Creating a synthetic spatial microdataset for zone design experiments using 2011 Census and linked administrative data

James Robards[*1], Chris Gale.[†2] and David Martin[‡1,2]

[1]ESRC National Centre for Research Methods (NCRM), University of Southampton, SO17 1BJ, UK
[2]ESRC Administrative Data Research Centre for England (ADRC-E), University of Southampton, SO17 1BJ, UK

Summary

New forms of administrative and linked data with high attribute and spatial detail present increased risk of information disclosure about individuals, potentially enabling identification. Evaluation of disclosure risk using real data is not feasible, as disclosive record-level data are understandably not accessible for such research. This paper details development of a synthetic microdataset for England and Wales with a realistic distribution of household locations and individual characteristics. We here exploit the synthetic dataset for assessment of alternative automated zone design solutions, with the eventual aim of improving researcher access to maximally useful data while minimising disclosure risk.

KEYWORDS: Synthetic data, privacy, census, automated zone design, statistical disclosure control

1. Introduction

Spatial aggregation has long been a standard approach to the statistical disclosure control (SDC) in population data such as those collected from a census of population. This results in small area aggregate datasets, which in the last two censuses in England and Wales have been aggregated to output areas (OAs) created using automated zone design (Cockings et al., 2011). The only way that researchers can access samples of census microdata is to accept low sample fractions and high levels of geographical aggregation (Tranmer et al., 2005), and are often still restricted to secure data laboratories. Increased interest in the exploitation of administrative and linked data (ONS, 2015) offers great potential for research and policy, but faces the same SDC challenges to an even greater degree. While researchers will often desire to undertake detailed spatial analysis, data owners will usually need to understand the likely risks of disclosure of individual records before deciding what levels of spatial aggregation must be applied. Standard population thresholds are used for census outputs but the decision is more complex for unique administrative and linked datasets where the potential risks and benefits are specific to particular combinations of variables. Our research is concerned to use automated zone design to aid this evaluation, but must generate realistic assessments without any risk to 'real' data. We here present the generation of a spatially detailed synthetic population dataset, illustrated for the County of Hampshire, as a basis for SDC experiments.

[*] James.Robards@soton.ac.uk
[†] C.Gale@soton.ac.uk
[‡] D.J.Martin@soton.ac.uk

2. Design of a synthetic spatial microdataset

The use of synthetic data offers the safest means of undertaking methodological experiments without any risk to actual data. We have designed an approach which begins with an available synthetic population of individual persons, the England and Wales Synthetic Population dataset (https://data.cdrc.ac.uk/dataset/synthetic-population) produced by the ESRC Consumer Data Research Centre (CDRC), by spatial microsimulation of records from the ONS 1% sample of census records released as a teaching file (ONS, 2014). The CDRC work already embodies substantial processing to provide a set of individual person records containing nine key variables (age, gender, economic activity, marital status, occupation, number of hours worked and general health), duplicated and constrained to match published census totals at the Middle Layer Super Output Area (MSOA) level (typical population 7,500).

Our own research has established a process to enhance this safeguarded dataset (or any equivalent source) to produce a synthetic spatial microdataset with household structure and plausible spatial locations, which can be used for a range of SDC zone design assessments. The entire process has been implemented in R (https://www.r-project.org/).

In order to assign a realistic spatial distribution for person-level synthetic microdata it is first necessary to group individuals into synthetic 'households' with plausible sizes and compositions, and then to assign these households to locations which closely reflect the true residential population distribution. We have achieved the first of these tasks by assembling a rule set which allows us to identify valid person types (from the microdata) to household types (from census aggregate data). Thus a two-person pensioner household must be assigned two persons of pensionable age; a single adult non-pensioner household must be assigned a single person of working age; a couple with two children must be assigned two children and two adults of working age, etc. We have worked with seven person types and 10 household types, the last of which accounts for residents of communal establishments. Based on this information, we are able to group the MSOA-level synthetic individuals into the correct number of households, matching the known distribution of household types at the OA level (typical population 325)

For the assignment of plausible spatial locations, we have commenced by using postcode locations from the census-based ONS Postcode Directory, which includes counts of households at each full postcode. These locations have been randomly perturbed within a 150m offset to produce unique locations. However, in order to further constrain the distribution, these locations have been constrained to fall within the populated 10m grid cells of the OpenPopGrid dataset (Murdock et al., 2015). This identifies populated residential areas at the 2011 census, based on weighted intersection of Ordnance Survey Vectormap Districts buildings and headcounts for Thiessen polygons generated around ONS postcode centroids. The result is a set of candidate household locations which match observed residential geography at 10m resolution. The correct numbers of synthetic households of each type are then allocated to address locations within each OA.

3. Preliminary mapping of household locations / data characteristics

Figure 1 shows the spatial distribution of synthetic households (n=730,699) for a subset of the national data covering Hampshire including the cities of Southampton and Portsmouth. This study area covers a wide range of urban and rural area types. The household distribution accurately reflects the real settlement pattern with sparse rural and linear settlements correctly reproduced. Our synthetic locations contain the correct number of communal establishments within each OA, although locations are randomly chosen from among the candidate points. Because of linkage of the CDRC data to include the household characteristics, we have a full distribution of all the original variables and household types, providing a rich synthetic distribution on which to evaluate the level of protection required in subsequent zone design solutions.

Figure 2 presents a small area within Southampton showing the way in which household locations relate to features such as rivers (e.g. the River Itchen is in the centre), roads, parks and other major features, reflecting the use of OpenPopGrid as a dasymetric mask for the offsetting of household locations from postcode centroids.

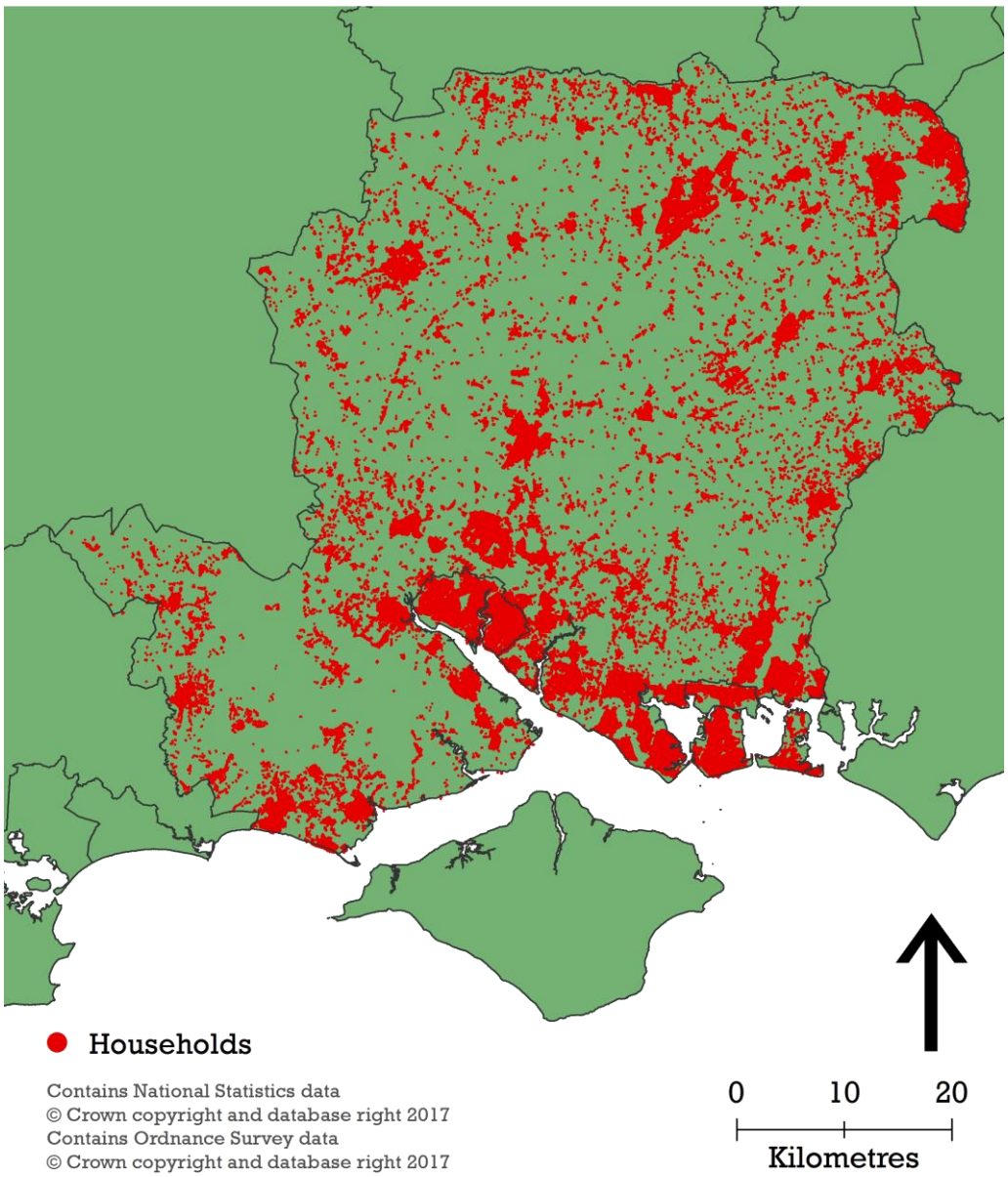

Figure 1 Spatial distribution of synthetic households in Hampshire

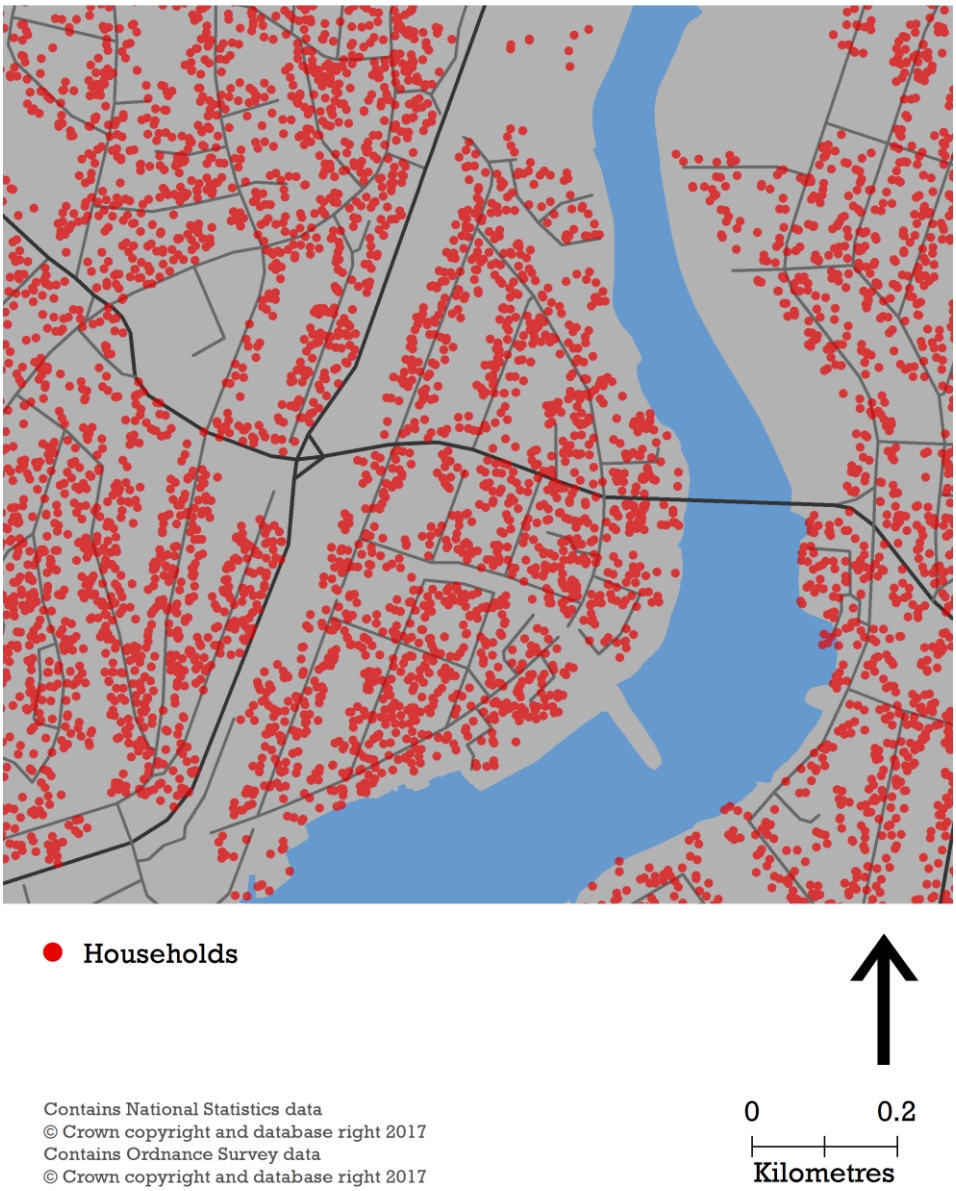

Figure 2 Spatial distribution of synthetic households around the St Denys area of Southampton

4. Automated zone design

Automated zone design methods are already established for construction of small areas for release of ONS census statistics. Design criteria for a particular zoning system may be specified in terms of parameters such as target population counts or overall number of zones and are set to meet SDC requirements such as threshold numbers of persons and households as well as factors such as alignment with existing areal units populations and real-world topographic features. Our synthetic dataset forms a highly realistic basis for assessing the data disclosure risks associated with different aggregations. Using AZTool (https://www.geodata.soton.ac.uk/software/AZTool/) the same software employed in the design of census OAs, it is possible to reassemble the synthetic data according to an enormous range of different design criteria, which can be compared with aggregations to standard areal units such as OAs and MSOAs.

Figure 3 shows a sample aggregation of households in Southampton, where the specifications were a minimum threshold population 750 and a target population of 1,500. The aggregation was built from

Thiessen postcode polygons (n=5,001, mean population 45). The optimal solution from the AZTool output was one with 152 polygons and mean population of 1,496 which met the target population and minimum threshold set. For data owners, there is a trade-off between the level of spatial detail and population characteristics that should be released to researchers. Our ability to evaluate many alternative zone design solutions using the synthetic dataset permits safe assessment of alternatives in order to find solutions which meet both research needs and SDC requirements.

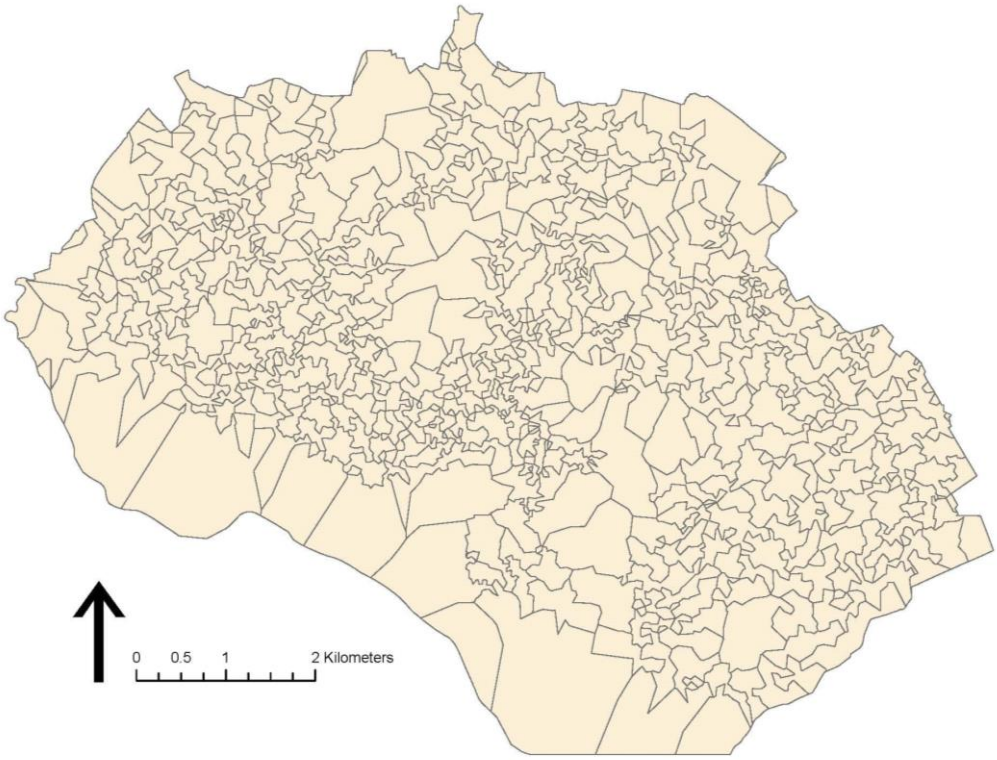

Figure 3 Sample zoning solution for Southampton with target population of 1,500 and a minimum threshold of 750

5. Acknowledgements

ESRC award: ES/L008351/1; 2011 Census synthetic individual level MSOA dataset (Source – ESRC Consumer Data Research Centre (CDRC), University of Leeds) The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 025, ES/L011840/1; ES/L011891/1; 2011 Census aggregate data from the Office for National Statistics licensed under the Open Government Licence v.3.0.; ONS Postcode Directory and digital boundaries contains National Statistics data © Crown copyright and database right 2017, contains OS data © Crown copyright and database right 2017; OpenPopGrid, 2015 product contains information from several Information Providers under Open Government Licence and Ordnance Survey OpenData Licence (http://openpopgrid.geodata.soton.ac.uk /Openpopgrid.xml)

6. Biographies

James Robards is a Research Fellow at NCRM at the University of Southampton. He has an extensive track record of using census micro-data in demographic, health and housing research, including testing of 2011 census microdatasets.

Chris Gale is a Research Fellow at the ADRC-E at Southampton. His PhD research undertaken at University College London led to creation of the Office for National Statistics' 2011 Output Area Classification.

David Martin is Professor of Geography at Southampton, involved in leadership of ESRC's UK Data Service, ADRC-E and NCRM. He has worked closely with the Office for National Statistics on automated zone design, including 2001 and 2011 census output geographies.

References

Cockings S, Harfoot A, Martin D and Hornby D (2011). Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales *Environment and Planning A* 43, 2399-2418.

Murdock AP, Harfoot AJP, Martin D, Cockings S and Hill C (2015). *OpenPopGrid: an open gridded population dataset for England and Wales.* GeoData, University of Southampton. http://openpopgrid.geodata.soton.ac.uk/.

ONS (2014). *2011 Census Microdata Teaching File User Guide*, ONS https://www.ons.gov.uk/file?uri=/census/2011census/2011censusdata/censusmicrodata/microdatateachingfile/microdatauserguide/2011censusmicrodatateachingfileuserguide_tcm77-349416.pdf.

ONS (2015). *ONS Census Transformation Programme Administrative Data Update*, ONS http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/census/2021-census/progress-and-development/research-projects/beyond-2011-research-and-design/research-outputs/administrative-data-update.pdf.

Tranmer M, Pickles A, Fieldhouse E, Elliot M, Dale A, Brown M, Martin D, Steel D and Gardiner C (2005). The case for small area microdata *Journal of the Royal Statistical Society Series A (Statistics in Society)* 168, 29-50.