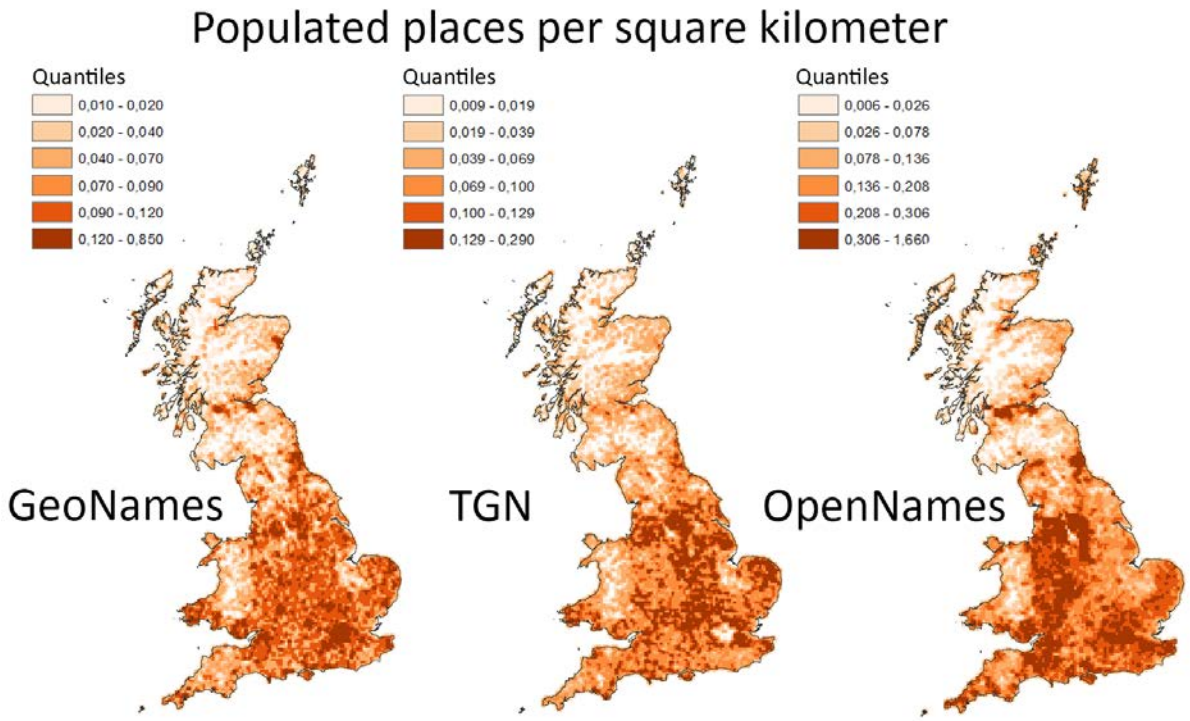

Figure 4 Number of populated places per square kilometer in GeoNames, TGN and OpenNames.

4. Concluding discussion

The results show how the quantity and spatial distribution of features in GeoNames and TGN in Great Britain varies compared to OS gazetteers. GeoNames and TGN provide only a fraction of the overall content of the OS gazetteers, though differences are less marked when looking only at populated places. TGN provides about half the amount of content in GeoNames, but the quantity of populated places is similar and correlates well across postcode areas, despite the coverage being spottier in TGN. This study thus demonstrates how inconsistent the coverage is, depending on the gazetteers, feature types, and regions taken into account.

To situate these results in the broader context, we need to consider that Great Britain is among the areas with a high concentration of features recorded in GeoNames, as well as in TGN. Therefore, these findings are significant not only for Great Britain, but also for the usage of global gazetteers in the analysis of areas covered in less detail.

These findings highlight the risk of perpetuating data-program-data cycles (Bowker, 2013), where the geographies of the output of an application are significantly influenced by the geographies of the gazetteer, and thus represent more the latter than the studied phenomenon. This is both a technical problem, as end-users would ‘see’ the gazetteer rather than a geographic phenomenon, and an ethical issue, as it reinforces broader information inequalities (Graham et al., 2015; Glasze and Perkins, 2015).

5. Biography

Stefano De Sabbata is a lecturer in Quantitative Geography at the University of Leicester, and a research associate of the Oxford Internet Institute of the University of Oxford. His research focuses on geographic relevance and location-based services, as well as critical GIS, and quantitative human geography, particularly information geographies.

Elise Acheson is a PhD student in the Geocomputation Unit at the University of Zurich, working on

automatically producing geographical models of text documents. Before starting her PhD in March 2015, she worked for 3 years as part of a software development team within ESRI.

References

Bowker, G.C., 2013. Data flakes: An afterword to “Raw Data”is an oxymoron. *Raw Data Is an Oxymoron*. MIT Press, pp. 167 – 172.

Glasze, G. and Perkins, C., 2015. Social and political dimensions of the OpenStreetMap project: Towards a critical geographical research agenda. In *OpenStreetMap in GIScience* (pp. 143-166). Springer International Publishing.

Graham, M. and De Sabbata, S., 2015. Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A*, *47*(6), pp.1254-1264.

Graham, M., De Sabbata, S. and Zook, M.A., 2015. Towards a study of information geographies:(im) mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, *2*(1), pp.88-105.

Jones, C.B., Alani, H. and Tudhope, D., 2001. Geographical information retrieval with ontologies of place. In *Spatial information theory* (pp. 322-335). Springer Berlin Heidelberg.

Jones, C.B., Purves, R.S., Clough, P.D. and Joho, H., 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), pp.1045-1065.

Leidner, J.L. and Lieberman, M.D., 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), pp.5-11.

Smart, P.D., Jones, C.B. and Twaroch, F.A., 2010. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Geographic Information Science* (pp. 234-248). Springer Berlin Heidelberg.

Yoshioka, M. and Kando, N., 2012. Issues for linking geographical open data of geonames and wikipedia. In Semantic Technology (pp. 375-381). Springer Berlin Heidelberg.