Journal of Maps

ISSN: (Print) 1744-5647 (Online) Journal homepage: http://www.tandfonline.com/loi/tjom20

Interactive mapping for large, open demographic data sets using familiar geographical features

Oliver O'Brien & James Cheshire

To cite this article: Oliver O'Brien & James Cheshire (2016) Interactive mapping for large, open demographic data sets using familiar geographical features, Journal of Maps, 12:4, 676-683, DOI: 10.1080/17445647.2015.1060183

To link to this article: http://dx.doi.org/10.1080/17445647.2015.1060183

JOURNAL OF MAPS, 2016
VOL. 12, NO. 4, 676–683
http://dx.doi.org/10.1080/17445647.2015.1060183

SOCIAL SCIENCE

OPEN ACCESS

Interactive mapping for large, open demographic data sets using familiar geographical features

Oliver O'Brien and James Cheshire

Department of Geography, University College London, Gower Street, London WC1E 6BT, UK

ABSTRACT

Ever-increasing numbers of large demographic data sets are becoming available. Many of these data sets are provided as open data, but are in basic repositories where it is incumbent on the user to generate their own visualisations and analysis in order to garner insights. In a bid to facilitate the use and exploration of such data sets, we have created a web mapping platform called DataShine. We link data from the 2011 Census for England and Wales with open geographical data to demonstrate the power and utility of creating a conventional map and combining it with a simple but flexible interface and a highly detailed demographic data set.

ARTICLE HISTORY
Received 17 March 2015
Accepted 28 April 2015

KEYWORDS
population; choropleth; open data; interactive; census; DataShine

1. Introduction

Until relatively recently, the creation of maps from demographic data sets was undertaken by geographic information systems (GIS) specialists who had access to complex software packages and the requisite skills to operate them. As online mapping interfaces have become more powerful, thanks to the likes of the Google Maps API and OpenLayers, it is now possible to create platforms that enable those without previous GIS training to produce detailed maps from a huge number of data sets. Whilst the creation and interpretation of any map requires some degree of spatial literacy, more people than ever can produce them guided by a series of well-established cartographic and statistical principles. We seek to apply these in the DataShine web mapping platform, as shown in the Main Map (datashine.org.uk) to facilitate the use and dissemination of spatially referenced open data in the UK. We apply a technique to ensure that such maps produced retain a geographic familiarity to a non-specialist audience used to conventional maps of local areas, towns or cities, by constraining the demographic data mapped to building footprints or urban extents. Interactive elements present in the website augment the mapped information and invite data exploration and knowledge discovery.

2. Methods

2.1. *Use of building footprints and urban extents*

Choropleth maps are a popular and straightforward way to show demographic data on a map. Each spatial statistical unit is coloured according to the proportion of a population within it that has a particular attribute, for example, the proportion of the working-age population that is in full-time employment. The colour ramps used are typically scaled from the lowest to the highest proportions across all the areas, or from 0% to 100% of the general population, and may be evenly banded (stepped/graduated, i.e. discrete) or use a continuous ramp. Alternatively, other methods of banding may be applied, such as Jenks natural breaks (1967) or having equal populations within each band.

The main problem with choropleth maps is that the statistical unit areas are shown conventionally, that is, with their full geographical extent, regardless of their population density. This means that rural and other low-population areas tend to dominate the map – particularly if the statistical unit areas are designed to have approximately equal populations, such as Output Areas (OAs) in the UK. This is shown clearly in Figure 1.

Similarly, attributes that do not have similar values across rural areas will cause the map to appear as a 'patchwork' of changing values, again visually emphasising change across rural areas that have a low population ahead of changes across urban extents with a high population, even though the latter may be of more significance.

A further limitation with 'conventional' choropleth maps is the way in which abrupt colour changes can emphasise the boundaries of statistical unit areas. Such boundaries are often unfamiliar to the non-specialist, and often do not correspond to any easily relatable feature on the ground. This can make orientation tricky without the addition of contextual labels and roads, which can often clutter the map.

CONTACT Oliver O'Brien o.obrien@ucl.ac.uk

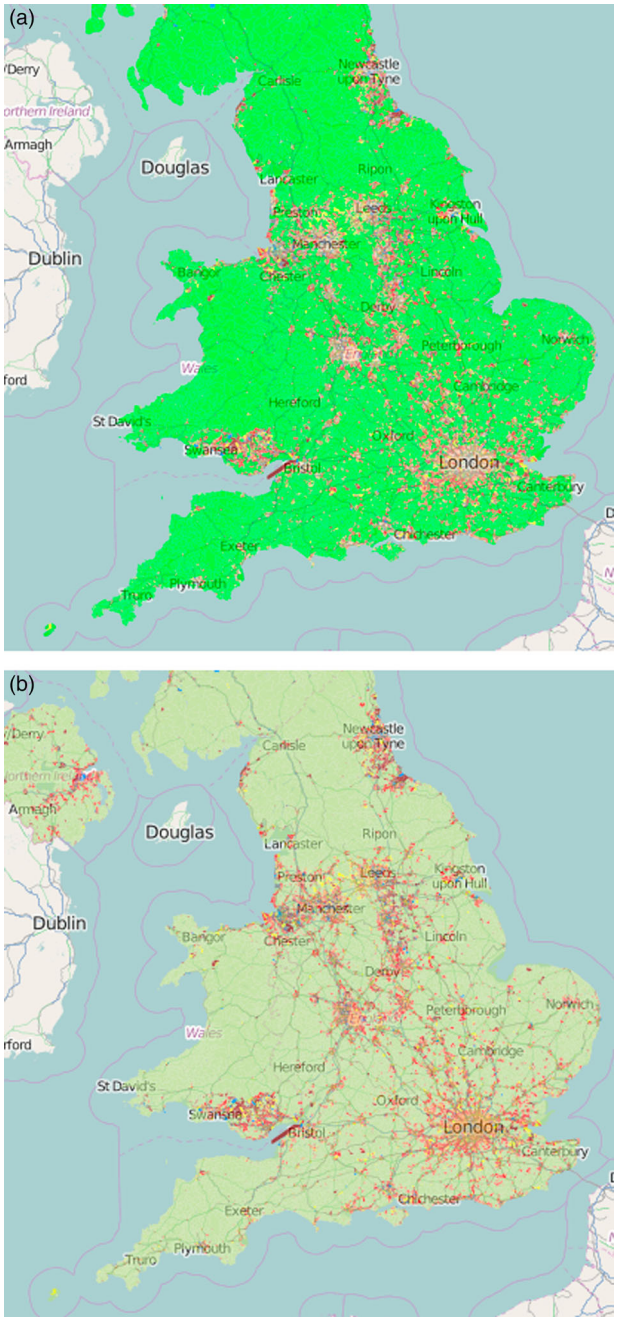

Figure 1. Output area classification (OAC) 2001 choropleth map examples for England/Wales, showing green colours predominating. Both maps here are showing the same data. The rural classification is shown in green. In the first map (a) it dominates the map even though the classification does not have more prominence than the other classifications. The second map (b) uses a less bright green that de-emphasises the rural areas. The background map and data are copyright of the OpenStreetMap contributors, licenced under a CC-By-SA licence (map) and the Open Database Licence (data).

We therefore set out to produce a map with colours representing the demographic proportion constrained to building footprints (at large scale) or urban extent (at small scale). This effect is achieved by creating a non-building/urban mask and layering a conventional choropleth map below so that the data 'shines through' the holes in the mask. The layers are not combined, so that different choropleth maps, which can be generated very quickly by software renderers, can be used with the same, more complex, mask.

The effect removes the overemphasis of unpopulated areas within the map, while retaining the actual geography, so making the map more relatable and familiar and facilitating knowledge discovery (see Figure 2). The non-building areas (e.g. roads, rivers and parks) can be used to augment the map at larger scales with regular features to further increase its familiarity and allow area recognition. In Great Britain, such data are freely and easily available through the Ordnance Survey Open Data suite of products. The data are rich in detail and allow the straightforward creation of a conventional geographic map as a layer that can then be combined with the choropleth map layer as described above.

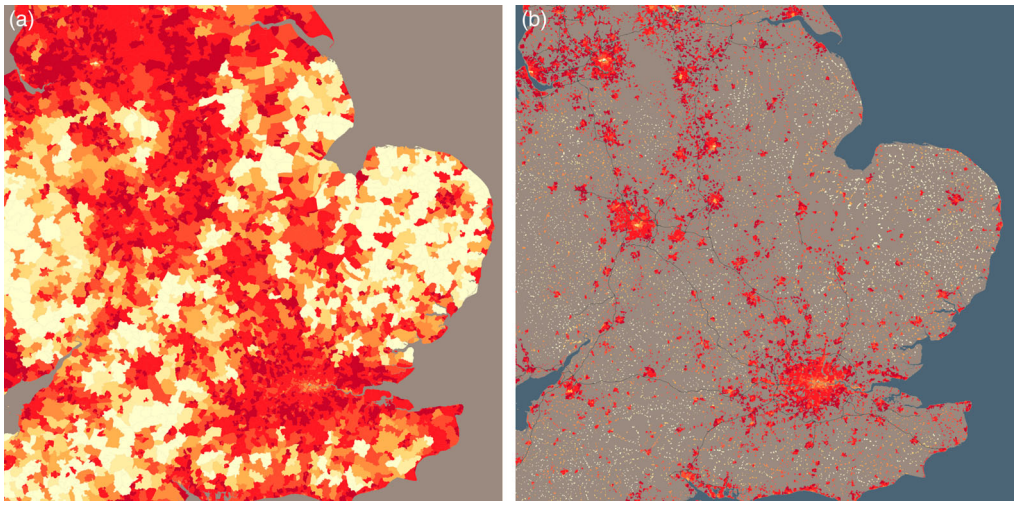

Figure 2. Gas central heating across England, (a) without and (b) with an urban extent layer overlaid. In the latter case, the urban/rural difference in results is more easily identifiable, as well as an inner city/outer city difference.

Cartographically, we think this offers a marked improvement on simple choropleth maps alone that ignore the built environment, but acknowledge that this generates false precision in the map since some users, on first glance, could observe the high level of detail and reasonably assume the data are being disseminated at the building level. This is not the case since the most fine-grained level of statistical data release in the UK is at the OA level (e.g. for the Census). An additional issue is that the open building footprint data set used also does not distinguish between residential houses and commercial/industrial buildings, and it does not indicate the number of storeys in each building. Disclosure controls would prevent the mapping of data at the building level, but it would be possible to provide a more nuanced impression of the built form by removing non-residential buildings. Such data exist but they are not 'open' and therefore, at present, cannot be easily integrated into a public platform. It is believed that these issues, while significant, are not sufficient to detract from the utility and visual improvement to the basic map. Additionally, there is scope for a partial improvement by combining with more attribute-rich open data sources such as from OpenStreetMap.

At smaller scales we use urban extents rather than building outlines, to increase the speed of map image production and because the individual detail would not have been discernable at such scales. For medium scales where building outlines are used, alpha blending of neighbouring pixels is applied so that buildings 'flare out' slightly, ensuring every building outline, which is otherwise surrounded by empty land, always contributes to at least one pixel in the resulting map – otherwise antialiasing would result in the smaller outlines vanishing altogether.

2.2. *Colour selection*

We also sought to facilitate appropriate colour selection. We make use of the ColorBrewer series of colour ramps (Brewer, 2015), defaulting to the 'YlOrRd' (yellow through orange to red) ramp as the change in lightness acted as a good quantitative indicator of measure variation, while the gradual shift of hue acts to additionally highlight the highest values in particular. We use eight bins of equal intervals, for simplicity in the key and display. Where it is considered (using an algorithm detailed below) that the standard deviation about the mean is likely the most interesting map to show, a diverging colour ramp 'RdYlGn' (red through yellow to green) is instead used. The user is always able to change the default ramp to another one, either by using buttons within the user interface or manually specifying a ColorBrewer colour identifier code as the appropriate parameter value in the URL.

For reasons of speed of generation of the maps, and to minimise a 'patchwork quilt' effect caused by having a large number of small areas with varying values, viewed across a wide area, two aggregation geographies are used – OAs (average population around 300) at larger scales and wards (average population around 6500) at smaller scales – we detail these in Table 1.

A side effect of varying the geographies, the modifiable areal unit problem (OpenShaw, 1983), means that localised results can disappear as the zoom level is varied. Additionally, because the average and standard deviation of the measure are calculated based on the areal grouping used, these values will change slightly when the alternative geography is used. This change in the overall average value for a measure may be counter-intuitive to a non-expert user. The historical nature of wards, meaning that their population can vary greatly due to boundaries being largely fixed for long periods of time, contrasts with OAs, statistical areas

Table 1. Statistical geographies used for our mapping for the 2011 census tables.

Statistical unit	Zoom levels	Approx. scales (72dpi, U.K.)	Number in England/Wales	Average population size	Notes
OA	12–14	1:70000 to 1:17000	181,408	309	Modern, may be redrawn to maintain minimum/maximum populations. OAs just have reference codes
Ward	6–11	1:4M to 1:140000	8570	6543	Historic, so some very small population wards. Wards have names was well as codes

that are combined/split periodically when their population suffers significant change. This can exacerbate the amount of change in the overall areal average value for many measures.

The limitations of a web browser as a dissemination platform are recognised and so an alternative solution has been engineered, whereby the current view can be redrawn onto a PDF, which is then presented to the user for printing or storage. This functionality is incorporated into the main user interface by means of a 'Print to PDF' button.

2.3. *2011 census for England and Wales*

We use the 'Quick Statistics' aggregate tables of Census 2011 data, released for England and Wales by the UK's Office of National Statistics (ONS), as our sets of measured data. There are 1558 measures across 67 tables, including a total population measure for each table that is used as the denominator when calculating percentages for binning and display – so leaving 1491 different maps that can be created for each geography.

We obtain the measures for two statistical geographies and visualise them using two physical geographies, shown in Tables 1 and 2, respectively.

We assemble DataShine (datashine.org.uk) with a simple 'drop-down' interface (see Figure 3) to access the tables and measures. All measures are presented in the same way, with on-the-fly choropleth maps created based on the percentage of the population (which is itself sometimes a subset of the residential population, depending on the table selected), and the built environment 'mask' (building block or urban extent, depending on the scale) overlain. The different statistical and physical geographies are switched automatically as the user adjusts the scale of the map.

All choropleth maps are presented with the values being placed in one of eight bins, shown using sequential or diverging ColorBrewer colour ramps. Different strategies are employed to split out each measure into the eight bins. These strategies are determined automatically, using an algorithm based solely on the percentage average of the measure across the currently used statistical unit areas, and the standard deviation of this average. The strategy did not use the maximum or minimum values, so it is possible for some measures that no areas fall into one or more of the outlier bins. The strategies used are shown in Table 3.

The Quick Statistics tables are designed to be as granular as possible, and do not aggregate categories together. So, for example, population age is shown in single year increments. The effect is that most measures therefore have a low average. Natural geographical variations across England/Wales, and, in particular, London's unusual demographic composition with respect to the rest of England/Wales (Gale, 2012), often result in measures having a relatively high standard deviation with respect to the average. This means that most measures are shown on DataShine with a colour ramp that shows multiples of the England/Wales mean, or Location Quotient – the first of the four strategies in Table 3.

We have chosen the YlOrRd and RdYlGn colour ramps, as discussed above, as they work well to show outliers and general trends, with yellow acting as a good contrasting colour to the other, strongly hued, colours used. One disadvantage, however, of these colours is their lack of neutrality; in other words, an association with meaning outside the measure being viewed. For example, red is often associated with 'bad' aspects, and green with 'good'. This may or may not be the case for the measure being viewed, but the automatic use of such colours may lead a casual viewer to draw incorrect conclusions. Other colour ramps can be selected by the user in DataShine, for example, to mitigate colour-blindness and colour-association issues, if desired.

The resulting map generation, based on the algorithm, generally displays an appropriately varied choropleth map, showing sufficient variation without overwhelming the display to the viewer. This means that the addition of the building/urban extent layer, which helps aid referencing the map and relating to real world locations, can be applied without the additional detail obscuring the measure being

Table 2. Physical geographies used for building/urban extent display in our mapping for the 2011 census tables. It is not appropriate to calculate an average population for each urban extent or building block in consultation with Table 1, because the two tables are for different extents and the physical units include non-residential buildings/areas.

Physical unit	Zoom levels	Approx. scales (72dpi, U.K.)	Approximate number in England/Wales	Source
Building outline	10–14	1:270000 to 1:70000	2,000,000	Ordnance survey vector map district
Urban extent	6–9	1:4M to 1:540000	20,000	Ordnance survey meridian 2

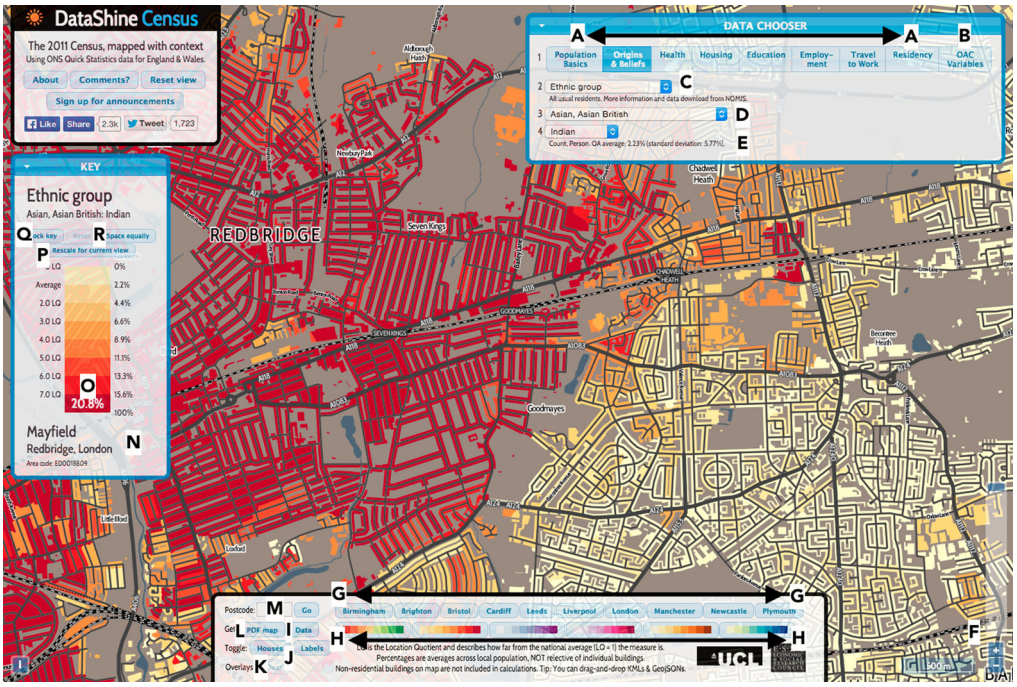

Figure 3. The DataShine website. A: Table groupings from the 2011 Census 'Quick Statistics'. B: Data used to construct the 2011 Area Classification of Output Areas 'OAC'. C: Table heading. D: Column group heading (only present for some tables). The overall population is specified below, along with a link to the source data webpage. E: Column heading. The population type and country-wide mean and standard deviation are specified below. F: Zoom control and scalebar. G: Quick jump buttons for key cities. H: Alternative colour ramps. I: Data download. J: Show/hide place-name labels or building/urban extents. K: Clear or fade any added KML/GeoJSON overlay polygons. L: Download a PDF map of the current view, for printing. M: Jump to postcode. N: Geographical attributes for the area under the cursor. O: The percentage value of the currently selected measure with respect to the population, for the area under the cursor. This is positioned on top of the bin it falls in to. P: Recalculate the bins and recolour the map by using the mean and standard deviation for just the area in view. Q: Lock the current bins to allow comparison across other measures. R: Use equal intervals for the bins, from 0% to 100%.

displayed or significantly diluting the impact of the data formed by the underlying choropleth map.

The map is primarily designed to be viewed in a website browser; further examination of the map can be performed using a PDF printing option that has been developed for the site, or printing directly from a web browser. It is likely that production of a paper map is an additional improvement towards the goal of creating maps that are as familiar as possible to casual users of maps, while including demographic data.

2.4. Uptake, feedback and examples

DataShine has been extremely well used. It was launched in July 2014, first in an edition showing the 2011 Census Quick Statistics tables as described above, and subsequently showing further Census-related data sets such as travel-to-work flows and geodemographics. In its first 6 months, it attracted nearly 112,000 users and 165,000 page views. The website continues to attract around 4000 users per month. We also provide a mailing list to keep users informed (as of

Table 3. Strategies used for colour binning the measure for display as a choropleth.

Mean[a]	Standard deviation of the mean	Strategy utilised	Default ColorBrewer colour code	Default colour ramp type	No of measures (OA)	No of measures (Ward)
Less than 12.5%	Greater than 2/3 of the mean	Multiples of the mean (also known as Location Quotient)	8-class YlOrRd	Sequential	1106	834
Greater than 87.5%	Greater than 2/3 of the difference of mean from 100%	Multiples of difference of the mean from 100%	8-class YlOrRd	Sequential	10	15
Between 12.5% and 87.5%	Greater than 2/3 of the mean or difference of the mean from 100%	Equal interval from 0% to 100% in 12.5% segments	8-class YlOrRd	Sequential	56	21
Any value	Less than 2/3 of the mean and difference of the mean from 100%	Divergence by standard deviations from the mean	8-class RdYlGn	Diverging	318	620

[a]Mean of the measure's percentage population across all statistical unit areas for the currently viewed geography.

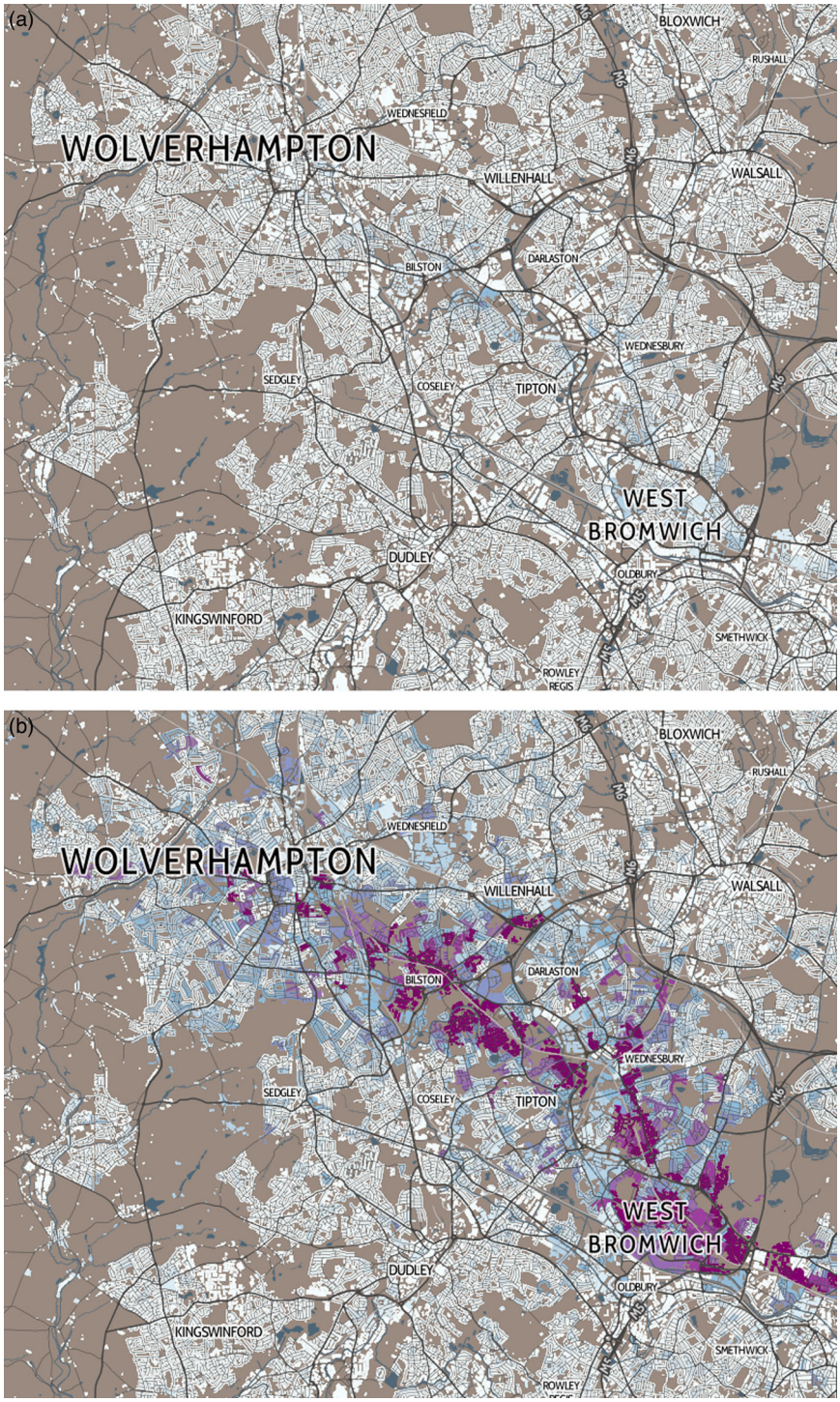

Figure 4. A map showing the proportions of people who travel to work by 'Underground, Metro, Light Rail or Tram' according to the 2011 Census (a) without and (b) with local area colour rescaling applied. In this case, the rescaling improves the map as it shows the impact of the Midland Metro system more clearly.

March 2015 it has 230 subscribers), 22 of whom have filled in a detailed questionnaire about their use of the website. We also operate an active comments page on the platform's blog (blog.datashine.org.uk). Analysis of this user base suggests the following key groups:

(1) Private sector – such as journalists, consultants at major engineering/architecture firms and consumer insight organisations and researchers at think tanks.
(2) Public sector – national and local government, planners, health-care providers and commissioning groups.
(3) Non-profit sector – charities looking to target their resources or raise awareness of particular issues, for example, transport mode-shift campaigners.
(4) Educators – widespread usage in both undergraduate and high school teaching.
(5) Members of the public – responding to features and social media mentions from the above as well as those with a general interest in their local area.

The amount of usage across these groups varies from single visits lasting less than a couple of minutes (the bulk of users) through to repeated intensive use,

such as making multiple maps for download, on a weekly basis. It is for these latter 'power users' that we added a series of additional functions at their request.

2.5. *Local area rescaling*

A function is included in DataShine to allow recalculation of the average percentage and the corresponding standard deviation using only data from the area shown in the extent of the web browser. This can result in the binning strategy changing, for local areas that are significantly divergent from the England/Wales averages. This is particularly useful if a user may be in a region where a particular demographic has very low (or high) values compared to the national average, but because the colour breaks are based on the national average, local variation may not be shown clearly.

For example, the popularity of London's underground network with its large population means that, for other cities with metros or trams, their usage is harder to pick out from the census. So, in Birmingham, the Midland Metro can be hard to spot (see Figure 4). Upon rescaling, just the local results are used when calculating the average and standard deviation, allowing usage variations, in this case along the route of the railway, to be more clearly seen. For transport planners in Birmingham this results in a much more useful map.

2.6. *Data download and overlay*

In addition to the ability to colour the map based on the statistical distribution of the data in a local area, DataShine offers users the option to download the data behind the map in a comma separated value format for further analysis. We consider this one of the most important tools in the DataShine suite since it gives users the chance to source only the data they need without the arduous process of navigating the lists of table names provided by the ONS website and then either using GIS or database functionality to exclude the areas that are not of interest.

The most recent functionality integrated into the platform is the ability to drag and drop boundary files, in KML and GeoJSON formats, onto DataShine maps. This is especially useful for those seeking to add their own annotations, and has been successfully used in the teaching of mixed methods research to first-year undergraduates at UCL Geography (Cheshire, 2015). This functionality is a quick fix for those who are interested in demographics within non-census boundaries such as clinical commissioning group areas in the National Health Service. It is not as good as aggregating the data to such areas, but it enables users to add their own geographical boundaries for reference and orientation purposes.

Continued feedback from users has enabled us to prioritise additional functionality which is due to be delivered in due course, such as the addition of an additional small-scale geography (likely at local authority level) and inclusion of other large, open demographic data sets in the socio-economic space, such as the Index of Multiple Deprivation, upon their next publication and availability on open data platforms.

3. Conclusions

There are various techniques that can plot demographic data such as the census aggregate statistics that we use here. The techniques have different limitations and shortcomings, and the need to balance clarity and effectiveness of the demographic data display with production of a map that is accessible, understandable and memorable is difficult. Our method, whereby a basic choropleth map is 'shone through' a regular map layer, colouring building blocks or urban extents only, provides a 'best of both worlds' combination of simplicity, clarity and relatability. While it is not without its issues, we believe that such treatment makes a positive contribution to disseminating large, spatially arranged demographic data sets such as the UK Census, particularly as the combination of an open toolstack, open geographic data and open demographic data lowers the barriers required to produce such a map.

Disclosure statement

The authors do not have any financial interest or benefit from the direct applications of this research.

Funding

This work was supported by the Economic and Social Research Council (ESRC) [grant number ES/K009176/1].

Software

The map is produced using Python bindings of the Mapnik C++ library, and is displayed in web browsers using the OpenLayers 3 and JQuery/JQueryUI Javascript libraries. Technical literature on the aspects of the technologies used behind DataShine can be found at http://oobrien.com/?s=datashine where a number of blog postings are available.

ORCID

Oliver O'Brien http://orcid.org/0000-0002-3413-0853
James Cheshire http://orcid.org/0000-0003-4552-5989

References

Brewer, C. (2015). *ColorBrewer 2.0*. Retrieved from http://www.colorbrewer.org/

Cheshire, J. (2015). *DataShine and GeoJSON*. Retrieved from http://vimeo.com/119639260

Gale, C. (2012). *Geodemographic output area classification for London, 2001–2011*. Edinburgh: GISRUK.

Jenks, G. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, *7*, 186–190.

Openshaw, S. (1983). *The modifiable areal unit problem*. Norwick: Geo Books.