National Centre for Research Methods Working Paper

5/17

Creating a synthetic spatial microdataset for zone design experiments

James Robards, Chris Gale and David Martin

NCRM National Centre for Research Methods

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Creating a synthetic spatial microdataset for zone design experiments

NCRM/ADRC-E working paper

James Robards[a], Chris Gale[b] and David Martin[ab*]
[a]National Centre for Research Methods, University of Southampton, SO17 1BJ, UK
[b]Administrative Data Research Centre-England, University of Southampton, SO17 1BJ, UK
*Corresponding author (telephone 023 8059 3808, email: D.J.Martin@soton.ac.uk).

Abstract

New forms of administrative and linked data containing high levels of attribute and spatial detail present increased risks 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. Data for the study comes from the England and Wales 2011 Census and are combined from multiple tables and sources to arrive at the final dataset. Our motivation for this work is exploit the synthetic dataset for assessment of alternative automated zone design solutions, with the eventual aim of improving researcher access to the most useful data while minimising disclosure risk. However, the synthetic microdataset, and the methodological approach used to produce it, potentially have wider utility than our automated zone design research. This working paper documents the generation of our synthetic dataset in a way intended to benefit others needing to conduct experiments on a non-disclosive population microdataset.

1. Introduction

New forms of administrative and linked data with high levels of both attribute and spatial detail have great analytical potential but present increased risks of information disclosure about individuals, potentially enabling identification. 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 (Martin et al., 2001; Cockings et al., 2011). The only way that researchers can access samples of census microdata is to accept small sample fractions and high levels of geographical aggregation (Tranmer et al., 2005), generally restricted to project-specific access within secure data laboratories. Increased availability of administrative, and particularly linked administrative, data (Office for National Statistics (ONS), 2015a) 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. Even where research access is granted to the most detailed data, the release of analysis results will still often require decisions about an appropriate level of spatial aggregation. Standard population thresholds are used for census outputs but the issue is more complex for unique record-level datasets where the potential risks and benefits are specific to particular combinations of variables. Our research is concerned with using automated zone design to aid this evaluation, but must generate realistic assessments without any risk to 'real' data. In this working paper we focus on one specific element of this research, namely the generation of a spatially detailed synthetic population dataset, as a basis for geospatial and SDC experiments. The specific requirements of this dataset will be further elaborated below.

The remainder of this paper is organised as follows. Section 2 outlines the target specification of a synthetic microdataset. Section 3 describes the source data that are to be combined. Section 4 outlines the design of the data processing and Section 5 details its implementation, Section 6 presents the results of this processing, focusing on a specific study area, and Section 7 summarises the broader contribution of our approach, both for our own research agenda, and related research areas.

2. Specification

The use of synthetic data offers the safest means of undertaking methodological experiments without any disclosure risk to real data. This working paper is concerned with establishing a process 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. Rather than create our own micro dataset, we have elected to enhance a safeguarded public use microdataset, although the approach presented here could be applied to any data source with broadly equivalent characteristics. We here set out the key requirements of a dataset that meets our needs:

- ***A realistic array of residential household locations (i.e. placing households only in locations which are residential).***

  We are seeking a dataset which has a distribution of households which closely follows that of actual residential locations. This requirement may be articulated in two different ways. Firstly, the synthetic data should aggregate to match plausible marginal counts across a range of geographical units used in official statistics, and secondly, the locations should reflect the actual spatial distribution and density of dwellings at the local level so that plausible results may be reproduced from any spatial analysis.

- *A high degree of attribute detail, across a range of variables (i.e. as per microdata information typical of a census record).*

  We are seeking a dataset which has a range of variables for households and individuals within the household reflecting those that would typically be included in a research data extract and are likely to pose an SDC challenge to a data provider.

- *A complete population (i.e. records reflecting the density and size of the entire population, not a sample).*

  The dataset should contain individual and household records that reflect the entire population over an extended area, containing the necessary ranges of population density and local and regional variations in attribute characteristics.

- *A realistic distribution of household types / structures per area.*

  The range of household types within the dataset should reflect the true make-up of the localities which they represent and include individual level data which corresponds with the household type. Within settlement variations in the types of household (e.g. single persons, family etc.) should therefore be reflected in the final dataset produced.

- *Plausible household structures.*

  Individuals within the synthetic microdataset should be grouped into synthetic 'households' with plausible sizes and compositions, which may then be assigned to locations with the required spatial distributional characteristics noted above, in keeping with the household structures and sizes of different localities. This will require the construction of a ruleset determining those individual and household types which can be combined with validity. In addition to the distribution of household types found across localities, Communal Establishments (i.e. types of Communal Establishments include: hospitals, care homes, prisons, defence bases, boarding schools and student halls of residence) (ONS, 2015b) are included in the plausible household structures and their processing detailed in this paper.

3. Description of available data

There is no standard dataset which meets all the requirements set out in Section 2. In order to arrive at a synthetic spatial micro dataset with all these characteristics it is necessary to combine several datasets each of which is georeferenced at some level and have some shared attributes, yet where each offers a new set of characteristics to the combined dataset, primarily by adding detail in one or more dimensions to what is available from the other sources. Georeferencing of the datasets varies between standard census geographies (e.g. middle layer super output area (MSOA), output area (OA)). Table 1 presents the six key datasets used in our processing. The table summarises the source

and contribution made by each dataset. There are many specific details in our implementation that are inevitably unique to the structure of census and postal geographies in England and Wales, and the way that households and communal establishments have been defined and recorded in the 2011 Census. Nevertheless, the broad approach followed here could be implemented using alternative data, and would thus be applicable, either in the UK using different data sources, or in other countries using entirely different datasets, but having the same principal characteristics.

Table 1: Data sources employed in generation of synthetic spatial microdataset

Dataset to be included in composition of final data	Purpose of the dataset	Data used in present study	Variables in data used for present study
1. Input microdata file (name in processing: 'SYNTH_POP')	Individual level data for population members by census area.	England and Wales synthetic population from ESRC Consumer Data Research Centre at MSOA level.	Zoneid (MSOA), personid, generalhealth, agecategory, maritalstatus, hoursworked, occupation, economicactivity, sex.
2. Small area aggregate counts (name in processing: 'OA_HHLD_TYPE')	Household types (living arrangements) by census area to enable the matching of individual level data (1) with households (4) to generate a realistic array of household types reflecting the residential make up of neighbourhoods.	ONS household types at OA level.	OA, household types, household type counts, communal establishment counts.
3. Communal establishment counts (name in processing: 'OA_COMMUNALS')	Provides counts of the communal establishment populations along with a total count of communal establishments.	2011 Census, Table ID QS420EW at OA level.	Communal Establishment Type: All categories Communal Establishment Type: Medical and care establishment Communal Establishment Type: Other establishment Communal Establishment Type: Establishment not stated.
4. Postcode household counts (name in processing: 'PC_HHLD_COUNTS')	Counts of households per postcode to be distributed around the locality defined in the residential settlement mask.	ONS Postcode Directory.	Postcode, Total (people), Males, Females, Occupied_Households, OA.
5. Residential settlement mask	Datasets used to define the residential locations where addresses are to be distributed to.	OS VectorMap District, OS Open Map – Local, OpenPopGrid.	-
6. OA to MSOA look-up file (name in processing: 'OA_MSOA')	Look-up between census geographies.	ONS 2011 Census Geography lookup tables (combining	OA, MSOA.

		ONS, NRS and NISRA 2011 Census Geography lookup tables).	

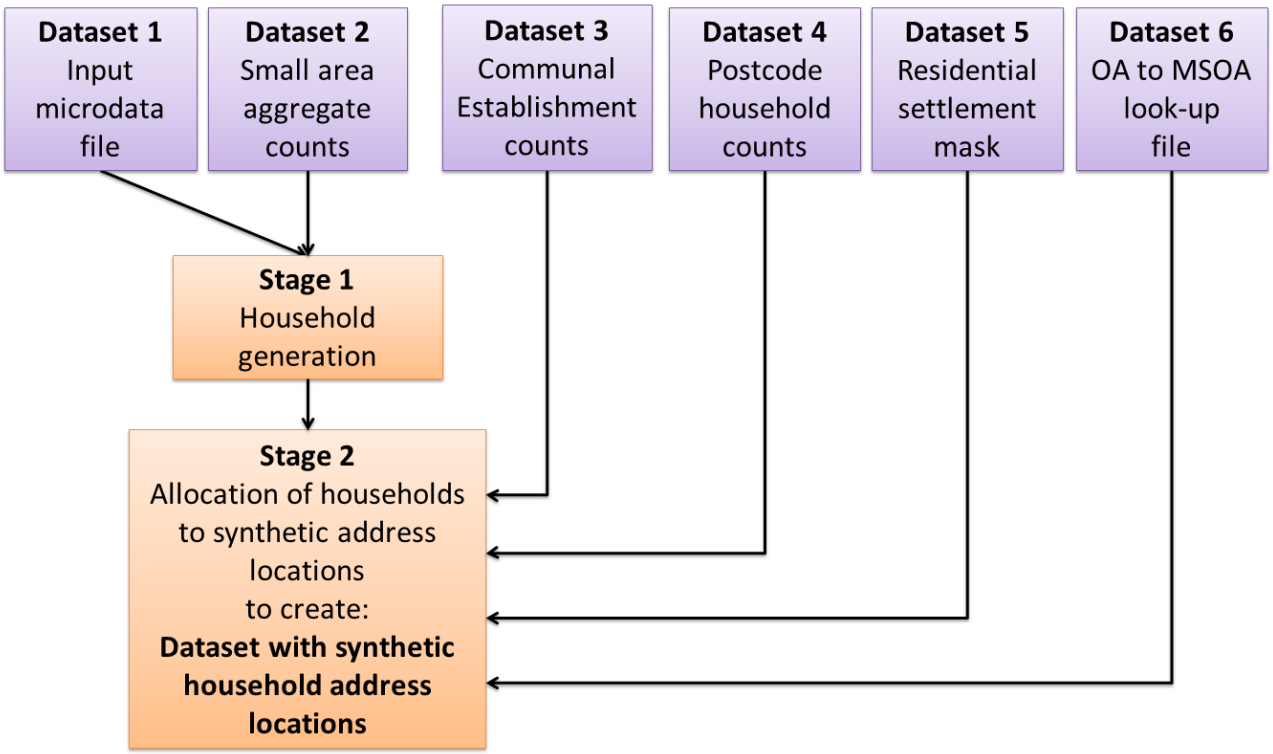
We have designed an approach which begins with an available synthetic population of individual persons, the England and Wales Synthetic Population dataset (https://data.cdrc.ac.uk/dataset/synthetic-population) produced by the Economic and Social Research Council (ESRC) Consumer Data Research Centre (CDRC), by spatial microsimulation of records from the ONS 1% sample of 2011 Census records released as a teaching file (ONS, 2014a). 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 MSOA level (typical population 7,500) (ONS, 2012). Figure 1 indicates the input datasets. The 'Input microdataset' (dataset 1), is the individual level synthetic data for population members, referenced to MSOA. This dataset provides the individual level characteristics required in the target dataset. These are the individuals who are to be allocated into households. Existing census microdatasets containing individuals already structured within households are themselves considered to be potentially disclosive and are only accessible by approved researchers within a secure data laboratory setting (ONS, 2014b), hence they are unsuitable for the purpose intended here. 'Small area aggregate counts' (dataset 2) provides counts of different types of households for a small area census geography, including a count of the number of communal establishments (e.g. types of communal establishments include: hospitals, care homes, prisons, defence bases, boarding schools and student halls of residence). 'Communal Establishment counts' (dataset 3) provides a count of communal establishments and their resident populations from 2011 Census data in census table QS420EW. Communal establishments present particular challenges, as they are spatially concentrated and may have unique characteristics which make them distinctive from the general household structure of a locality. This means that they may present particular disclosure risks in a real world, record-level dataset. There are complexities with the processing of communal establishments because of inconsistencies within census data; these will be explained within the detailed methodology within this document. It is our intention to include synthetic communal establishments within the target dataset. 'Postcode household counts' (dataset 4) provides counts of households and populations at the finest spatial granularity available from the 2011 Census data in England and Wales, the unit postcode level. Across England and Wales there are on average 18 households per residential unit postcode, with households in the same unit postcode sharing spatial coordinates. These locations play a key role in the subsequent generation of plausible spatial

locations for synthetic households. The process of distributing households makes use of (dataset 5) 'Residential settlement mask' which provides a dasymetric mask (Fisher and Langford, 1996) in the form of a high resolution spatial grid indicating whether 10m grid cells have a residential population. Finally, a look-up of OA to MSOA geographies (datatset 6) is used within the spatial allocation.

4. Design / outline of process

Figure 1 provides an overview of our processing of the datasets introduced in Section 3. Two distinct processing stages are apparent in the Figure. These are (1) the generation of synthetic households and (2) the allocation of households to synthetic address locations. This section outlines the design of our data processing, while Section 5 provides technical details of the implementation. The entire processing sequence has been implemented in the R program and language (https://www.r-project.org/).

Figure 1: Diagram of production of the synthetic spatial microdataset



The first block of processing code, 'Household generation' uses the information on the numbers of households of each type per locality in the input microdataset and small area aggregate counts to allocate individuals from the microdataset into households, in keeping with the household type profile (see Table 3) of the area. We assembled a rule set which allows us to identify valid person types (from dataset 1) for each household type (from the census aggregate data (dataset 2)). These person types relate to the type of living arrangements in a particular household. In addition,

information about the range of household sizes (i.e. the number of individuals living in a household unit) is necessary to inform the numbers of individuals to be allocated into each household, for example ensuring that only one individual is assigned into single person households or two into couples, etc. 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 created seven person types and 10 household types, the last of which accounts for residents of communal establishments. Based on this information, MSOA level synthetic individuals can be grouped into the correct number of households, matching the known distribution of household types at the OA level (typical population 309) (ONS, 2012). Depending on the type of household to be constructed, different numbers of individuals from the list of person types will be selected. The first stage in the data processing is a form of matching code for the input microdataset (dataset 1) and small area aggregate counts (dataset 2). The code links the individual level data with the household area data to place individuals within households in the locations where they reside according to the small area aggregate counts. The purpose of this is for the household structure of locations to be reflected in the generation of households and for the inclusion of the microdata for synthetic individuals to be accurate with reference to these households.

Table 2 details the seven person types identified from the descriptions of individuals in the input microdataset (dataset 1) to enable joining to the small area aggregate counts (dataset 2). The last two columns show the age and marital status variables used in each type. These types have been selected so as to align individuals with household types provided in the small area aggregate data. This is necessary as the ONS classification of household types for the small area aggregate cross tabulations is coarser than that which can be derived from the input microdataset. Note that one problem in relation to the person type classification is the age classification and number of people aged 16 and 17 years and the inability to distinguish between individuals living with family and in full time education and others working full time and not living with family (hence the note after Table 2).

Table 2: List of person types to which individuals in the input microdataset are coded

Person Type	Description	Age variable filter	Marital Status variable filter
1	Adult – living either alone or not	16 to 64	Single OR Divorced or formerly in a same-sex civil partnership which is now legally dissolved
2	Adult – living alone	16 to 64	Separated (but still legally married or still legally in a same-sex civil partnership) OR

			Widowed or surviving partner from same-sex civil partnership
3	Adult 65+ either alone or not	65 and over	Single OR Divorced or formerly in a same-sex civil partnership which is now legally dissolved
4	Adult 65+ alone	65 and over	Separated (but still legally married or still legally in a same-sex civil partnership) OR Widowed or surviving partner from same-sex civil partnership
5	Adult – together	16 to 64	Married or in a same sex civil partnership
6	Adult 65+ together	65 and over	Married or in a same sex civil partnership
7	Child	0-15	<16 years old therefore ineligible to marry

Note that where adults in the input microdata are in age categories 16-24 and economic activity is either "Economically active – student" or "Economically inactive – student" these cases are considered to be candidates to be recoded to become 'children' in the household allocation process as the age group spans both children and adults. If this were not done, there would be insufficient individuals coded as children to match the household composition data in the small area aggregate dataset.

Table 3 shows the list of household types to which the small area aggregate counts logically collapse to from pre-defined categories used in ONS (2014c). Within the variables there are a number of cases which are classified as 'complex other' and also communal establishments.

Table 3: List of household types derived from small area aggregate file

Household Type	Description	Variable coding in ONS file used to produce household type
1	One adult - To age 64 (inclusive)	F1814
2	One adult – Aged 65 and over	F1812
3	Two adults - To age 64 (inclusive)	F1819, F1825
4	Two adults – Aged 65 and over	F1818, F1840
5	One adult and 1 child	F1868
6	One adult and 2+ children	F1870
7	Two adults and 1 child	F1848, F1858, F1864
8	Two adults and 2+ children	F1850, F1860, F1866
9	Complex other	F1821, F1828, F1833, F1872, F1874, F1838, F1842
10	Communal establishments	F1298

The second block of processing code, 'Allocation of households to synthetic address locations' matches Communal Establishment counts (dataset 3), Postcode household counts (dataset 4) and uses a Residential settlement mask (dataset 5). This stage is focuses on the location of households and the need for these to have a realistic neighbourhood level geospatial distribution. For the assignment of plausible spatial locations we have used the unit postcode centroid locations provided

as part of the ONS Postcode Directory, and appended these to the unit postcode counts of households from the 2011 Census. Random locations are generated around the postcode centroids for each household (households are assigned to MSOA and OA and then within the MSOA and unit postcode within that OA coordinates are granted), constrained to fall within residential areas defined by the intersection of relevant spatial mask datasets, the primary input being populated 10m grid cells from 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 therefore allocated to address locations within each OA.

5. Implementation

This section explains the implementation of the design described in the previous section using R, and the outcome at each step. Subsections in this section correspond to the two key processing steps: household generation and household address location generation. Table 1 has already detailed data sources employed in constructing the synthetic spatial microdataset dataset, the source of the data and the way in which it is used in this study.

1. *Household generation – Matching the input microdataset and Small area aggregate counts to generate households*

In order to assign each household a unique ID number the total number of unique household types in each MSOA in England and Wales (7,201) are used to create a new dataset called 'DataCategories'. For example, as there are 1,946 households of Type in MSOA E02000001 there are 1,946 rows in 'DataCategories' belonging to MSOA E02000001 and Household Type 1. As such, each row in 'DataCategories' represents a unique household in the dataset and can therefore be given a sequential ID reference number (H1, H2... etc.)

Table 4: Example lines for dataset 1 (Input microdataset)

zoneid	personid	sex	agecategory	generalhealth	maritalstatus	hoursworked	occupation	economicactivity
E02004760	7806039	1	6	4	3	5	5	7
E02004760	7806039	2	4	2	5	7	5	6
E02004760	7806039	2	1	1	3	6	6	8
E02004760	7806039	1	1	4	2	2	7	4
E02004760	7806039	1	7	5	3	7	2	9
E02004760	7806039	1	6	3	1	7	4	6
E02004760	7806039	1	3	1	3	4	8	10

Note this is an example of the dataset and does not contain values from the input data itself. The values are randomly generated in Excel for the range of categories in the actual dataset.

Table 5: Example lines for dataset 2 (Small area aggregate counts) and dataset 3 (Communal establishment counts)…

MSOA	AV_PER_PER_HHLD	HH_SIZE_ALL_CAT_UNIT	HH_SIZE_1_P_HHLD	HH_SIZE_2_P_HHLD	HH_SIZE_3_P_HHLD	HH_SIZE_4_P_HHLD	HH_SIZE_5_P_HHLD	HH_SIZE_6_P_HHLD	HH_SIZE_7_P_HHLD	HH_SIZE_8_PLUS_P_HHLD	F1297__ALL_Communal_ests_NUMBER
E02000001	1.6	4385	2472	1356	339	153	36	17	7	5	42
E02000002	2.5	2713	814	785	492	362	175	66	11	8	3
E02000003	2.6	3834	1039	1077	721	563	281	100	30	23	2
E02000004	2.6	2318	609	702	412	374	160	46	7	8	4
E02000005	2.7	3183	808	854	601	539	248	103	18	12	0
E02000007	2.5	3441	1064	936	597	481	221	105	21	16	4
E02000008	2.5	4591	1398	1235	819	670	308	125	24	12	2

…Table 5 (continued): Example lines for dataset 2 (Small area aggregate counts) and dataset 3 (Communal establishment counts)

hhld_type_1	hhld_type_2	hhld_type_3	hhld_type_4	hhld_type_5	hhld_type_6	hhld_type_7	hhld_type_8	hhld_type_9	hhld_type_10	hhld_type_all_EXC_COM_EST	hhld_type_all_INC_COM_EST
1946	526	864	153	73	18	174	132	499	188	4385	4573
405	409	217	203	229	220	174	317	539	51	2713	2764
573	466	423	238	215	193	321	455	950	12	3834	3846
286	323	317	188	104	104	217	291	488	245	2318	2563
437	371	365	156	214	229	258	433	720	0	3183	3183
635	429	310	142	316	291	221	422	675	119	3441	3560
783	615	456	218	381	293	368	566	911	5	4591	4596

The two datasets used in the England and Wales implementation are:

1. **Input microdataset containing 56,075,912 rows, each relating to one synthetic population member. It contains the 'pers_type_append' column which is used in the code to assign persons to households based on the set of rules created.**
2. **'DataCategories' details synthetic households and contains 23,425,076 rows. Each row contains a unique reference number and the type of household.**

Processing proceeds by looping through both datasets one MSOA at a time. Columns in the small area aggregate data provide a count of the total number of people in each household type (defined in the typology listed above) and how many households are present, of each size from 1 to 8+ persons. These counts are essential in order to ensure that the synthetic households follow the actual household size distribution.

Objects are created to represent the total number of people living in communal establishments (taken from the Small area aggregate counts dataset) and the total number of children. The latter is the sum of all people in the MSOA assigned to either group 7 (child) or 8 (originally an adult, but now recoded as a child to ensure there are sufficient children to match the households) in the input microdataset.

A summary dataset is created to represent the total count of the 10 different household types present in the MSOA. This is derived from 'DataCategories'. A matrix is created to divide the counts of the 10 different household types between the differently sized households. The only values that can be specified at this stage are those where the household types have precisely defined numbers of individuals (e.g. 1 person living by themselves, a couple, etc.). Household types that include 2+ children cannot be established yet as they require further calculation, nor can the population who live in communal establishments.

If the MSOA contains a population that live in communal establishment(s) then that population is divided between the number of establishments present. In most MSOAs there are at least the same number of people as there are establishments, meaning that each establishment will contain at least 1 person. However, in certain MSOAs there are establishments present with no population or more establishments than population. In these instances, the entire population is assigned to a single establishment, resulting in the other establishments having zero population. We are not concerned to replicate the populations of actual communal establishments, but rather to allocate a population of the correct size and plausible composition to communal establishment addresses within each OA. The treatment of communal establishments is important due to the large distorting effects they can have on local population characteristics, for example where a neighbourhood may contain student halls of residence or residential care homes. There is an artefact in reported census data whereby in

some locations, although communal establishments were identified (i.e. in address data), no census returns / population counts were received from those communal establishment addresses. This leads to some addresses not having population, causing the issues noted here.

To complete the matrix for each MSOA the numbers of children to be assigned to different household sizes for Household Types 6, 8, 9 and communal establishments needs to be calculated. The total number of unallocated children at this stage is the sum of those that have not been allocated to Household Types 5 and 7. The minimum number of children in Household Types 6 and 8 (2 children per household) is subtracted from the total unallocated children count. In creating person type 8 only the minimum number required to ensure households in that MSOA have enough children are allocated (there is no way of knowing how many more children there should be). For MSOAs where there are still remaining unallocated children after this process (in 5,969 of the 7,201 MSOAs in England and Wales this was the case), the remaining unallocated children are assigned into three possible categories:

1. **If the communal establishment population in the MSOA is greater than 0 then a random number of children (between 1 and the count of remaining unallocated children) is selected.**
2. **Between 20% and 50% of the new count of remaining unallocated children after step 1 are added to those children already assigned to be in either Household Type 6 or 8. (The total number of children assigned to be in Household Types 6 and 8 are therefore calculated by the minimum required children to ensure there are at least 2 children in each household plus a random value (if possible) so these household types can have more than 2 children).**
3. **Any remaining children left after this process are assigned to Household Type 9 (complex other).**

With a total number of children to be allocated to Household Types 6 and 8, they need to be assigned into different household sizes (as they are merely recorded in the aggregate data as having two or more children). If the number of Household Type 6 and 8 children is the minimum possible, then they will be assigned into 3 person sized households for Household Type 6 and 4 person sized households for Household Type 8. The only exception in these instances is where the allocation of children in this way will result in too many 3 or 4 person sized households (compared to how many are known to be present in each MSOA). In these instances children are assigned to larger sized household belonging to Household Types 6 and 8. In these instances more children are required; however this has been taken into account when creating person type 8, therefore there will always be enough children in each MSOA to assign them to Household Types 6 and 8 without over-assigning to smaller households.

The allocation of Household Types 6 and 8 to different household sizes, when there are more than the minimum number of children to be allocated, is decided at random. As the total number of

Household Types 6 and 8 is already known (derived from the Small area aggregate counts dataset), each household is randomly assigned to be in a 3 to 8 person sized household if they belong to Household Type 6, or be in a 4 to 8 person sized household if they belong to a Household Type 8. The division of Household Types 6 and 8 to different household sizes are then randomly allocated and swapped until two criteria are fulfilled. Firstly, the total number of children assigned to the different sized households is the same as the number of children allocated to Household Types 6 and 8 for each MSOA. Secondly, the random assignment of Household Types 6 and 8 to different household sizes does not exceed the total number of actual households of that size for each MSOA.

At this stage, Household Type 9 is the only type to which no households have been assigned. To complete this task, the total numbers of Household Types 1 to 8 by 1 to 8 person household sizes are cross-tabulated. These values are then compared to the total number of 1 to 8-person household sizes known to exist in the MSOA. The difference between these counts for each household person size corresponds to the number of households of each size required of Type 9. The total number of households belonging to each type and each size has been calculated for the MSOA. The allocation of individual synthetic population members into Household Types 1 to 10 can now be undertaken, commencing with the adults (person types 1 to 6) as detailed in Table 6. They are allocated as detailed in Table 6.

Table 6: Summary of process applied to each household type derived from small area aggregate file

Household Type	Description	Process
1	One adult - to age 65	First assigns Person Type 2 (adult to age 65 alone) and if this is not enough also includes Person Type 1 (adult to age 65 either alone or not).
2	One adult - 65+	First assigns Person Type 4 (65+ alone) and if this is not enough also includes Person Type 3 (65+ either alone or not).
3	Two adults - to age 65	First assigns Person Type 5 (adult to age 65 together) and if this is not enough also includes Person Type 1 (adult to age 65 either alone or not).
4	Two adults - 65+	First assigns Person Type 6 (65+ together) and if this is not enough also includes Person Type 3 (65+ either alone or not).
5	One adult and 1 child	First assigns Person Type 1 (adult to age 65 either alone or not) and Person Type 3 (65+ either alone or not) and if this is not enough also includes Person Type 2 (adult to age 65 alone) and Person Type 4 (65+ alone).
6	One adult and 2+ children	
7	Two adults and 1 child	First assigns Person Type 5 (adult to age 65 together) and Person Type 6 (65+ together) and if this is not enough also includes Person Type 1 (adult to age 65 either alone or not) and Person Type 3 (65+ either alone or not).
8	Two adults and 2+ children	
9	Complex other	Assigns a random selection of remaining Person Types 1 to 6.
10	Communal establishments	Assigns a random selection of remaining Person Types 1 to 6 for the adults. The number of children allocated reflects the number of children assigned to communal establishments in a previous step.

Any adult individuals that have not been assigned to a Household Type by this point will instead be added to 8+ person sized households later in the processing. From the earlier calculations, the household destinations for all the children present in the MSOA have now been decided.

The next stage of the assignment process is the allocation of individual adults to households. This is done by assigning people to households based on their household size (not household type) using a loop function which runs for the total number of people who live in households of each size in the current MSOA. Depending on the household size, varying criteria are used to identify both individuals who can belong a household of that size and Household Types that can accommodate that size of household. Before the commencement of each loop, a subset of suitable individuals and household types are created. Each iteration selects at random a household that still has space for at least one more person (i.e. its assigned person count is below its target size). An unallocated person record is then randomly selected to be assigned to this household. The suitability of the person to be added to the household will depend on both the size of the household and the persons already allocated to it. The rules are as follows:

- **1 person households**
    - **Household Type 1: 2 (adult to age 65 alone) then 1 (adult to age 65 either alone or not) if no more available 2. If no more available 2 or 1 then 5 (adult to age 65 together) are selected (this accommodates any inconsistences in the census data).**
    - **Household Type 2: 4 (65 and over alone) then 3 (65 and over either alone or not) if no more available 4. If no more available 4 or 3 then 6 (65 and over together) are selected (this accommodates any inconsistences in the census data).**
- **2 person households**
    - **Household Type 3: 5 (adult to age 65 together) then 1 (adult to age 65 either alone or not) if no more available 5.**
    - **Household Type 4: 6 (65 and over together) then 3 (65 and over either alone or not) if no more available 4.**
    - **Household Type 5: 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 7 (child), 8 (child – former adult) if random household empty; 1 or 3 if already contains child or 7 or 8 if already contains adult.**
    - **Household Type 8: 1 to 6 i.e. any combination of adults only, but in most cases person types 2 and 4 (those living alone) will have been already assigned to one person sized households.**
- **3 person households**
    - **Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than 2 children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains two children.**
    - **Household Type 7: 7 (child) or 8 (child – former adult) if random household empty; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains a child.**
    - **Household Type 9: 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not).**
- **4 person households**
    - **Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than three children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains three children.**
    - **Household Type 8: 7 (child) or 8 (child – former adult) if random household contains less than two children; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains two children and less than two adults.**
    - **Household Type 9: Any random combination of individuals that results in households containing either 4 adults (person type 1 to 6) or 3 adults (person type 1 to 6) and 1 child (person type 7 or 8).**

- **5 person households**
    - **Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than four children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains four children.**
    - **Household Type 8: 7 (child) or 8 (child – former adult) if random household contains less than three children; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains three children and less than two adults.**
    - **Household Type 9: Any random combination of individuals that results in households containing either 5 adults (person type 1 to 6), 4 adults (person type 1 to 6) and 1 child (person type 7 or 8) or 3 adults (person type 1 to 6) and 2 children (person type 7 or 8).**
- **6 person households**
    - **Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than five children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains five children.**
    - **Household Type 8: 7 (child) or 8 (child – former adult) if random household contains less than four children; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains four children and less than two adults.**
    - **Household Type 9: Any random combination of individuals that results in households containing either 6 adults (person type 1 to 6), 5 adults (person type 1 to 6) and 1 child (person type 7 or 8), 4 adults (person type 1 to 6) and 2 children (person type 7 or 8) or 3 adults (person type 1 to 6) and 3 children (person type 7 or 8).**
- **7 person households**
    - **Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than six children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains six children.**
    - **Household Type 8: 7 (child) or 8 (child – former adult) if random household contains less than five children; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains five children and less than two adults.**
    - **Household Type 9: Any random combination of individuals that results in households containing either 7 adults (person type 1 to 6), 6 adults (person type 1 to 6) and 1 child (person type 7 or 8), 5 adults (person type 1 to 6) and 2 children (person type 7 or 8), 4 adults (person type 1 to 6) and 3 children (person type 7 or 8) or 3 adults (person type 1 to 6) and 4 children (person type 7 or 8).**

- 8 person households
    - Household Type 6: 7 (child) or 8 (child – former adult) if random household contains less than seven children; 1 (adult to age 65 either alone or not) or 3 (65 and over either alone or not) if household already contains seven children.
    - Household Type 8: 7 (child) or 8 (child – former adult) if random household contains less than six children; 1 (adult to age 65 either alone or not), 3 (65 and over either alone or not), 5 (adult to age 65 together) or 6 (65 and over together) if household already contains six children and less than two adults.
    - Household Type 9: Any random combination of individuals that results in households containing either 8 adults (person type 1 to 6), 7 adults (person type 1 to 6) and 1 child (person type 7 or 8), 6 adults (person type 1 to 6) and 2 children (person type 7 or 8), 5 adults (person type 1 to 6) and 3 children (person type 7 or 8), 4 adults (person type 1 to 6) and 4 children (person type 7 or 8) or 3 adults (person type 1 to 6) and 5 children (person type 7 or 8).

The assignment of individuals to communal establishments uses a subset of all individuals previously randomly selected to belong in these establishments and a subset of the Household Types already identified as being communal establishments (Household Type 10). From these subsets, an individual not already assigned is selected at random and added to a random communal establishment. This process continues until all individuals are assigned. As the total number of people in each communal establishment is unknown (unless there is only one in the MSOA), there are no restrictions on what the minimum and maximum number of people in each communal establishment are.

Any remaining unallocated adults and children are added to households containing 8 people (as at this point the maximum non-communal establishment household size is 8). Any remaining adults are in the first instance distributed among only Household Type 9 households. If none of these exist with 8 people in the MSOA then the excess adults are distributed amongst Household Types 6 and 8 that have 8 people (or if there are no 8 person sized households in that MSOA, the largest sized household that does exist). The same process applies to any excess children, except there is no initial restriction to being allocated to only Household Type 9.

This completes the allocation of individuals to the different Household Types and household person sizes. The output of this process is an amended data frame based on the input microdataset, called 'MyData225'. This contains the same 56,075,912 rows as the input microdataset, with several additional columns. Each individual has now been assigned to a Household Type (1 to 10), a Household ID (H1 to H23425076), a count (all values should be 1) and an allocation (used during the

assignment process to make sure the individual was assigned into the correct Household Type). The count and assignment columns are no longer required.

The 'DataCategories' data frame has also been appended, and for each of the 23,425,076 rows three columns have been added: Count, Children and Adults. The Count column gives the total number of people assigned to that household and the Children and Adults give the total number of children and adults in the household.

A consequence of the random assignment of people into households is that the number of adult male-male or female-female pairs in 2 adult households is higher than in reality. To counter balance this, a swapping process is applied in each MSOA. The number of male-male and female-female adult pairs in Households Type 3 (two adults - to age 65), 4 (two adults – 65 and over), 7 (two adults and 1 child) and 8 (two adults and 2 or more children) in each MSOA is counted. The value of the pairs that has the smallest sum is used to swap that number of male/females with the same number of the opposite sex from the pairing with the highest sum. So, for example, if there are 200 male-male pairs and 250 female-female pairs, then 200 males will be swapped with 200 females. This will create an additional 200 male-female pairs and leave 0 male-male pairs and 50 female-female pairs. If the number of pairs is the same then the same swapping process is undertaken, but the result will be no male-male or female-female pairs will remain. This process is undertaken for each of the applicable four Household Types in turn. The result of this swapping process is an amended 'MyData225' that contains the same assignments as before, but with a more realistic gender balance between the adult couples.

2. ***Allocation of households to synthetic address locations – matching households to postcode population and household counts, communal establishment counts and the residential settlement mask to generate addresses***

The following outlines the steps taken to assign households and communal establishments (and individuals) in the synthetic population dataset to unit postcode and XY coordinates. These steps take place once the individuals in the MSOA level synthetic population dataset has been assigned to a Household Type and given a Household ID using the previously outlined method (section 1). Table 9 outlines the datasets to be used in this part of the data processing.

Table 7: Example records for dataset 3 (Communal Establishment counts)…

date	geography	geography code	Rural Urban	Communal Establishment Type: All categories: Communal establishment management and type; measures: Value	Communal Establishment Type: Medical and care establishment: Total; measures: Value	Communal Establishment Type: Medical and care establishment: NHS: Total; measures: Value	Communal Establishment Type: Medical and care establishment: NHS: General hospital; measures: Value	Communal Establishment Type: Medical and care establishment: NHS: Mental health hospital/unit (including secure units); measures: Value	Communal Establishment Type: Medical and care establishment: NHS: Other hospital; measures: Value	Communal Establishment Type: Medical and care establishment: Local Authority: Total; measures: Value	Communal Establishment Type: Medical and care establishment: Local Authority: Children's home (including secure units); measures: Value	Communal Establishment Type: Medical and care establishment: Local Authority: Care home with nursing; measures: Value
2011	E00082111	E00082111	Total	0	0	0	0	0	0	0	0	0
2011	E00082113	E00082113	Total	0	0	0	0	0	0	0	0	0
2011	E00082114	E00082114	Total	0	0	0	0	0	0	0	0	0
2011	E00082115	E00082115	Total	0	0	0	0	0	0	0	0	0
2011	E00082118	E00082118	Total	0	0	0	0	0	0	0	0	0
2011	E00082112	E00082112	Total	0	0	0	0	0	0	0	0	0

…Table 7 (continued): Example records for dataset 3 (Communal Establishment counts)…

Communal Establishment Type: Medical and care establishment: Local Authority: Care home without nursing; measures: Value	Communal Establishment Type: Medical and care establishment: Local Authority: Other home; measures: Value	Communal Establishment Type: Medical and care establishment: Registered Social Landlord/Housing Association: Total; measures: Value	Communal Establishment Type: Medical and care establishment: Registered Social Landlord/Housing Association: Home or hostel; measures: Value	Communal Establishment Type: Medical and care establishment: Registered Social Landlord/Housing Association: Sheltered housing only; measures: Value	Communal Establishment Type: Medical and care establishment: Other: Total; measures: Value	Communal Establishment Type: Medical and care establishment: Other: Care home with nursing; measures: Value	Communal Establishment Type: Medical and care establishment: Other: Care home without nursing; measures: Value	Communal Establishment Type: Medical and care establishment: Other: Children's home (including secure units); measures: Value	Communal Establishment Type: Medical and care establishment: Other: Mental health hospital/unit (including secure units); measures: Value	Communal Establishment Type: Medical and care establishment: Other: Other hospital; measures: Value	Communal Establishment Type: Medical and care establishment: Other: Other establishment; measures: Value	Communal Establishment Type: Other establishment: Total; measures: Value
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

…Table 7 (continued): Example records for dataset 3 (Communal Establishment counts)

Communal Establishment Type: Other establishment: Defence; measures: Value	Communal Establishment Type: Other establishment: Prison service; measures: Value	Communal Establishment Type: Other establishment: Approved premises (probation/ bail hostel); measures: Value	Communal Establishment Type: Other establishment: Detention centres and other detention; measures: Value	Communal Establishment Type: Other establishment: Education; measures: Value	Communal Establishment Type: Other establishment: Hotel: guest house; B&B; youth hostel; measures: Value	Communal Establishment Type: Other establishment: Hostel or temporary shelter for the homeless; measures: Value	Communal Establishment Type: Other establishment: Holiday accommodation (for example holiday parks); measures: Value	Communal Establishment Type: Other establishment: Other travel or temporary accommodation; measures: Value	Communal Establishment Type: Other establishment: Religious; measures: Value	Communal Establishment Type: Other establishment: Staff/worker accommodation only; measures: Value	Communal Establishment Type: Other establishment: Other; measures: Value	Communal Establishment Type: Establishment not stated; measures: Value
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Table 8: Example records for dataset 4 (Postcode household counts)

Postcode	Total	Males	Females	OSLAUA	Occupied_ Households	OSEAST	OSNRTH	OA2011	LSOA2011	MSOA2011
BH235QQ	6	2	4	E07000091	4	422230	94794	E00116842	E01022996	E02004794
BH237AG	16	9	7	E07000091	4	417995	97173	E00116840	E01022994	E02004794
BH237AH	4	2	2	E07000091	1	418046	97114	E00116840	E01022994	E02004794
BH237AJ	28	13	15	E07000091	10	417938	97340	E00116840	E01022994	E02004794
BH237AT	84	43	41	E07000091	8	415821	96797	E00117196	E01023063	E02004794
BH237AU	16	6	10	E07000091	8	415658	96741	E00117196	E01023063	E02004794
BH237AX	24	13	11	E07000091	9	415652	96997	E00117196	E01023063	E02004794

Table 9: Datasets for assignment of households and communal establishments to unit postcodes and XY coordinates

Dataset	Contents / Description
SYNTH_POP	56,075,912 unique individual records in the synthetic population dataset. Each individual is assigned to an MSOA, Household Type and Household ID.
OA_HHLD_TYPE	Counts of the 10 different Household Types used in NCRM processing for every OA in England and Wales.
OA_MSOA	A lookup for every OA in England and Wales to LSOA, MSOA, Local Authority and English Region.
OA_COMMUNALS	Total numbers of communal establishments and communal population as recorded by the 2011 Census in England and Wales at the OA level.
PCD_HHLD_Counts	Total numbers of people, males, females and occupied households at the unit postcode level for England and Wales. (This is not a full list of all unit postcodes active on census day in 2011, only the 1,308,776 that contained a population).
OA_PWC_PCD_Intersect	Out of the 181,408 OAs in England and Wales 70 do not contain any unit postcode centroids within their boundaries. To allocate OA populations to unit postcodes every OA in England and Wales we require it to have at least one associated unit postcode. To achieve this Thiessen polygons were constructed using the centroids of unit postcodes from 'PCD_HHLD_Counts'. The intersection of each OA's population weighted centroid with the unit postcodes Thiessen polygons means all the 181,408 OAs can be assigned to at least one unit postcode. As some unit postcode centroids share the same XY coordinates, some OAs have more than one unit postcode assigned to them in this dataset (this is performed as an external GIS operation).

MSOA to OA allocation

The first step is to ensure that the same geographies are included across the datasets so they can be merged and integrated with ease and also to aggregate some categories in some of the datasets. 'PCD_HHLD_Counts' is merged with 'OA_MSOA' so the 2011 Census postcode counts for England and Wales have OA, LSOA, MSOA, LA and Region codes assigned to them. 'OA_COMMUNALS' is then merged with 'OA_MSOA' so the OAs have LSOA, MSOA, LA and Region codes assigned to them. 'OA_HHLD_TYPE' is merged with 'OA_MSOA' so the OA level counts of Household Types have LSOA, MSOA, LA and Region codes assigned to them. The postcode level counts of individuals and households ('PCD_HHLD_Counts') are aggregated to OAs (to match 'OA_COMMUNALS'). There are 70 OAs that do not contain any postcodes with populations, and these must be added to the OA

level dataset with zero population totals. Within this part of the coding and allocation there is a substantial amount of cross-checking of household and population counts.

The 'SYNTH_POP' dataset is next aggregated by unique Household ID. For each Household ID the total number of people in the household is calculated (i.e. the number of times that Household ID appears in the 'SYNTH_POP' dataset) and what Household Type it belongs to. This new dataset output is called 'SYNTH_POP_HHLD_ID'. The 'OA_HHLD_TYPE' dataset is then recast so that each instance of each Household Type in every OA is assigned a row. For example, as OA E00085919 contains 17 Household Type 1's a total of 17 unique rows were created. This new dataset is called 'AREA_OA_HOUSEHOLDS'.

'SYNTH_POP_HHLD_ID' (derived from the synthetic population dataset) contains 23,423,261 unique households while 'AREA_OA_HOUSEHOLDS' (derived from 2011 Census data) contains 23,425,076. The difference arises from there being 1,815 fewer communal establishments in the 'SYNTH_POP_HHLD_ID' dataset, as the synthetic population creation had to deal with a greater number of communal establishments in some MSOAs than there are individuals who reside in them (i.e. 20 establishments with zero population or 100 establishments with a population of 90). In these circumstances the entire communal establishment population of an MSOA was assigned to a single communal establishment. This results in some communal establishments recorded in the 2011 Census not being included in the synthetic dataset.

Apart from the 1,815 fewer households of Type 10 (communal establishments), there are the same number of households for each Household Type in 'SYNTH_POP_HHLD_ID' and 'AREA_OA_HOUSEHOLDS'). A simple merge can therefore be performed to allocate the Household IDs in 'SYNTH_POP_HHLD_ID' to the unique households in 'AREA_OA_HOUSEHOLDS' using the steps outlined below:

- The row order of the 'SYNTH_POP_HHLD_ID' and 'AREA_OA_HOUSEHOLDS' datasets are randomised.
- A new sorting column is added to 'AREA_OA_HOUSEHOLDS' from 2 to 23,425,077 (i.e. 1 to the number of rows in 'AREA_OA_HOUSEHOLDS' plus one).
- As there would be 1,815 unmatched communal establishments at the end of the merge process, priority in the matching process is given to OAs that contained unit postcodes with no occupied households but containing a population (suggesting the population of that postcode lived in communal establishments). Across England and Wales there are 4,401 such unit postcodes in 3,845 OAs. Any household assigned to Household Type 10 in 'AREA_OA_HOUSEHOLDS' in one of these 3,845 OAs is assigned '1' in the sorting column. 'AREA_OA_HOUSEHOLDS' is then ordered by the sorting column. This means communal

establishments in these 3,845 OAs are at the top of the dataset for the merge process which operates from top to bottom.

- 'SYNTH_POP_HHLD_ID' and 'AREA_OA_HOUSEHOLDS' both have a new column added 'I' (for "iteration"). This value from 1:n creates a unique number sequence based on the MSOA code and Household Type. For example, the first instance of E02004754 and 'Household Type 1' was numbered 1 in 'I', and the second instance numbered 2 and so on.
- Adding this column means 'SYNTH_POP_HHLD_ID' and 'AREA_OA_HOUSEHOLDS' can be merged based on three identifiers: MSOA code, Household Type and I. The combination of these three identifiers creates a unique reference for each household in both datasets.

This process allows all 23,423,261 households in the 'AREA_OA_HOUSEHOLDS' dataset to be assigned to a unique Household ID derived from the synthetic population dataset. The additional 1,815 Household Type 10 (communal establishments) that are not in the synthetic dataset remain in the 'AREA_OA_HOUSEHOLDS' dataset but are not allocated a Household ID.

Priority in the matching process of communal establishments was given to the 3,845 OAs across England and Wales that contained unit postcodes with no occupied households but did contain a population. Of these 3,845 OAs, 5 did not have any communal establishments assigned to them after the merge process. This is because the MSOAs to which these OAs belonged had fewer communal establishments assigned to them in the synthetic population dataset compared to the total number recorded in the 2011 Census. This issue is explained in the 'Limitations with communal establishment assignment' section below.

After the MSOA to OA allocation the 'AREA_OA_HOUSEHOLDS' dataset contains 23,425,076 unique households of which 23,423,261 have been assigned to a unique Household ID based on the synthetic population dataset. This means each of the 23,425,076 unique households from the synthetic population dataset are allocated to one of 181,408 OAs in England and Wales.

OA to Unit Postcode Allocation

The basis of the OA to unit postcode allocation are the values in 'PCD_HHLD_Counts'. The synthetic population is distributed among the 1,308,776 postcodes included in this dataset. The number of occupied households in 'PCD_HHLD_Counts' are aggregated to OA level and the number of households in 'AREA_OA_HOUSEHOLDS' are also aggregated to OA level. These two aggregated datasets are combined and the difference calculated for each OA. OA values calculated from 'AREA_OA_HOUSEHOLDS' are treated as correct as they are derived from the synthetic population dataset. Positive differences indicate more households need to be added to the postcodes in that OA, while negative values indicate households need to be removed.

As no unit postcode level counts of communal establishments are available, the total number of communal establishments in each OA as derived from the synthetic population dataset are randomly distributed among the postcodes in 'PCD_HHLD_Counts' for that OA. Due to 70 OAs being included in the 'AREA_OA_HOUSEHOLDS' that are not in 'PCD_HHLD_Counts' (as described when discussing the 'OA_PWC_PCD_Intersect' dataset), additional postcode records are added to 'PCD_HHLD_Counts'. Using 'OA_PWC_PCD_Intersect', 172 unit postcodes for the 70 OAs are added to 'PCD_HHLD_Counts'. Adding these records means there are duplicates of unit postcodes in 'PCD_HHLD_Counts'. An additional column is added to 'PCD_HHLD_Counts' that classifies the postcode source as either being from the 2011 Census records ('PCD_HHLD_Counts') or from the population weighted centroid and Thiessen polygon intersection ('OA_PWC_PCD_Intersect'). The revised 'PCD_HHLD_Counts' has 1,308,948 rows, with each row containing a unique unit postcodes and OA combination.

As the revised 'PCD_HHLD_Counts' now contains at least one unit postcode for every OA, households and communal establishments can be distributed from OAs to unit postcodes. The rules used for this process are:

- If the number of households needs to increase in an OA then only postcodes which already contain occupied households (according to 2011 Census records) are used. This means unit postcodes which contain a population but have no occupied households do not have any households assigned to them (the assumption being that these postcodes contain communal establishments). After the allocation process the 4,484 unique unit postcodes that contained no occupied households in the 2011 Census records still had no households assigned to them.
- If the number of households needs to decrease in an OA then priority is given to decreasing the counts of postcodes with more than one occupied household. As long as it was mathematically possible, then all postcodes that contain at least one occupied household in the 2011 Census records will retain at least one household. If, however, this is not possible then some unit postcodes will have all households removed from them. However, all unit postcodes that contained at least one occupied household in the 2011 Census records (1,304,292 of them) retained at least one household after the allocation process.
- The assignment of communal establishments to unit postcodes prioritises postcodes that contain a population but have no occupied households in the 2011 Census records. Once all postcodes without any occupied households have been assigned at least one communal establishment then the remainder of the communal establishments are randomly distributed among all the unit postcodes in the OA. If there are more unit postcodes that contain people but have no occupied households compared to the number of communal establishments needed to be assigned then a random selection of these unit postcodes is used. This resulted in some postcodes not being assigned any communal establishments (or households) despite the 2011 Census records indicating they contained a population.

The result of the allocation process is that every unit postcode in 'PCD_HHLD_Counts' is assigned a total number of households and communal establishments based on the synthetic population dataset. Due to the allocation process, some unit postcodes included in 'PCD_HHLD_Counts' did not have any households or communal establishments assigned to them. In addition, as some duplicated postcodes were required through the use of 'OA_PWC_PCD_Intersect', some unit postcodes were duplicated in the dataset. In total, 1,308,740 unique unit postcodes contained at least one household or one communal establishment once the assignment process had completed. Matching the number of households and communal establishments per unit postcode in 'PCD_HHLD_Counts' with the Household Type in 'AREA_OA_HOUSEHOLDS' requires the modification of 'PCD_HHLD_Counts'. 'PCD_HHLD_Counts' is recast so each individual household or communal establishment is given its own row in the dataset. This creates a dataset with 23,423,261 records which matches the number of unique households in 'AREA_OA_HOUSEHOLDS' that have been assigned characteristics from the synthetic population dataset.

As 'PCD_HHLD_Counts' and 'AREA_OA_HOUSEHOLDS' both have the same number of rows they can be matched using each row's OA reference. A new column is created in both datasets, 'OA_ID', which contains the OA code and the iteration of that OA each dataset i.e. E00000001_1, E00000001_2 etc. to create 23,423,261 unique IDs in both datasets. The two datasets are then merged using this unique ID to create a new dataset, 'SYNTH_HHLD_PCD_ALLOCATION'. This dataset contains 23,423,261 unique household and communal establishments, with each one assigned to a Household Type and one of 1,308,740 unit postcodes.

Unit Postcode to XY Coordinates Allocation

As the total number of households and communal establishments based on the synthetic population dataset is calculated at the unit postcode level, the next stage is to give each of these households or communal establishments in 'SYNTH_HHLD_PCD_ALLOCATION' unique XY coordinates (the coordinates are OS Eastings and OS Northings, here referred to as XY coordinates). To achieve this, a Shapefile (ESRI, 1998) containing all the required postcodes is used. This Shapefile contains building outlines that cover the majority of unit postcodes (an alternative of building outlines is used in postcodes without any permanent buildings). Within the building boundaries in each unit postcode, the 'spsample' function from the 'sp' package in R randomly allocates the required number of points and records the XY coordinates. This means the assignment of XY coordinates in each unit postcode could only occur in places where the population could be (i.e. not in fields, parks, woods etc.).

The following steps have been followed to create the Shapefile used in this process:

- Thiessen polygons are created around the 1,465,006 postcodes active in March 2011. All active postcode centroids are used to ensure the area for each unit postcode was as realistic as possible. There are no definitive spatial boundaries for unit postcodes, hence the need for this approximation using Thiessen polygons.
- OpenPopGrid (a 10x10m raster layer (Heywood et al. 2002)) is imported along with OS Open Roads. Both are converted to 5x5m raster layers. OpenPopGrid (Murdock et al., 2015) was selected as it had already been processed to remove commercial buildings.
- The cells of the OS Open Roads raster layer are then subtracted from the OpenPopGrid layer (to remove any sections of the rasterised buildings which overlap with roads).
- The OpenPopGrid 5x5m raster layer, minus roads, is then converted to a vector layer and intersected with the postcode Thiessen polygons. The buildings that are within each postcode polygon are then merged into a single object, resulting in one or more separate building features within in each postcode polygon being treated as a single object.
- Of the 1,308,740 unit postcodes that needed to be processed, 2,419 were not included in this dataset because they contained no buildings in the OpenPopGrid dataset.
- To supplement the OpenPopGrid buildings three additional data sources are used: 'OS VectorMap District', 'OS Open Map – Local' and 40x40 metre squares. Each of these sources are converted into 5x5m raster layers and have the rasterised version of OS Open Roads subtracted from them. They are then vectorised and intersected with the Thiessen polygons with the separate buildings in each polygon being merged into a single object.
- OS VectorMap District is used in the first instance, as this derives from the same data source as OpenPopGrid and provides a more generalised outline of buildings. This resulted in 516 additional postcode polygons having some building outlines within their boundaries.
- OS Open Map – Local was then checked against the remaining 1,903 postcode polygons. The buildings outlines in this data source are provided at a finer spatial scale, therefore smaller buildings which may not have been in OS VectorMap District may be included.
- The OpenPopGrid, VectorMap and Open Map Local datasets do not all align precisely in time with one another or with the 2011 Census, but our objective here is to provide the plausible locations for all synthetic households, within areas most likely to contain residential buildings. This resulted in a further 1,719 postcode polygons having some form of building outlines within their boundaries.
- Finally, the remaining 184 postcode polygons without any buildings in their boundary are assigned a 40x40 metre square centred on unit postcode's centroid. These typically relate to unit postcodes that have no permanent buildings, such as those containing marinas (i.e. localities with residential populations but no buildings present in any of the map data sources). This resulted in 184 postcode polygons having a single 40x40 metre square to assign the XY coordinates.

- **The final Shapefile contained building outlines, or a 40x40 metre square, for each of the 1,308,740 unit postcodes. With 1,306,321 (99.8%) of unit postcodes using OpenPopGrid data, 516 (0.04%) of unit postcodes using OS VectorMap District data, 1,719 (0.13%) of unit postcodes using OS Open Map – Local data and 184 (0.01%) of unit postcodes being assigned to a 40x40 metre square.**

Each unit postcode in the Shapefile is read in turn and a set of random XY coordinates are generated. Rather than process all of England and Wales in one go, the task is divided into the English (former Government Office) Regions and Wales. This makes it easier to process in R due to only having to import a smaller section of the Shapefile at any one time. Random XY coordinates are generated for each of the 23,423,261 households and communal establishments within the 1,308,740 unique unit postcodes, recorded in a dataset called 'SamplePoints_Regions'. This contains three columns: unit postcode, X coordinate and Y coordinate. A new ID column, 'PCD_ID', is added to 'SYNTH_HHLD_PCD_ALLOCATION' and 'SamplePoints_Regions' containing the unit postcode (without any spaces) and a sequential number within that unit postcode in each dataset i.e. AL100AB_1, AL100AB_2 etc. This creates 23,423,261 unique ID's that can be matched together.

'SYNTH_HHLD_PCD_ALLOCATION' and 'SamplePoints_Regions' are merged together based on the shared 'PCD_ID' column to create 'HHLD_Synth_Geo_Allocation_to_Postcodes', containing 23,423,261 rows, with each row having a unique set of XY coordinates, Household ID, and PCD ID; and a non-unique Household Type, unit postcode, OA, LSOA, MSOA, LA and Region. The number of people in each household is also recorded (as this was calculated earlier when 'SYNTH_POP_HHLD_ID' was created). The 'HHLD_Synth_Geo_Allocation_to_Postcodes' dataset represents the final stage of allocation of 23,423,261 unique households in the synthetic population dataset from the MSOA level to unit postcode.

Household Unit Postcode and XY Coordinates to Individual Household Unit Postcode and XY Coordinates Allocation

To assign the details of each household or communal establishment to each of the 56,075,912 individuals from the 'SYNTH_POP' dataset the 'HHLD_Synth_Geo_Allocation_to_Postcodes' dataset is recast and a new ID column added to both datasets.

The 'HHLD_Synth_Geo_Allocation_to_Postcodes' dataset is recast so that each individual in each unique household or communal establishment is assigned their own row. For example, as Household 'H18446750' contains 4 people a total of 4 rows were created. This new dataset is called 'HHLD_Synth_Geo_Allocation_to_Postcodes_Expanded' and contains 56,075,912 rows.

The new ID column, 'UID', added to 'SYNTH_POP' and 'HHLD_Synth_Geo_Allocation_to_Postcodes_Expanded' is based on the Household ID. It consists of the Household ID and the iteration of that Household ID in each dataset. For example, the 4-people living in Household 'H18446750' are assigned: 'H18446750_1', 'H18446750_2', 'H18446750_3' and 'H18446750_4'. This creates a unique ID for the 56,075,912 rows in both datasets.

Using 'UID' the 'SYNTH_POP' and 'HHLD_Synth_Geo_Allocation_to_Postcodes_Expanded' datasets are merged together to create the 'HHLD_Synth_Geo_Allocation_to_Indivduals' dataset. This consists of 56,075,912 rows and contained the columns from 'SYNTH_POP' with the unit postcode and XY coordinates of each individual household or communal establishment appended (along with OA, LSOA, MSOA, LA and Region geographies). Table 10 provides an overview of the outputs files.

Table 10: Output Files (from R processing)

Dataset output from process	Contents / Description
HHLD_Census_and_Synth_ Allocation_Postcode_Counts	Population, household and communal establishment counts for unit postcode in England and Wales derived from 2011 Census records and the synthetic population dataset.
HHLD_Synth_Geo_Allocation_t o_Individuals	Details of the 56,075,912 unique individuals in the England and Wales synthetic population dataset, including the Household Type, Household ID each individual has been assigned to and the geographies and XY coordinates of the household each individual has been assigned to.
HHLD_Synth_Geo_Allocation_t o_Postcodes	Details of the 23,423,261 unique households in the England and Wales synthetic population dataset, including the Household Type, Household ID, geographies and XY coordinates that each household has been assigned to.

Limitations to communal establishment assignment

As already detailed, the number of communal establishments in the synthetic dataset do not match the 2011 Census records for certain MSOAs. This variation is most pronounced in MSOAs where there are a greater number of communal establishments than the communal population. The rule that if there were more communal establishments than people in an MSOA then the entire communal population would be assigned to a single communal establishment means that 108 MSOAs in England and Wales (out of 7,201) are assigned a single communal establishment in the synthetic dataset despite having more than one communal establishment according to 2011 Census records.

This rule has had an impact on the assignment of households and individuals to unit postcodes, for example, the 2011 Census records for MSOA E02000970 indicate a communal population of 67 and 100 communal establishments. Records from the 2011 Census show that out of the 35 OAs that are

within MSOAE02000970, 6 of them contained communal establishments with a population (four of these six did not contain any households), plus 16 are recorded as having communal establishments without a population. These six OAs with a communal establishment population contain a total of 234 census day postcodes in the dataset. Due to the aforementioned rule, only one of these OAs and one postcode has been assigned a communal establishment, which cannot be said to be representative of the spatial distribution of communal establishments in MSOA E02000970.

The reality for MSOA E02000970 is that on the 2011 Census day anywhere between 1 and 67 of the recorded communal establishments contained a population, with anywhere between 33 and 99 not having any population. This uncertainty caused by having a greater number of communal establishments than communal population means the spatial distribution of the individuals in MSOA E02000970 who live in communal establishments is likely to be less reflective of reality than the individuals in MSOA E02000970 who do not live in communal establishments. The same will apply to the other 107 MSOAs in England and Wales with a greater number of communal establishments than communal population, although this accounts for only 1.5% of MSOAs in England and Wales. There is no way in which the inconsistent census counts can be fully reconciled in these situations.

6. Results

Spatial distribution of household types

Figure 2 shows the spatial distribution of household types for the western part of Southampton while Figure 3 shows a larger scale map for the locality of St. Denys, Portswood and Bitterne Park allowing a more detailed view of the distribution of the synthetic households. The nested structure of the households is evident with a mix of households at the OA level evident. Given the data is constrained by OAs there is perhaps an overconcentration of household types, although the pattern matches the required totals and the household distribution is plausible. These aspects are very positive in relation to use of the data for evaluating access measures such as distance to a GP or local school so the comparability to genuine census data makes the dataset a promising basis for intended automated zone design and other experiments. In Figure 3 the degree of household clustering is clear and so too is the way in which industrial locations are left blank, for example the grey building outline with no households located in the north central area of the map is a supermarket and to the south of the map the location with large buildings and no households is an industrial estate.

Figure 2: Distribution of household types in the synthetic dataset

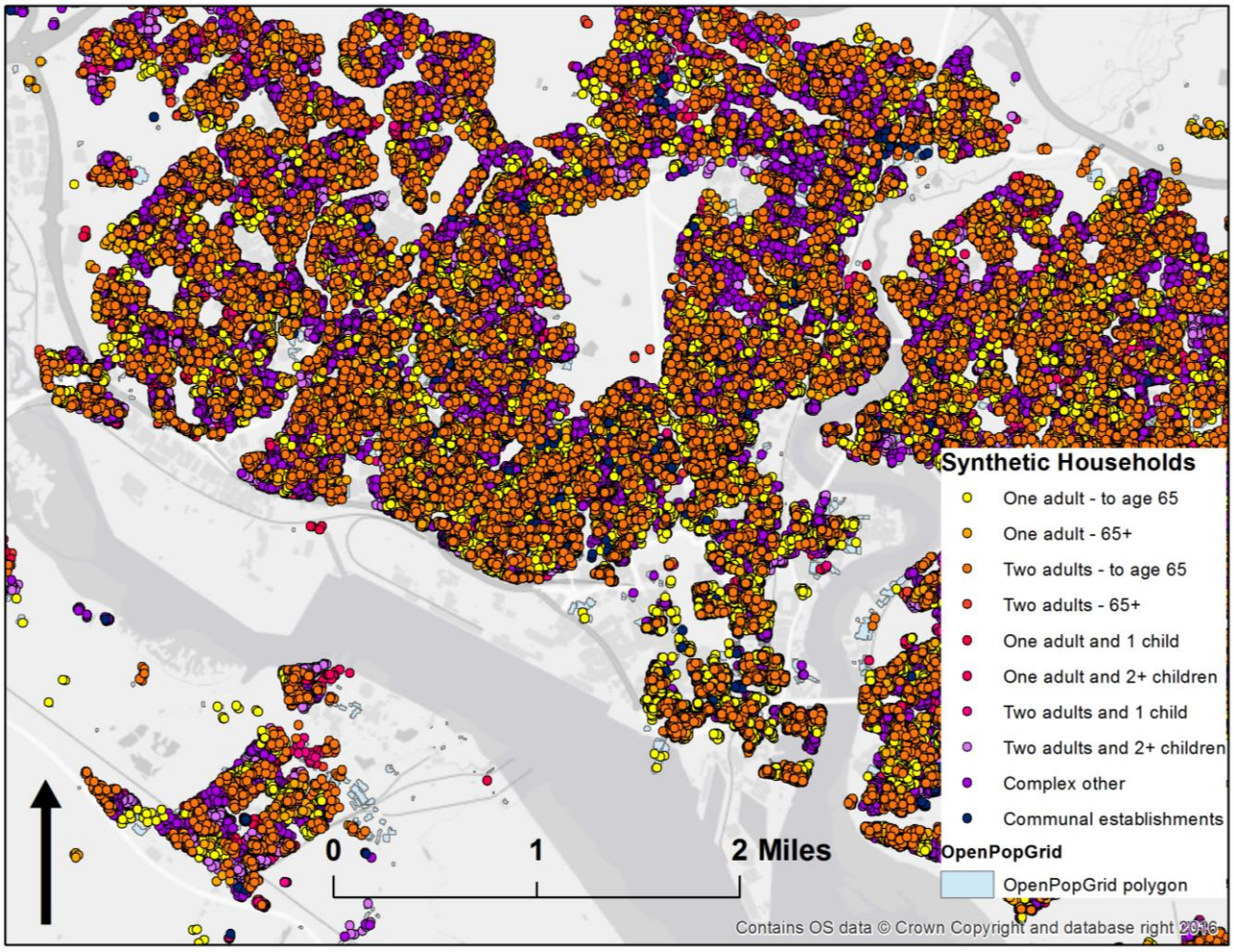
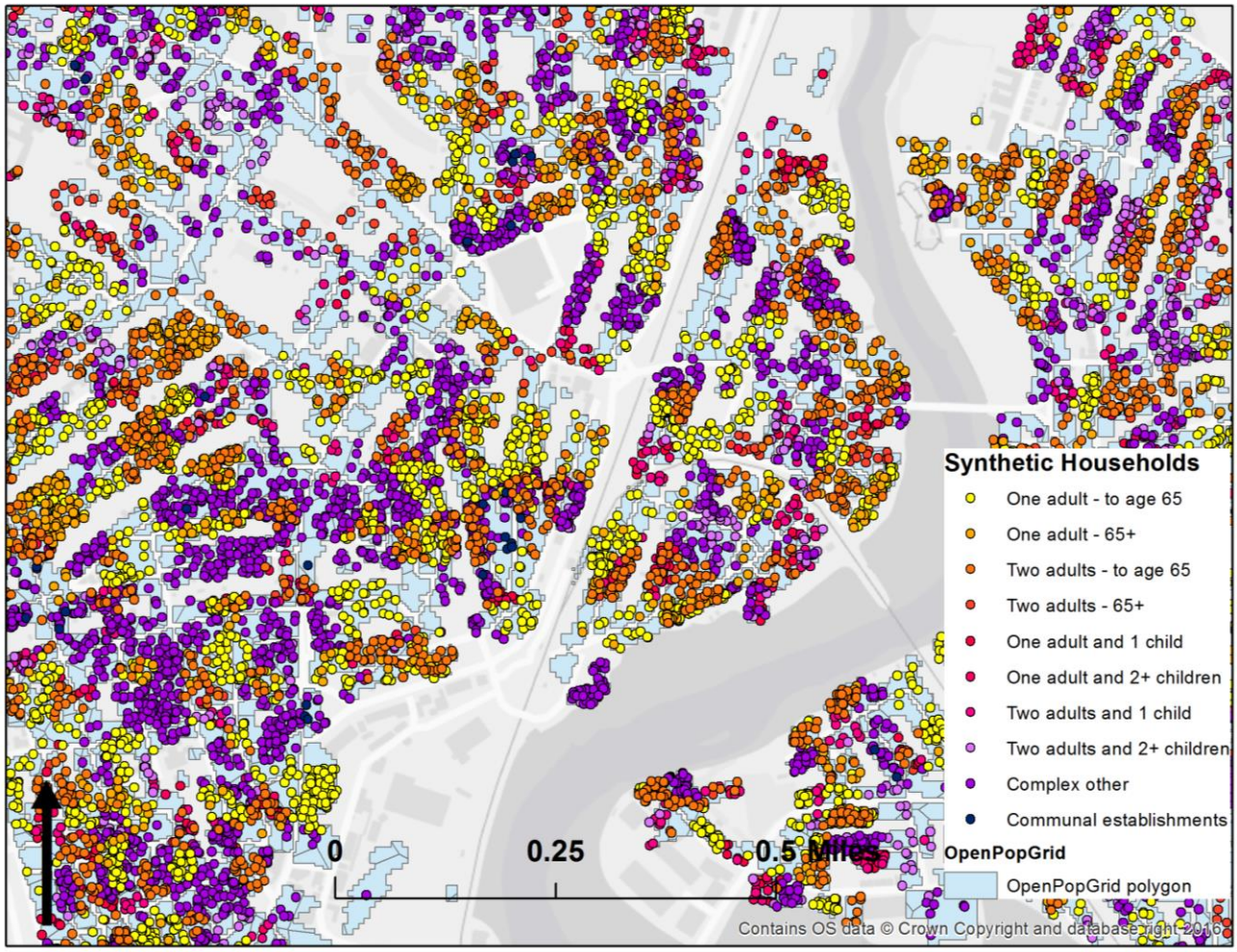


Figure 3: Distribution of household types in the synthetic dataset – small area case study



Comparison of synthetic data to aggregate 2011 Census data

In order to assess the representativeness of the new synthetic dataset, comparisons may be made between census aggregate data and re-aggregation of key variables from the new synthetic dataset. These include variables which have been used to determine the construction of the dataset and also those which have not been used to constrain or to shape the data. Figure 4 presents a scatterplot of the number of people per OA in the new synthetic data compared to the 2011 Census data for Hampshire. The method is not expected to lead to perfect correspondence; not least because the synthetic address locations nested within postcodes are not constrained in any way to fall within the boundaries of their 'true' OAs. We would anticipate that variables involved in the microsimulation will be more closely associated with the census aggregate values.

The first two of our four comparison variables are counts of persons per OA in Figure 4 and households per OA in Figure 5. For these two variables we see clustering of points and a high degree of comparability between the two because these variables have constrained the construction of the synthetic data. Figures 6 and 7 compare two variables not involved with the construction of the dataset. For the percentage of an OA aged 65 and over (Figure 6) we see that there is a relatively

strong association, reflecting the importance of age in the design of the synthetic data to constrain totals for the two household types aged 65 and over: living alone or with another person aged 65+. Although not explicitly included in our processing, we would expect the structure of the synthetic microdata to largely replicate that found in the original census data. The final scatterplot (Figure 7) shows individuals reporting poor health, a variable entirely excluded from our processing. Here there is a weak association with the OA aggregate counts, although it should be noted that these values will still be correct at the MSOA level; it would appear that poor health is therefore subject to OA-level clustering that is not closely related to the household and person type structures used here. This suggests that other variables which have played no part in the processing will not be so accurately distributed at the OA level, whereas at the MSOA level the patterns in the original data will be preserved.

Figure 4: Scatterplot of OA person counts in synthetic and 2011 Census data, Hampshire

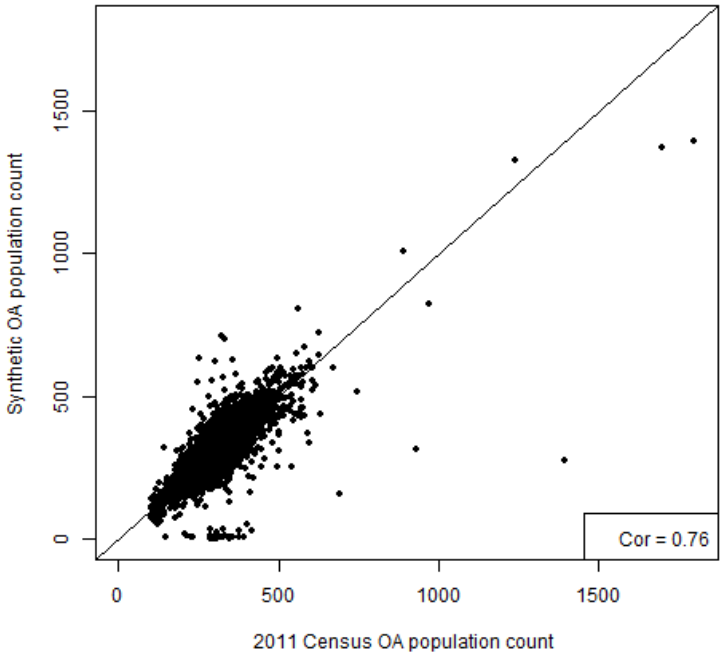


Figure 5: Scatterplot of OA household counts in synthetic and 2011 Census data, Hampshire

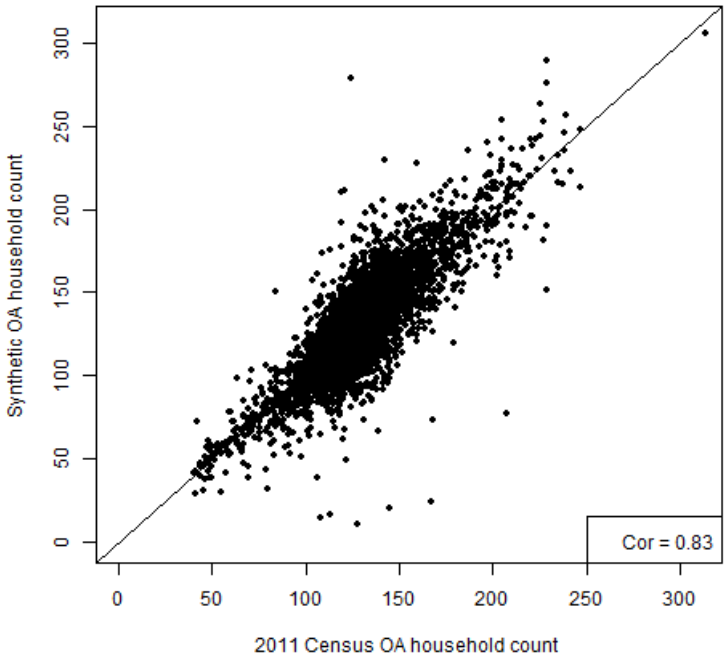


Figure 6: Scatterplot of OA percentage of population aged 65 and over in synthetic and 2011 Census data, Hampshire

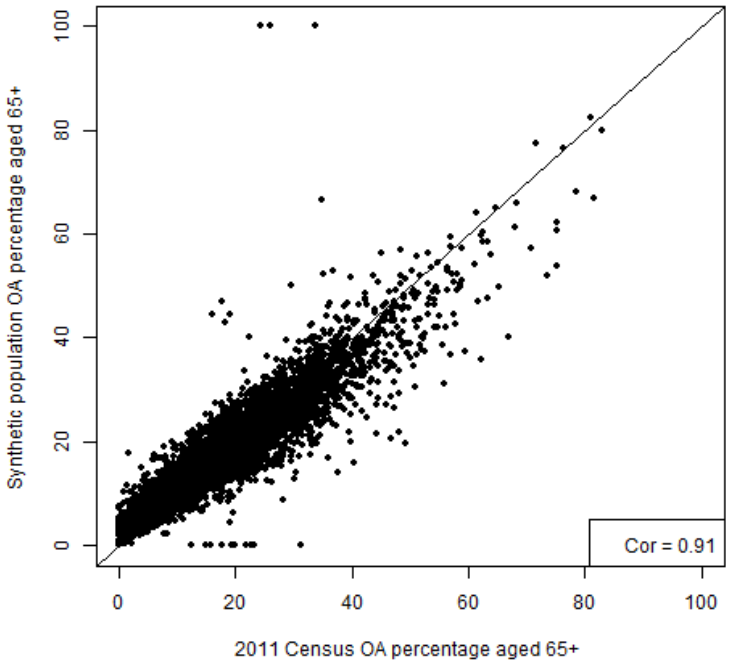
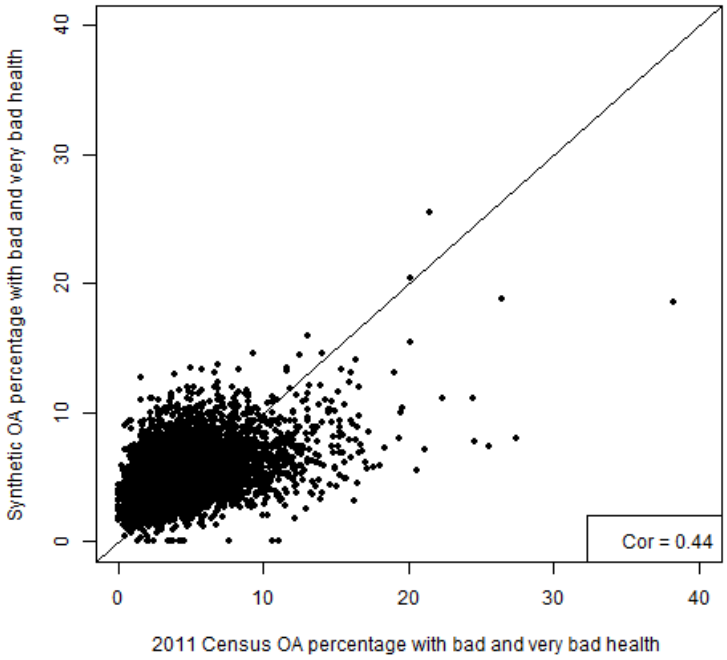


Figure 7: Scatterplot of OA percentage of population in bad or very bad health in synthetic and 2011 Census data, Hampshire



7. Discussion and Conclusion

The dataset which has been constructed combines synthetic population data at the MSOA level with other 2011 Census data and residential settlement data to arrive at a set of households, each with its own composition in keeping with the locality, positioned in areas which have a residential make-up. This paper has explained the rationale, design, input data and coding to develop a synthetic spatial microdataset suitable for zone design experiments and other spatially focussed tests. The design of the processing and account provided has been written in such a way that it is hoped the same methodology can be applied in other countries and contexts where similar data can be sourced. Data created for the present study has the potential to be used in many other studies in the geospatial aggregation and anonymisation of census / linked data topic areas.

We have illustrated the spatial distribution of the different household types at a local level and presented some examples of the levels of association between the new data and aggregate 2011 Census data at the OA level. In particular, these show that the OA level associations for some variables from the synthetic data and census data are strongly reflective of the census aggregate values while for others not included in the processing, the associations are weaker (although all original values are preserved at the MSOA level, reflecting the input data from CDRC). Importantly, the synthetic dataset has a highly realistic spatial distribution at the level of individual streets and buildings which is an essential requirement for any application which involves assessment of alternative zone designs and geoprivacy issues.

The input MSOA CDRC microdataset and the data resulting from the methodology outlined in this paper, deposited with CDRC, are both 'safeguarded' datasets within the CDRC classification of data (which also corresponds to the data levels described in the UK Data Service's three tier access policy (https://www.ukdataservice.ac.uk/get-data/data-access-policy). While not personally-identifiable, these safeguarded data have access restricted due to license conditions and in order to ensure that the synthetic populations are used in an appropriate research context. Access is available via a remote service with registration and project approval requirements. For further information on how to access CDRC data, visit: https://www.cdrc.ac.uk/wp-content/uploads/2015/07/User-Guide-V5.pdf.

The construction of the synthetic spatially referenced dataset with a realistic population make-up is of use, not only in relation to studies on protecting identities of individuals and households, but for further research on the construction of indicators and accessibility of local services. This dataset further highlights the potential for synthetic data to be used in studies on geoprivacy and the anatomy of disclosure risk.

Acknowledgements

David Martin and James Robards were supported by the National Centre for Research Methods, ESRC Award ES/L008351/1. David Martin and Chris Gale were supported by the Administrative Data Research Centre-England, ESRC Award ES/L007517/1.

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. Details of the construction of the synthetic population data can be found in Morris and Clark (in press). For more information on data being held by the ESRC CDRC please see https://data.cdrc.ac.uk/.

OpenPopGrid is licensed under the Open Database License. You are free to use under the terms of the OdBL licence subject to citation. Any rights in individual contents of the database are licensed under the Database Contents License.

Ordnance Survey - VectorMap District, StreetView © Crown copyright and database right 2014, 2016. OS OpenData is free to use under the Open Government Licence (OGL).

Office for National Statistics - 2011 Census Area boundaries, 2011 Census Postcode headcounts available under the Open Government Licence v3.0

Office for National Statistics - ONS Postcode Directory (Open) May 2011 © Crown copyright and database right 2014, © Royal Mail copyright and database right 2014

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 doi: 10.1068/a43601

Cooke, D.J. and Philip, L. 2001. To treat or not to treat? An empirical perspective. In: Hollin, C.R. ed. Handbook of offender assessment and treatment. Chichester: Wiley, pp. 3-15.

ESRI (1998) (Environmental Systems Research Institute) *ESRI Shapefile Technical Description: An ESRI White Paper*, July 1998 https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Fisher, P. F. and Langford, M. (1996) Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping, *The Professional Geographer,* 48(3): 299-309. http://dx.doi.org/10.1111/j.0033-0124.1996.00299.x

Heywood, I., Cornelius, S., and Carver, S. (2002) *An Introduction to Geographic Information Systems,* Second Edition, Prentice Hall, Harlow.

Martin, D., Nolan A. and Tranmer, M. (2001) The application of zone design methodology to the 2001 UK Census *Environment and Planning A* 33, 1949-1962 doi: 10.1068/a3497

Morris, M.A. and Clark, S. *In press.* A big data application of spatial microsimulation for neighbourhoods in England and Wales (Chapter 20). In Chen, Z. and Schintler, L. (Eds). *Big Data for Regional Science*, Routledge.

Murdock, A.P., Harfoot, A.J.P., 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/OpenPopGrid_ProductDocumentation.pdf

ONS (2012) *Statistical bulletin:2011 Census: Population and Household Estimates for Small Areas in England and Wales, March 2011* (23 November 2012) https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuspopulationandhouseholdestimatesforsmallareasinenglandandwales/2012-11-23

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

ONS (2014b) *ONS Innovation Laboratories* Working Paper Series No 1, ONS Methodology, Office for National Statistics, published December 2014 http://webarchive.nationalarchives.gov.uk/20160105160709/http:/www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/ons-working-paper-series/index.html

ONS (2014c) *Households and Household Composition in England and Wales: 2001-11,* Office for National Statistics, published 29 May 2014. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/articles/householdsandhouseholdcompositioninenglandandwales/2014-05-29#household-composition

ONS (2015a) *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 (October 2015).

ONS (2015b) *2011 Census Analysis: What Does the 2011 Census Tell Us About People Living in Communal Establishments?* Released: 11 February 2015 http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/census/2011-census-analysis/what-does-the-2011-census-tell-us-about-people-living-in-communal-establishments-/story-what-does-the-2011-census-tell-us-about-people-living-in-communal-establishments-.html

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.