Spatial Demography (2023) 11:10
https://doi.org/10.1007/s40980-023-00121-9

Linking Individuals to Areas: Protecting Confidentiality While Preserving Research Utility

Paul Norman[1] · **Jessie Colbert**[2] · **Daniel J. Exeter**[2]

Accepted: 21 October 2023

Abstract

Modern computational capabilities have brought about concerns about risks associated with the level of information disclosed in public datasets. A tension exists between making data available that protects the confidentiality of individuals while containing sufficiently detailed geographic information to underpin the utility of research. Our aim is to inform data collectors and suppliers about geographic choices for confidentiality protection and to balance this with reassurance to the research community that data will still be fit-for-purpose. We test this using simple logistic regression models, by investigating the interplay between two geographical entities (points for the observations and polygons for area attributes) at a variety of scales, using a synthetic population of 22,000 people. In an England and Wales setting, we do this for individuals located by postcodes and by postal sector and postal district centroids and link these to a variety of census geographies. We also 'jitter' postcode coordinates to test the effect of moving people away from their original location. We find a smoothing of relationships up the geographical hierarchy. However, if postal sector centroids are used to locate individuals, linkages to Lower/Medium Super Output Area scales and subsequent results are very similar to the more detailed unit postcodes. Postcode locations jittered by 500–750 m in any direction are likely to allow the same conclusions to be drawn as for the original locations. Within these geographic scenarios, there is likely to be a sufficient level of confidentiality protection while statistical relationships are very similar to those obtained using the most detailed geographic locators.

Keywords Geographical linkages and scale · Confidentiality and geoprivacy · Obfuscation · Simulation · Misclassification · Modifiable areal unit problem

Extended author information available on the last page of the article

Published online: 28 November 2023

1 Introduction

Geographic location is a critical consideration when integrating social science with epidemiological investigations to inform research and policy. There is strong evidence that the characteristics of particular locations and types of places influence variations in health and social outcomes geographically. For example, poorer people may have poorer access to local resources and facilities (Macintyre et al., 2008), there may be ethnic employment penalties in more deprived neighbourhoods (Jivraj & Alao, 2023), and public sports facilities may not be evenly accessible (Higgs et al., 2015).

Decisions made in the research process regarding how geographic location is incorporated into the data and analysis can directly affect the results of the research. When data is collected and disseminated for research purposes, geographical characteristics may be incorporated into the planning of data collection, and geographic variables may be included in the subsequent dissemination of the data released. For information collected about individuals (people or households), individual records may be linked to residential neighbourhoods, census areas, or administrative units such as local government areas. Individual level data may then be released at the neighbourhood or area level, for confidentiality reasons. However, the geographical boundaries chosen by suppliers to disseminate data can influence the results found in subsequent research. Data collection and dissemination must involve consideration of the different geographical scales available, their definition and the characteristics being measured to ensure facets of people's lives which influence a particular outcome are captured in enough detail to demonstrate spatial distributions, while minimising the risk of confidentiality breaches (Galster, 2001; Macintyre et al., 2002; Exeter et al., 2014; Petrovic et al., 2022). For area-level data, issues also revolve around geographic scale and how larger areas are subdivided into smaller units. The 'Modifiable Areal Unit Problem' (MAUP) warns that different conclusions may be drawn depending on which geographical zones are used in a study (Flowerdew et al., 2008; Openshaw, 1981).

These considerations are relevant for government agencies, data custodians and researchers. As data stewards, national statistical and data agencies (e.g. Office for National Statistics, Stats NZ, US Census Bureau, etc.) are responsible for aggregating individual-level data for public dissemination while ensuring small-cell counts in tables are minimised. For the research community, those focussed on place-based initiatives and 'putting people into place' (Entwisle, 2007) face the scientific art of defining what scale(s) constitute 'neighbourhoods' and how these impact on the granularity of the socio-demographic characteristics available for population research.

A range of relevant elements in the research process are about 'geography'. There may be a geography to the planning of data collection (census or survey strategy), or the geography used for the aggregation of administrative records, or the geography chosen for dissemination by suppliers or by researchers for analysis. In this paper, we aim to inform data collectors and suppliers about geographic choices for confidentiality protection and to balance this with information for the research

community as to whether the data specification will still be fit-for-purpose (Exeter et al., 2014; Terashima & Kephart, 2016; Ajayakumar et al., 2019; Schmutte & Vilhuber, 2020).

2 Background

When data collection is planned, an organisation or individual researcher typically must make decisions regarding what information to collect, and how that information will be recorded. For example, an individual's age may be recorded as date of birth, the exact number of years since birth, or which age band the individual belongs to from a predetermined set of age bands. Similarly, information collected about where an individual lives or works may be recorded as a specific address with coordinate information, or more broadly as general place information, such as a suburb or city location. Decisions made at the data collection stage can affect a respondent's willingness to provide information, and if suboptimal decisions are made, analyses of curated data may be constrained (Boyle & Dorling, 2004).

Following data collection, data may be offered or uploaded to a repository, curated and metadata added, and finally disseminated or released for research purposes. Through these stages from collection to release, decisions may be made on the geographical identifiers attached to records which may later affect the research utility of the data. These decisions may be made by national/local government organisations, academic researchers, or commercial organisations, for example. For any data set, different people or agencies may make data specification decisions at any of the stages along the collection to dissemination processing progression.

Paramount to data dissemination and the integrity of research is that the confidentiality of individuals is preserved. Confidentiality considerations include safe projects, safe users of the data, safe settings for access to data, safe data specifications and safe outputs (Desai et al., 2016; UK Data Service, 2021; Mills et al., 2022). For the researcher, there would be a preference to have ready access to the data on their own hard drive, a secure network or at a safe setting within their own institution. Therefore, there is a need for data sets provided to be safe and non-disclosive of individual's information. However, geographical variables within a dataset can compromise the confidentiality of individuals by increasing the risk of identification. Without geographic variables, specific combinations of personal attributes (age, sex, tenure, ethnicity, health status, etc.) may lead to unique observations in the population or sample, where for those unique observations there is higher risk of identifying the individual. In a released data set, for any one variable there is a risk that too much detail will enable an individual to be identified, and this is exacerbated through the detail of a cross-tabulation of two or more variables. When geographical location is then included, disclosure risk is heightened (Griffiths et al., 2019). This is because an unscrupulous user would know where to look to identify a particular person and the size of the known geographical area may be critical (Greenberg & Voshell, 1990; Mills et al., 2022). A tension therefore exists between making data available that protects the

confidentiality of individuals while containing sufficiently detailed socio-demographic and geographic information to underpin the utility of research.

Typically, the level of detail used in personal attributes and geographical indicators included in a publicly available dataset will be appraised prior to release by official agencies for census microdata (e.g. the UK's Census Samples of Anonymised Records) and large scale surveys (e.g. Health Survey for England); often in consultation with expert users (Dale & Elliot, 2001). The decision-making process on the specification of the data released has a basis in the probability of uniqueness (Skinner & Elliot, 2002), but thresholds for data release vary. Regardless of whether the source of the data is official, administrative or Big Data, in order to access individual level records, the researcher may need to negotiate with the supplier the specification of the data that are released. Geographical indicators may, however, be incorporated which would not be the researcher's choice; the specification of the geographical area units or boundaries may be unclear, unfamiliar or dated, or the reliability of locations and linkages between individual points and geographic boundaries of unknown quality (MacEachren et al., 2005). It is possible that the data supplier has deliberately blurred geographical locations and linkages as a confidentiality preserving measure (Ajayakumar et al., 2019; Scheider et al., 2020) though still with the aim of providing high-quality data (Franklin, 2022). Whether or not the geographical detail is negotiated, it is likely that both supplier and user do not know the degree to which the decisions made affect the utility of research results.

There are several commonly used privacy protection measures which may be implemented to protect the confidentiality of individual's information. These broadly fall into three categories: data suppression, data coarsening (e.g. aggregation), and noise infusion (e.g. random perturbation). These three measures were assessed by Schmutte and Vihuber (2020) in terms of balancing individual level data privacy and data usability. In the context of geographical elements, suppression can occur when the locations of observations are not available to researchers, coarsening when a respondent's specific location is released for a less local/more regional geographic area units, and noise infusion (random perturbation) when the location is blurred in some way by 'jittering' (shifting) the centroids of an area a specific or random distance from the original point location. These methods introduce a degree of spatial uncertainty which in itself is confidentiality preserving (Delmelle et al., 2022) with location blurring helping to ensure privacy (Armstrong et al., 1999; Goodchild, 2018; McKenzie et al., 2022). However, any insertion of error into data may have repercussions for geographers, civic stakeholders, policy makers (Franklin, 2022) and the populations for whom decisions are made based on those data. Thus, there have been substantial developments in the privacy protection measures available for protecting individual's location information including the development of geomasking and obfuscation techniques (e.g. Seidl et al., 2015, 2018), and using obfuscation to protect individual's location while maintaining usability for Location Based Services (Duckham & Kulik, 2005a, b). Recently, differential privacy has begun to be implemented as a privacy protection measure instead of more traditional methods such as noise infusion (jittering). However, as evident in the US 2020 Census, there

is still an ongoing debate surrounding the appropriate usage and applicability of differential privacy to research and social sciences, as outlined by Hawes (2020).

Formally collected data such as census, social survey, or health records which may then be held in a repository for researchers to apply for use or have as open access download, needs a pragmatic assessment of options which both suppliers and researchers can consider when specifying geographical identifiers and geographical scale of analysis. The release of the 'Goldacre Review' (Goldacre & Morley, 2022) highlighted the importance of data protection in the context of public health data held by the National Health Service (NHS) in the UK. They identified that current privacy techniques in use are out-of-date and have a high re-identification risk, placing public health data in high risk. While in the long-term Trusted Research Environments (TREs) have been proposed as the way forward in ensuring data privacy while balancing utility (Goldacre & Morley, 2022; Lehoux & Rivard, 2022), until these are implemented there is still a need to reduce the risk of re-identification in individual level data.

To the best of our knowledge, no paper has specifically focused on investigating a geomasking/obfuscation method for assessing both privacy of individual level information and research utility, that is easy to use, practical and can be regularly used by data holders and/or researchers without assistance from third parties, applications, or extensive mathematical knowledge. In this paper we will specifically assess the effect of an obfuscation approach on associative relationships, not just the spatial patterns of the original dataset. While many papers have focussed on maintaining spatial patterns in the dataset when obfuscating location data, there is little attention on whether relationships are maintained between variables such as outcomes for individuals and area characteristics. This is important to investigate, as linking obfuscated points to areas may result in the observation being linked to the wrong area, so that an associated composite measure (e.g. level of deprivation) is incorrect. This would represent misclassification of exposure and lead to misclassification bias in the reported relationships (Peat, 2002).

Based on the assumption that data would not be made publicly available by agencies with coordinates for respondents' addresses, the underlying question we will investigate in this paper is:

If there is a relationship between outcomes for individuals and area characteristics, how far away from residential locations can geographical identifiers be displaced (accidentally or deliberately) before a different conclusion would be made?

To investigate the implications of the balance between what information has been collected and is available in a public dataset and whether research utility is affected by any compromise in accuracy of geographical identifiers, in this work:

- We will specify the geographies (UK area units) in England and Wales which are relevant to this work. This will include information about address conventions and postal geographies, and the geographies which are used for the dissemination of census data.
- We will introduce the resources we use to explore the linkages of individual level data to area data. Specifically, these are *synthetic microdata* in which observations (c. 22,000) can be located by a variety of point locations and linked to the

area data which comprise a variety of GIS boundary files relevant to census area data.

- In the analytical work, we are working with two geographical *entities* (points for the observations and polygons for area attributes) at a variety of *scales* (points obfuscated in different ways linked with areas coarsened from local to larger areas).

  - In order to learn of the implications of linking people using different location information we will first link the synthetic microdata using the most detailed information which *might* be available to researchers, the unit postcode, to a range of area scales from very local up to larger areas. We carry out a series of simple statistical procedures to determine relationships between outcomes for individuals and level of area deprivation.
  - We then repeat the same procedures but with successively less resolved geographic detail for the synthetic individual locations and repeat the statistical procedures. This emulates the situation if truncated address information had been collected in the first place or the location coarsening choices which data custodians/disseminators might have made before releasing data to researchers.
  - A simulation is undertaken whereby the point location of the residential postcode is moved incrementally away from the original point. This 'jittering' of the points acts as a blurring mechanism and this simulation is to determine how much noise can be applied before the expected statistical relationships no longer hold.

3 Methods

3.1 Explanation of Geographies Being Used

UK address convention is to have a house number (or name) with the street name, the town or city and a postcode. The postcode forms part of postal geography and is a device used for the delivery of mail though has a long history of use in geographically related research (Raper et al., 1992). Postcodes may be for residences or businesses and here we focus on the former. Although a postcode is geolocated as a point in space, in reality it will be a set of addresses (average c. 20 in England and Wales). The 'unit' postcode is an alphanumeric code such as BD23 1UH and there are around 1,300,000 postcodes in England and Wales. Again, for the purposes of organising the delivery of mail, going up the geographical hierarchy from smaller to larger areas, there are c. 8,200 postal sectors (BD231), c. 2,300 postal districts (BD23) and 164 postal areas (BD).

Figure 1 illustrates postal geographies in the vicinity of postal district BD23 (the polygon shaded grey). Other postal districts also have bold lines and the within-district postal sectors are the polygons with thinner lines. Unit postcodes are located as red dots and their densities are a proxy for population distribution across the urban—rural gradient (Norman et al., 2003).

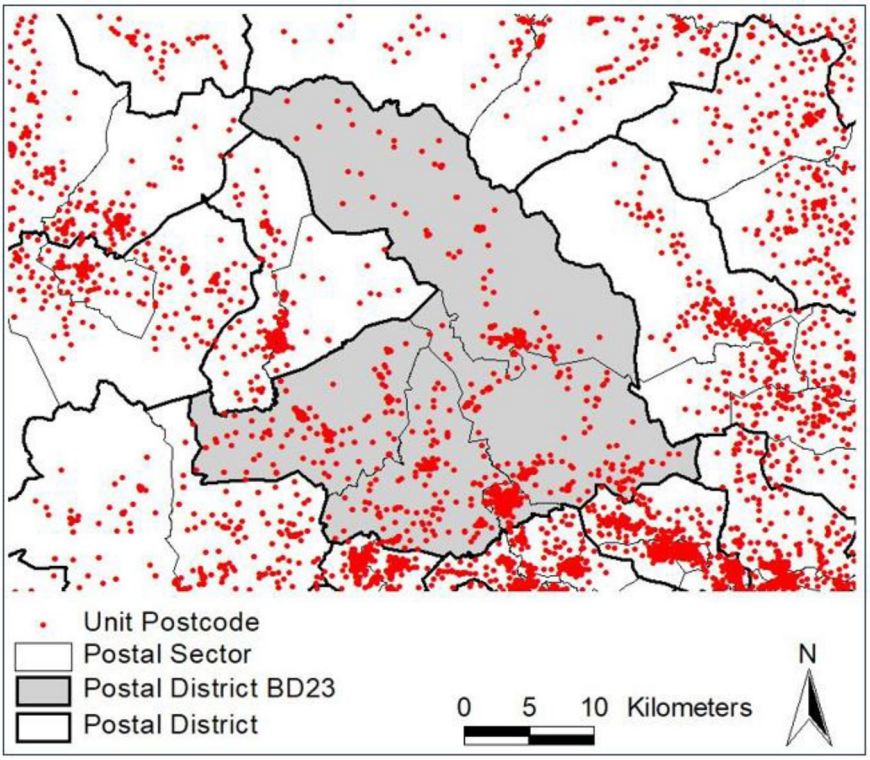


Fig. 1 Postal geographies

Figure 2 illustrates the same extent as in Fig. 1 but here the grey polygon is Craven, a local authority district in North Yorkshire. Adjacent local government areas also have bold lines. The within district geography are the Lower Super Output Areas (LSOAs). The LSOAs are part of the hierarchy of areas from Output Areas and LSOAs to Middle Super Output Areas (MSOAs) which nest within each other and the local government areas (ONS, 2016). From smaller to larger areas, the OAs, LSOAs and MSOAs are designed for the release of census data with a balance being struck between (more > less) geographic detail and (less > more) socio-demographic detail.

As noted above, individual records are often linked to geographical areas so that, for example, variations in an outcome might be investigated in relation to area characteristics. If the unit postcode is available, this can be geocoded. BD23 1UH has the GB grid reference (x) 398946 (y) 452057. If postal sector is available, the geometric centroid of the polygon can be geocoded (x) 398730 (y) 451,842. The centroid of the postal district is (x) 396097 (y) 455741. These points can be associated with the census area geographies.

For illustrative purposes, Fig. 3 shows the location of the unit postcode (red square) and the sector centroid (blue triangle). The straight-line distance between

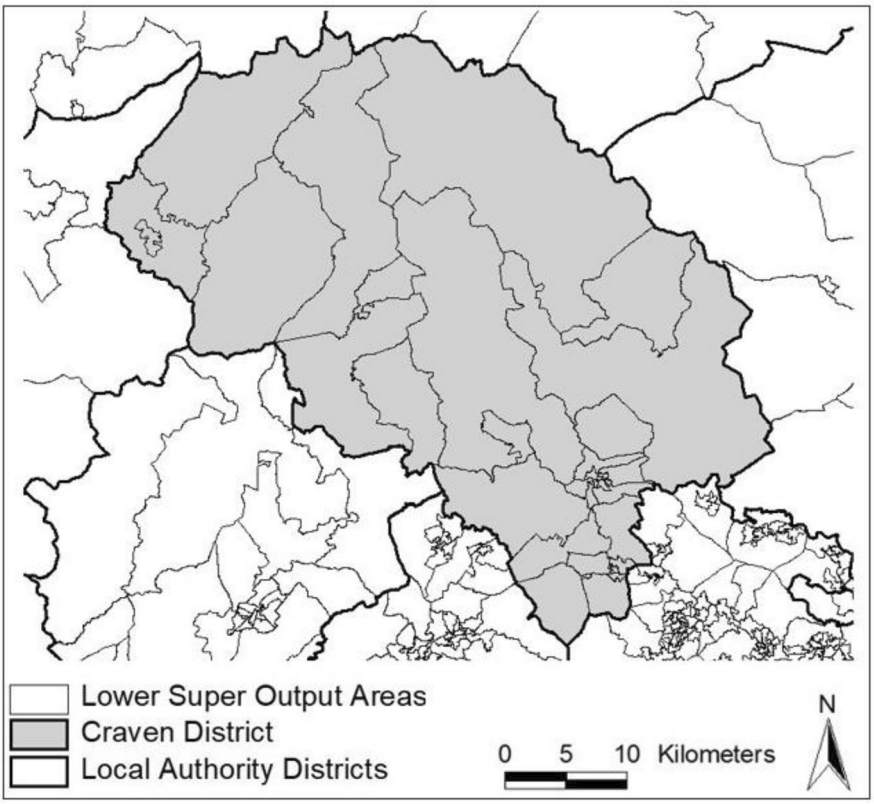


Fig. 2 2011 Census geographies

the unit postcode location and the sector centroid is just over 300 m. The postal district centroid (green circle) is over 4.5 km to the North. Figure 3a illustrates the census Output Area geography. The postcode and postal centroid locations are close together but would be associated with different OAs. This is not the case though for the LSOA and MSOA geographies (Fig. 3b and c) where the link would be to the same area even though the point is in a different place. The three points geocoded in the different ways would all be linked with Craven local government district. Further descriptive statistics on the postal and census geographies are in Appendix 1.

3.2 Synthetic Data

We have devised a synthetic, individual level data set of ~ 22,000 individuals living in England and Wales with variables defined to enable the testing of linkages across various geographical scales.

The synthetic microdata file has a small number of individual level variables similar to those in the ONS microdata teaching file, itself used to create a

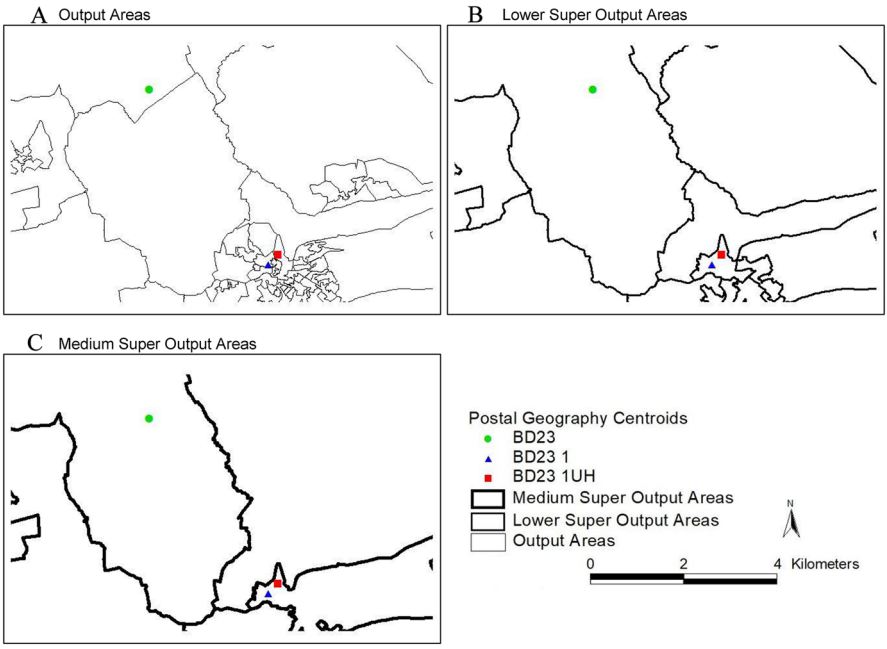


Fig. 3 Linking postal geography derived locations to different census geography scales

synthetic longitudinal dataset (Dennett et al., 2016). Our synthetic data are fictional apart from being based on the probability that a person with a range of attributes might live at a postcode in a particular Output Area. This is the basis of spatial microsimulation (Lomax & Smith, 2017). The individual level variables include whether or not the person owns their home, their age, sex, ethnicity, qualifications, etc. (Appendix 2 explains how the synthetic microdata have been devised and details the variables included).

In the synthetic microdata, individuals are located by residential postcodes; the finest resolution of location information which may be available to researchers. These postcode locations can be linked with the England and Wales census geographies: Output Area (OA), Lower Super Output Area (LSOA), Medium Super Output Area (MSOA) and Local Authority (LA) district (ONS, 2011, 2016). In addition to the detailed unit postcode locations, individuals are also located using the postal sector and postal district centroids. These less resolved point locations can also be linked to the census area geographies (Appendix 1, Fig. 6 has a schematic comparing postal and census geographies).

For an area measure which can be attached to the synthetic microdata, we use population weighted quintiles of the Carstairs deprivation index (Carstairs & Morris, 1989). The input variables (rates of male unemployment, low social class, no car and household overcrowding) can be obtained for each census geography, OA, LSOA, MSOA and LA, with deprivation calculated at each level. The distribution of the synthetic individuals is ~ 20% by deprivation quintile.

3.3 Analysing Variations in Individual Location Area Scale Linkages

Ideally, accurate and precise data are available to carry out reliable research.

In population geography/demography/epidemiology, linking a person's residential location to area data is regularly carried out whether at source by a data supplier or by a researcher. This may be to provide aggregate, area level data, to calculate rates of an 'outcome' using numerators of the outcome and the denominator of the population 'at risk' of that outcome and/or to determine the relationship between the type of place and a particular outcome for individuals. The researcher may want the most detailed geographic information (location of the individual and the most local level zones). The public and the data holder may require that the information is not so detailed geographically. Here, we want to establish whether relationships vary when individual locations, variously geocoded, are linked to areas at different scales.

The outcome of interest we are testing here is whether an individual owns their house or not, in relation to the level of deprivation in an area. We explore whether the relationship varies by both distance from original postcode location and by geographical scale. To achieve this, there are two broad phases of analysis. During each phase, we test variations in rates of homeownership which occur when microdata georeferenced in various ways are linked to different geographical scales. For every individual location/area combination we run simple binary logistic regressions to predict the odds that someone owns their home as the outcome. We include each individual's age, sex and qualifications as explanatory variables in the models however the relationship between area deprivation and each individual's home ownership is our primary focus.

In Phase 1, we start by linking the unit postcodes to the hierarchy of census geographies: OA, LSOA, MSOA and LA. We regard this as a best-case scenario against which other georeferencing specifications can be compared. Next, we use the postal sector centroid to locate individuals and link this to all the census geographies. We then use the postal district centroid for the linkages. What these less geographically detailed postal sector and district centroids emulate are the choices a data supplier might make as confidentiality relevant alternatives to releasing data for unit postcodes. Do these less geographically resolved locations lead to different relationships apparent between homeownership and area deprivation?

In Phase 2, we introduce some 'noise' into the unit postcode locations by exploring a range of distances away from the original grid reference. To achieve this, for each synthetic individual observation, we 'jitter' (adjust) the grid reference coordinates away from the original unit postcode point locations and link these to area deprivation for the different census geographies.

Compared with the truncated postcode centroid locations, this will provide a greater range of 'noise' possibilities for assessing the distance away from original before the outcome/deprivation relationship changes. We systematically jitter the x and y coordinates incrementally, at random plus or minus ($\pm$) a specified distance from the original point. Since both x and y are changed, the new locations are jittered diagonally away from the original. We iterate this process 100 times at each distance: 100, 250, 500, 750, 1,000, 1,250, 1,500, 1,750, 2,000, 2,500, 3,000, 3,500, and 4,000 m.

For all of the above, variations in the odds ratios of homeownership by quintile of deprivation output from the series of simple logistic regressions will reveal whether the choices of geolocating individuals affects the relationships.

4 Results

4.1 Phase 1: (a) Unit Postcode

Each observation in the synthetic microdata is the full unit postcode at its original location and is linked to the census geographies of increasing size (Output Area up to Local Authority) with the level of deprivation calculated at each scale. This represents the best choice scenario for researchers whereby the most geographically resolved location of the observations is available to be linked to a range of geographies.

Figure 4, panel A, illustrates the odds ratio (and confidence intervals) of homeownership (controlling for age-group, sex and qualifications) by level of deprivation. By quintile of deprivation at OA level there is a steep gradient with increasingly lower likelihood of homeownership with increasing deprivation. There is also a gradient for LSOAs and MSOAs but less steep with increasing area size. At LA level the most deprived quintile 5 is shown to have a lower level of homeownership but there is little difference between quintile 1 (the reference category) and quintiles 2, 3 and 4.

4.2 Phase 1: (b) Postal Sector Centroid and (c) Postal District Centroid

Each microdata observation is located as the postal sector centroid and then as the postal district centroid with, for both specifications, linkages made to the range of census geographies and area deprivation. This emulates situations in which decisions had been made by collectors of individual level data (or custodians releasing data) to not make the detailed unit postcode available for research, instead opting to provide a higher level of geography to locate the observations.

Links between postal sector centroids and OAs result in a reduced difference between the reference category quintile 1 (Fig. 4, Panel B) and the other quintiles and a much flatter gradient compared to the links using unit postcodes. Similarly, for LSOAs and MSOAs there are gradients with deprivation but these are less steep than in Fig. 4, Panel A, while the pattern for LAs is very similar to that using unit postcodes and the same conclusions as for sectors and unit postcodes would be drawn.

For linkages using postal district centroids the next hierarchical postal geography, the gradient of odds across the quintiles flattens (Fig. 4, Panel C) with little difference between adjacent quintiles for OAs, LSOAs and MSOAs. At the LA level, the most deprived quintile 5, as with the linkages for unit postcodes and postal sectors, has lower odds of homeownership.

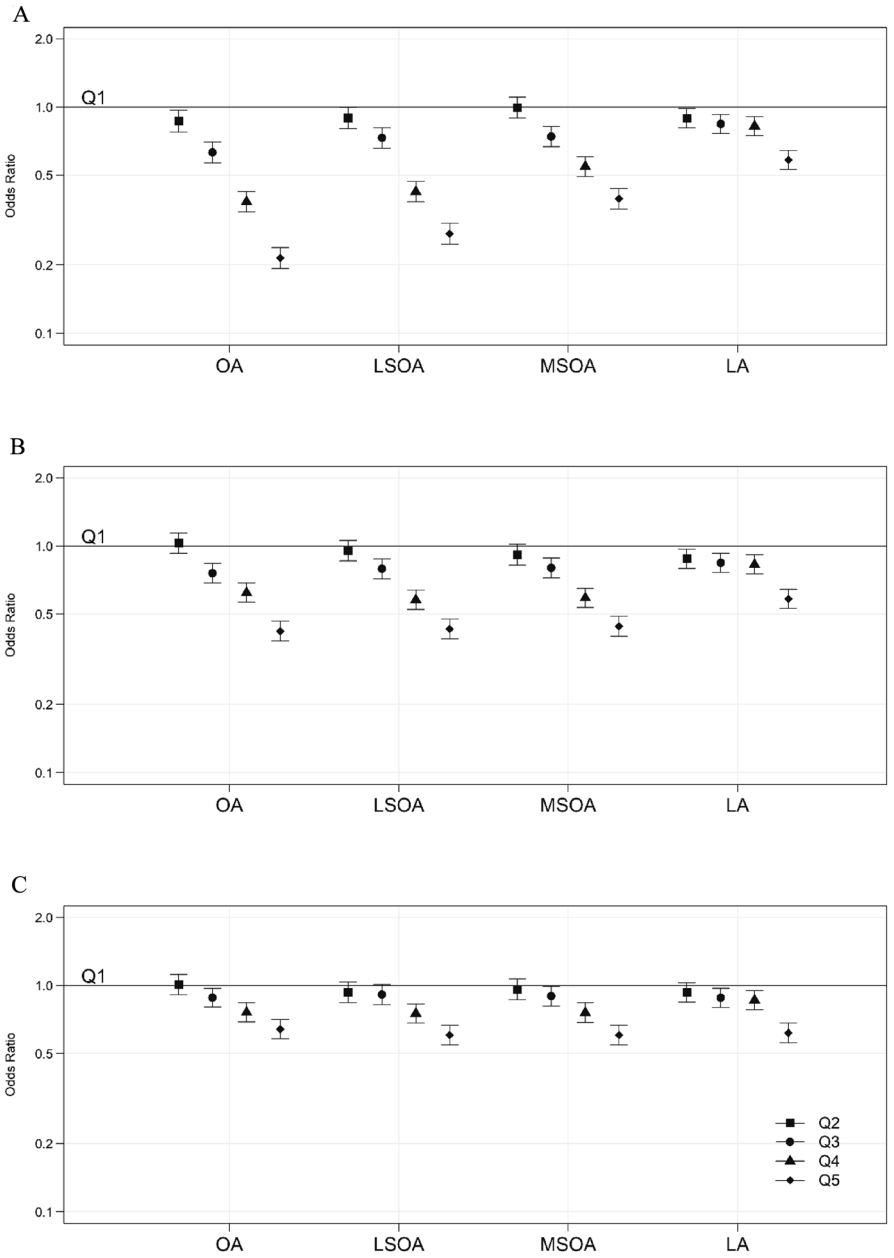


Fig. 4 Modelled odds of homeownership, for links between **A** unit postcodes, **B** postal sectors, **C** postal districts, and census geographies by deprivation quintile

4.3 Phase 2: Jittering the Unit Postcode Location

As specified above, we introduced 'noise' into the unit postcode locations by exploring a range of distances away from the original grid reference. This location jittering represents an alternative approach for data suppliers to release geolocated individual records and thereby assure respondents that their data are being kept confidential since their original postcode location would not be made available.

The jittered unit postcode locations are linked to the OA, LSOA, MSOA and LA geography and the statistical procedures from Phase 1 are repeated, thus exploring the relationships between quintile of deprivation and homeownership for jittered locations. The multiple model outputs are graphed with the coefficient distributions displayed as simplified boxplots (median value, minimum/maximum value error bars and any outliers) by quintile of deprivation by distance from the original point for each geographical level (Fig. 5).

Results at the OA level (Fig. 5, Panel A) retain the same pattern as we observed in Fig. 4, with reduced differences in effect size as distance from the original location increases. The overall differences do not change very much beyond 500–750 m from origin, but the variability of distributions within each quintile tends to increase. Results for both LSOA and MSOA levels reveal very similar patterns, while at the LA level, the most deprived areas have the lowest rates of homeownership at any distance from the original but there is little difference from the less deprived quintile.

In summary, we found that for the OA, LSOA and MSOA geographies, linking unit postcode locations jittered by 500–750 m or less is likely to allow effectively the same conclusions to be drawn as for the original locations. Beyond 750 m however, the patterns are likely to be very similar, although the variations between simulations make any findings less reliable should a 'one-off' ad-hoc linkage be carried out with a jittered point. The effect of jittering points from the origin dilutes associations with the outcome of interest, but remains apparent for OA, LSOA and MSOA scales. For the larger LA geography, we found little difference between relationships with distance away from the original unit postcode point.

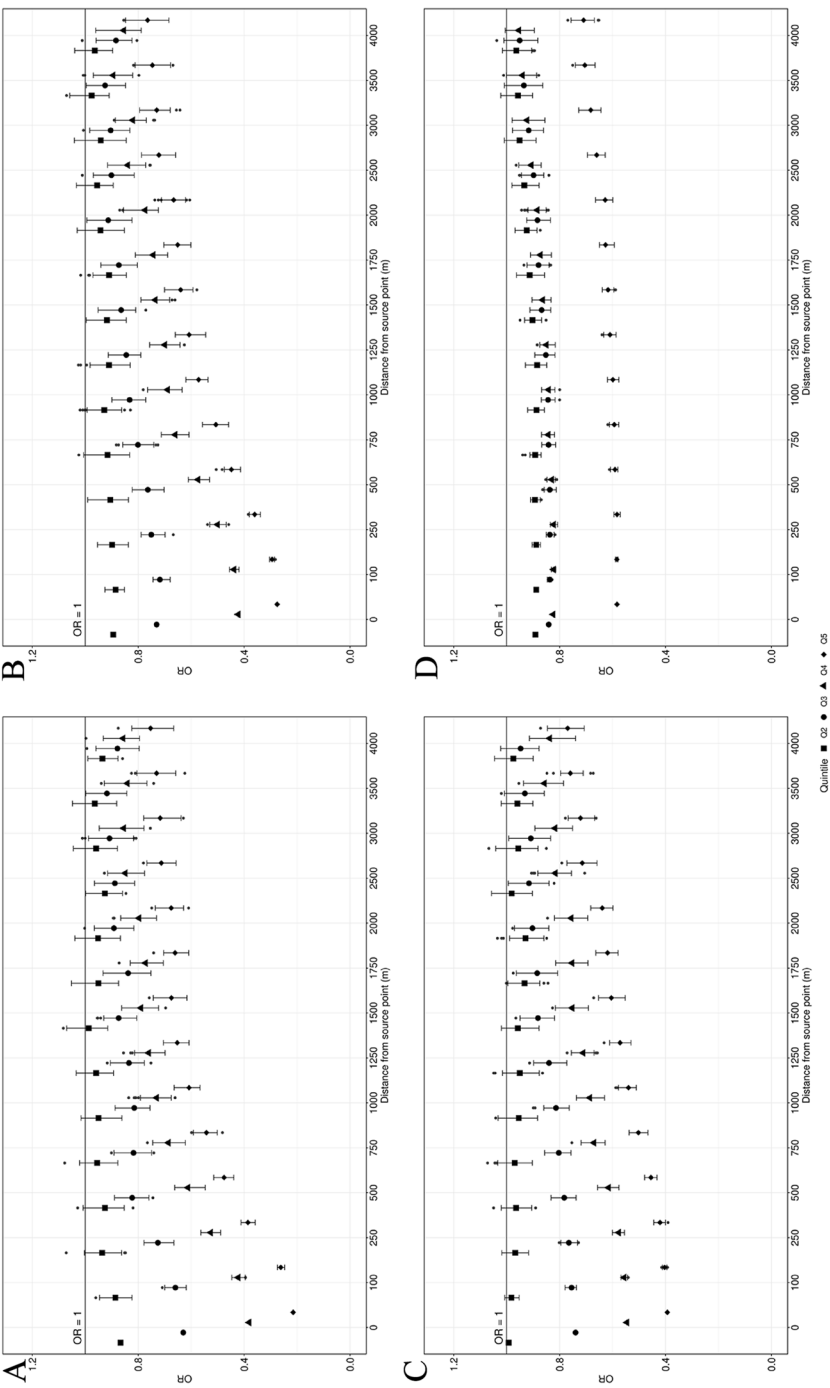


Fig. 5 Modelled odds of homeownership: links between jittered unit postcode coordinates and the census geographies by deprivation quintile for **A** Output Area, **B** Lower Super Output Level; **C** Middle Super Output Level; and **D** Local Authority geographies

5 Discussion

This paper investigated the interplay between two geographical entities (points for the observations & polygons for area attributes) at a variety of scales, to emulate choices a data collector/supplier might make and using an analytical approach researchers might use to explore spatial relationships of sociodemographic phenomena.

Using a synthetic microdata population, in two phases, we:

(i) Linked synthetic microdata at the unit postcode, postal sector, and postal district levels to UK census area geographies and calculated the odds of home ownership by level of deprivation controlling for individual attributes.
(ii) Ran a simulation where the postcode location was jittered incrementally away from the original point, and the odds of home ownership calculated again for each jittering distance and simulation run.

We found that protecting confidentiality by linking data to geographical areas larger than the most local unit available (unit postcodes in the UK), resulted in reduced effect size and reduced differences between area types as the level of administrative geography increased in size. The use of postal sector or district centroids instead of postcodes for locating the individual, another tactic of protecting confidentiality, is not advised for linking to smaller geographies, while linking to larger geographies will give similar results to equivalent postcode links. In blurring the postcode coordinates to protect confidentiality, we found that jittering up to and around 500–750 m away from the postcode point provides very similar results to using the original location.

Our findings for Phase 1 align with prior research relating to the effect of scale on research results, specifically the Modifiable Area Unit Problem (MAUP). Previous work has found that smaller geographical units tend to have greater within area population homogeneity but are more likely to require suppression of results for privacy, whereas larger geographical units tend to protect individual information better but also have more population heterogeneity (Manley et al., 2006; Mills et al., 2022). In this paper, we saw in Phase 1, a) unit postcodes, as the level of geography linked to increases in size, the difference in effect size between deprivation quintiles decreased. Similarly, in Phase 1, b) postal sector and c) postal district, we found evidence of decreasing differences in effect size as the point location used for locating individuals decreased in detail and specificity.

In Phase 2, we used an input noise infusion obfuscation technique, referred to as 'jittering' here, to protect the privacy of individual locations (Schmutte & Vilhuber, 2020). Obfuscation techniques, sometimes referred to as 'geomasking', have been applied to spatial data and individual point locations in previous studies. Zandbergen (2014) reviewed the different geomasking techniques that have been used and emphasised that the effectiveness of a given technique used depends on the reidentification risk of the original locations from the obfuscated dataset. In turn, the reidentification risk is dependent on the amount of displacement used. Typically, Zandbergen (2014) found that the greater the displacement, the lower the risk of reidentification, however this may not always hold in larger, rural areas where the risk is higher due to lower

population density. Furthermore, given our interest in providing guidance to data collectors/suppliers, greater displacement may decrease the utility of the output dataset for research purposes. Seidl et al. (2015) also reviewed existing geomasking techniques and compared their effectiveness to a new method they developed, Voronoi polygon masking (Seidl et al., 2018). They found that while their approach balances privacy at the household-level while maintaining the original spatial distribution of location data (Seidl et al., 2015), there is greater risk of false identification when using the Voronoi method than other methods (random perturbation, donut masking and Military Grid Reference System masking) (Seidl et al., 2018). This supports our chosen obfuscation approach, random perturbation, used in this paper.

Hampton et al. (2010) utilised a 'donut method' of geomasking where, similar to our approach, the location is displaced in a random direction between a minimum and maximum distance. The random perturbation of point locations was found to be more effective in protecting privacy than aggregation to the centroid of a census unit. We do not randomly perturb the location within a circle, however rather in a diagonal direction from the original point, by a specific distance. Therefore, the risk of reidentification may be greater than if a circle of displacement was used.

The effectiveness of an obfuscation approach is typically assessed using *k*-anonymity, which measures the risk that an individual could be correctly identified (e.g. Hampton et al., 2010). Typically, an approach is considered effective based on the extent that spatial patterns in the dataset are maintained. Recently, differential privacy has also emerged as a way to ensure an acceptable level of anonymity is reached, however there is still ongoing debate as to the applicability to social science research and health data (Hawes, 2020; Franklin, 2022). Further research is required to assess the effectiveness of the obfuscation approach used here in protecting privacy of individual records, and the reidentification risk.

Ajayakumar et al. (2019) showed that it is possible to balance individual privacy protection and maintain accuracy of spatial data relationships. They used a random distance offset in their obfuscation approach (translation) and then re-transformed the points by the same distance. Within the translation step they also ensured that obfuscated points were at least a minimum distance from the actual location. While we did not use translation in our obfuscation method, this could be a possible extension to our method in the future.

More recently, McKenzie et al. (2022) presented options for obfuscation of individual's locations via a mobile application. Similar to the jittering methods used in this paper, they include an option for randomly shifting the location of an individual by a distance offset. This is considered as the method preferable for retaining the greatest level of detail while obtaining some security. Other options proposed include using the region of the location represented by a circle of a set distance radius where the centroid of the circle is randomised and the actual location is contained within the circle, or using the official geography that the location is contained within. These options are useful for personal data protection but may not provide the level of specificity required for use in research. Here we tested the research applications by running simulated methodology on the jittered output.

The results of this work provide alleviation of the tension between protecting individual level information while maintaining sufficient detail for research. We have

demonstrated that it is possible to both protect the location of an individual and ensure research utility is maintained by using the obfuscated locations. However, this balance is only achieved up to a maximum jittering distance, and beyond this research utility may be reduced. For those living in rural areas, obfuscating a location by jittering it randomly by 500–750 m may be less effective for protecting confidentiality.

Linking geocoded individual observations to areas has potential for (deliberate or accidental) classification error (Terashima & Kephart, 2016). Misclassification is a common source of statistical bias (Espeland & Hui, 1987; Peat, 2002) and in the context of this paper, misclassification can be said to have occurred when individuals are linked to/associated with a level of deprivation which is different to their residential location. This can occur if an 'individual' is associated with the 'wrong' area and therefore the 'wrong' level of deprivation is associated with their record, which is more likely to occur with increasing distance from the individual's area of residence. The extent of misclassification may differ for urban and rural areas, due to census geographies in urban areas tending to be smaller in areal extent than those in rural areas. Therefore, the distances to the 'wrong' area may be shorter in urban and longer in rural areas. In future work, the effect of urban and rural area classifications on the results of this research will be investigated by replicating the work presented here for OAs up to LAs differentiating between local authorities which are urban or rural. Using our synthetic data, the extent of misclassification of exposure related to area size and similarity of neighbouring areas can be quantified (differences in links to the various levels of area deprivation).

6 Conclusion

The most detailed geolocation in the United Kingdom, unit postcodes, and lowest scale of geography (OAs) lead to the greatest heterogeneity in between area results. There is a smoothing of the relationships up the geographical hierarchy. If postal sector centroids are used to locate individuals, linkages to OAs, since they are inherently smaller, may not be appropriate. However, if the geography of analysis was either LSOAs or MSOAs then results are very similar to the more resolved links but with reduced effect sizes. Postal district centroids do not differentiate between individuals and their locations to enable analyses below LA level.

For the OA, LSOA and MSOA geographies, linking unit postcode locations jittered by 500–750 m is likely to allow effectively the same conclusions to be drawn as for the original locations. At further distances away, whilst the patterns are likely to be similar, the variations between simulations make any findings less reliable should a 'one-off' linkage be carried out with a jittered point. There is a dilution of effect with any distance away from the original location.

Thus, researchers can be re-assured that if choices have been made in a manner similar to the above, the relationships observed in the data are likely to be similar, though somewhat diluted, to the relationships which would be found using the most detailed geographic locations. However, the research approach we have taken is relatively simple and more thought might need to be given in both multilevel and longitudinal frameworks. In a hierarchy of person > neighbourhood > local government,

effect sizes might be diluted for the lower but not the higher geography. In a cross-classified model (e.g. school catchment and residential neighbourhood) both geographies might lead to dilution. For longitudinal models, if individuals are linked to neighbourhoods over the lifecourse (e.g. Murray et al., 2021; Pearce et al., 2018), then with every change in location there might be variations in individual/area relationships due to linkages carried out for different time points. However, if individuals are linked to more than one geography, whether cross-sectionally (e.g. neighbourhood and workplace) or longitudinally (e.g. change in residential location), then this may heighten the risk of identification.

For reassurance to data holders and the general public, in the long-term, secure data platforms such as Trusted Research Environments (TREs), or Safe Havens, will be commonplace for the storage and protection of public data (Goldacre & Morley, 2022), and the need for approaches such as presented here will be lessened. TREs are already in place in countries such as the UK, where they are used for purposes such as health and social care by government organisations such as the NHS (Zhang & Kamel Boulos, 2022). New Zealand's Integrated Data Infrastructure (IDI) is a TRE providing academics, researchers and government policy analysts access to de-identified individual and/or household routine data along with national surveys including the 2013 and 2018 Censuses, General Social Surveys and Health Surveys (Stats NZ, 2022). The Australian Bureau of Statistic's (ABS) DataLab is another example of a TRE, which gives researchers access to de-identified detailed microdata in a secure environment controlled by the ABS (Australian Bureau of Statistics, 2021). Individual level information stored in TREs are deidentified and linked to other key personal information, and can be used by researchers who must apply for access. However, if geographic identifiers are used in data releases, then our approach is still relevant for protecting privacy. Additionally, as highlighted by Affleck et al. (2022), while TREs help to reduce risk of re-identification, there are still caveats that must be addressed to ensure that public data is robustly protected.

The applications of this research are wide-ranging, for informing data collectors and suppliers about geographic choices for confidentiality protection which in turn can provide reassurance to survey/census respondents and inform researchers on whether the data may be fit-for-purpose. While this research may focus on England and Wales, utilising the local official geography and postal geography system, the methods used in this research can be applied to any country and postal/census geography system. In particular, the jittering approach used here could be applied to any point location dataset to better understand the associations between individual demographic structure or compositional variables and the spatial distribution of phenomena at an area-unit scale.

Appendix 1

See Fig. 6.

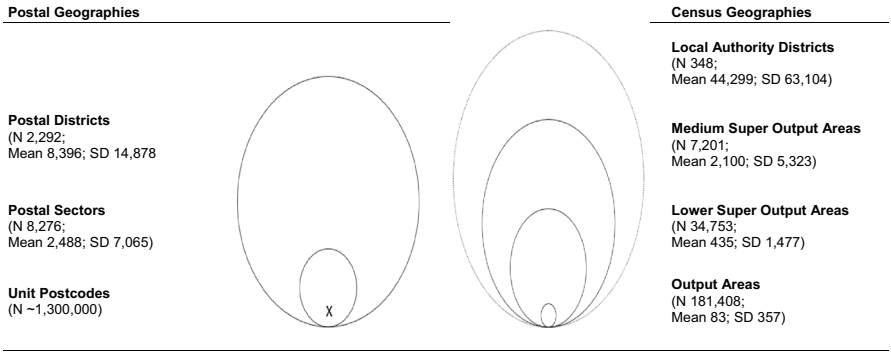
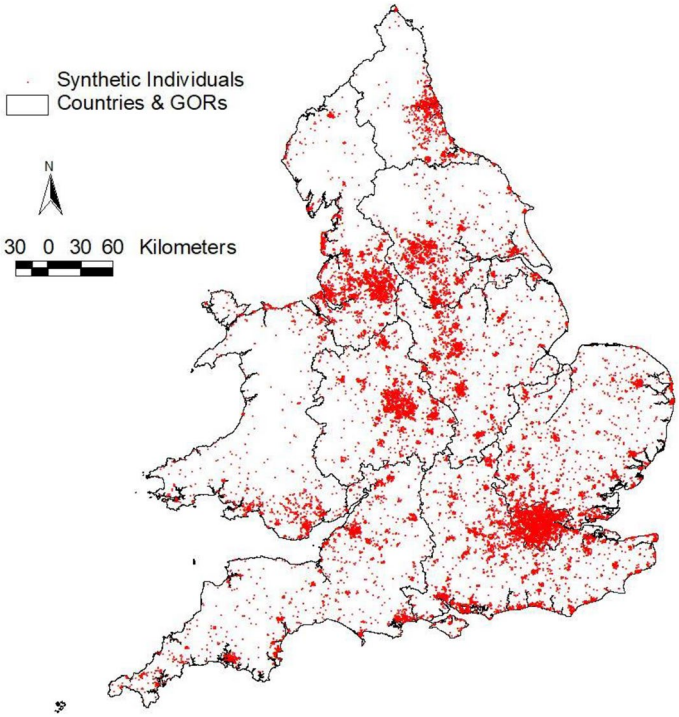


Fig. 6 Hierarchies of postal and census geographies: England and Wales. Note Total number of each (N) and their Mean area in hectares and Standard Deviation (SD). Not to scale and just to give an indication of relative size though the Local Authority Districts line is dotted as should be much larger here

Appendix 2

The geographic distribution of the synthetic microdata is based on the ONS, 2011 unit postcode to Output Area (OA) lookup file (ONS, 2011) for England and Wales. The population weighted quintile of deprivation for OAs was attached to the postcode file and a series of 1% samples was selected at random from within each quintile. Finally, a sample was extracted from that series such that there are 19,941 unique postcode locations and 22,111 records out of around 1,300,000 postcodes in the original ONS file. Duplications of records for the same postcode represent different people in effectively the same location but not necessarily at the same address. Around 1.5% of the England & Wales postcodes are included in the resulting file.



Individual level attributes as categorical microdata have been estimated based on OA area level variables. The aim is for sufficiently plausible attributes rather than necessarily being strictly representative. The proportions of a small set of area level variables have been obtained for each OA. Any binary variables (e.g. whether the synthetic person is male or female) are determined using random numbers (between 0 and 1). If the random number is greater than the lower proportion (i.e. is more likely to be the case) then this is the category selected. In the first row of Table 1, the random number is greater than the overall proportion of females in the OA so the synthetic person at that postcode is deemed to be male. In the second row, the random number is less that the proportion of females so that synthetic person is deemed to be female. For variables with more than two categories, there is a similar process with the largest proportion category determined the same way and then the other categories in order of the OA proportion for where the random number falls between proportions. Thus, for example,

Table 1 Estimation of individual level attribute allocation

Location	Proportion Males in OA	Proportion Females in OA	Random number	Microdata category allocation
XXXX1	0.61	0.39	0.54	Male
XXXX2	0.62	0.37	0.10	Female

Table 2 Example descriptives for the synthetic microdata and for England and Wales, 2011

Variable category	Synthetic data		Census data
	Frequency	Percentage	Percentage
16–24	2,613	11.8	16.3
25–34	3,729	16.9	16.8
35–44	3,626	16.4	17.3
45–54	4,047	18.3	17.2
55–64	3,148	14.2	14.6
65 +	4,948	22.4	17.8
Male	10,599	47.9	49.2
Female	11,512	52.1	50.8
White British Ethnicity	18,885	85.4	80.5
Non-White Ethnicity	3,226	14.6	4.4
Non-home owner	9,258	41.9	36.1
Home owner	12,853	58.1	63.9

age structure tends to be older in less deprived areas and more people have HE qualifications compared with more deprived areas.

Table 3 Variable list for the synthetic microdata

Variable	Code	Category label	Frequency	Percent
Outcome	0	Non-home owner	9,258	41.9
	1	Home owner	12,853	58.1
age_band	1	16–24	2,613	11.8
	2	25–34	3,729	16.9
	3	35–44	3,626	16.4
	4	45–54	4,047	18.3
	5	55–64	3,148	14.2
	6	65 +	4,948	22.4
Sex	1	Male	10,599	47.9
	2	Female	11,512	52.1
ethnic_gp	1	White British Ethnicity	18,885	85.4
	2	Non-White Ethnicity	3,226	14.6
qual_he	0	No HE qualification	14,822	67.0
	1	HE qualified	7,289	33.0
job_status	1	Professional occupations	6,797	30.7
	2	Intermediate occupations	2,832	12.8
	3	Lower occupations	12,482	56.5
submig	1	Non-migrant	20,280	91.7
	2	Migrant	1,831	8.3
uk_born	1	UK born	20,453	92.5
	2	Overseas born	1,658	7.5

Table 2 has example descriptives for age, sex, ethnic group and home ownership for the synthetic microdata estimations (for ages 16 and over) and for the percentages for England and Wales from the original 2011 Census data (for all ages). The frequency distributions in the synthetic data provide plausible percentages compared with the national level percentages even though there is not a close match. Further variables (Table 3) are in the synthetic data for whether or not the person is a subnational migrant, born in the UK or not, together with their qualifications and job status.

Table 3 has a full variable list with the category codes and labels. Each individual observation has a postcode and the associated (x) Easting and (y) Northing. The microdata are for researching variations in relationships which occur when the synthetic individuals are linked to areas using different location identifiers.

Acknowledgements The authors thank the reviewers and editor for their very useful comments. The authors are grateful to Richard Feltbower, University of Leeds, for asking some of the questions which motivated this work and for comments on an early draft.

Funding The work was unfunded.

Data Availability All relevant resources for this work will be made freely available on publication.

Declarations

Conflict of interest The authors declare they have no competing interests.

References

Affleck, P., Westaway, J., Smith, M., & Schrecker, G. (2022). Trusted research environments are definitely about trust. *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2022-108678

Ajayakumar, J., Curtis, A. J., & Curtis, J. (2019). Addressing the data guardian and geospatial scientist collaborator dilemma: How to share health records for spatial analysis while maintaining patient confidentiality. *International Journal of Health Geographics, 18*(1), 1–12.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine, 18*(5), 497–525.

Australian Bureau of Statistics (2021). DataLab. https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab

Boyle, P., & Dorling, D. (2004). Guest editorial: The 2001 UK census: Remarkable resource or bygone legacy of the 'pencil and paper era'? *Area, 36*(2), 101–110.

Carstairs, V., & Morris, R. (1989). Deprivation: Explaining differences in mortality between Scotland, England and Wales. *British Medical Journal, 299*, 886–889.

Dale, A., & Elliot, M. (2001). Proposals for 2001 samples of anonymized records: An assessment of disclosure risk. *Journal of the Royal Statistical Society: Series A (statistics in Society), 164*(3), 427–447.

Delmelle, E., Desjardins, M. R., Jung, P., Owusu, C., Lan, Y., Hohl, A., & Dony, C. (2022). Uncertainty in geospatial health: Challenges and opportunities ahead. *Annals of Epidemiology, 65*, 15–30. https://doi.org/10.1016/j.annepidem.2021.10.002

Dennett, A., Norman, P., Shelton, N., & Stuchbury, R. (2016). A synthetic longitudinal study dataset for England and Wales. *Data in Brief, 9*, 85–89. https://doi.org/10.1016/j.dib.2016.08.036

Desai, T., Ritchie, F., & Welpton, R. (2016). Five safes: Designing data access for research. University of the West of England. https://uwe-repository.worktribe.com/output/914745

Duckham, M., & Kulik, L. (2005a). A formal model of obfuscation and negotiation for location privacy. In *International conference on pervasive computing* (pp. 152–170). Springer.

Duckham, M., & Kulik, L. (2005b). Simulation of obfuscation and negotiation for location privacy. In *International conference on spatial information theory* (pp. 31–48). Springer.

Entwisle, B. (2007). Putting people into place. *Demography, 44*, 687–703.

Espeland, M. A., & Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics, 43*(4), 1001–1012.

Exeter, D. J., Rodgers, S., & Sabel, C. E. (2014). "Whose data is it anyway?" The implications of putting small area-level health and social data online. *Health Policy, 114*(1), 88–96. https://doi.org/10.1016/j.healthpol.2013.07.012

Flowerdew, R., Manley, D. J., & Sabel, C. E. (2008). Neighbourhood effects on health: Does it matter where you draw the boundaries? *Social Science & Medicine, 66*(6), 1241–1255.

Franklin, R. (2022). Quantitative methods I: Reckoning with uncertainty. *Progress in Human Geography, 46*(2), 689–697.

Galster, G. (2001). On the nature of neighbourhood. *Urban Studies, 38*(12), 2111–2124.

Goldacre, B., & Morley, J. (2022). Better, broader, safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care. Department of Health and Social Care.

Goodchild, M. F. (2018). A giscience perspective on the uncertainty of context. *Annals of the American Association of Geographers*. https://doi.org/10.1080/24694452.2017.1416281

Greenberg, B., & Voshell, L. (1990). *Relating risk of disclosure for microdata and geographic area size.* US Bureau of the Census Selected Papers: 1990 Meeting of the American Statistical Association, pp. 450–490.

Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., & Woods, C. (2019). *Handbook on statistical disclosure control for outputs*. Online accessed 13 December 2021. https://ukdataservice.ac.uk//app/uploads/thf_datareport_aw_web.pdf

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L. & Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069.

Hawes, M. B. (2020). Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review, 2*(2).

Higgs, G., Langford, M., & Norman, P. (2015). Accessibility to sport facilities in Wales: A GIS-based analysis of socio-economic variations in provision. *Geoforum, 62*, 105–120.

Jivraj, S., & Alao, C. (2023). Are ethnic employment penalties mitigated in deprived neighbourhoods and in ethnically dense neighbourhoods? *Population, Space and Place*. https://doi.org/10.1002/psp.2646

Lehoux, P., & Rivard, L. (2022). Major public works ahead for a healthy data-centric NHS. *BMJ, 377*, o1018.

Lomax, N., & Smith, A. (2017). Microsimulation for demography. *Australian*. *Population Studies, 1*(1), 73–85. https://doi.org/10.37970/aps.v1i1.14

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science, 32*(3), 139–160. https://doi.org/10.1559/1523040054738936

Macintyre, S., Ellaway, A., & Cummins, S. (2002). Place effects on health: How can we conceptualise, operationalise and measure them? *Social Science and Medicine, 55*(1), 125–139. https://doi.org/10.1016/S0277-9536(01)00214-3

Macintyre, S., Macdonald, L., & Ellaway, A. (2008). Do poorer people have poorer access to local resources and facilities? The distribution of local resources by area deprivation in Glasgow, Scotland. *Social Science & Medicine, 67*(6), 900–914.

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems*, *30*(2), 143–160.

McKenzie, G., Romm, D., Zhang, H., & Brunila, M. (2022). PrivyTo: A privacy-preserving location-sharing platform. *Transactions in GIS*. https://doi.org/10.1111/tgis.12924

Mills, O., Shackleton, N., Colbert, J., Zhao, J., Norman, P., & Exeter, D. (2022). Inter-relationships between geographical scale, socio-economic data suppression and population homogeneity. *Applied Spatial Analysis & Policy, 15*, 1075–1091. https://doi.org/10.1007/s12061-021-09430-2

Murray, E. T., Nicholas, O., Norman, P., & Jivraj, S. (2021). Life course neighborhood deprivation effects on body mass index: Quantifying the importance of selective migration. *International Journal of Environmental Research and Public Health*, *18*(16), 8339.

Norman, P., Rees, P., & Boyle, P. (2003). Achieving data compatibility over space and time: Creating consistent geographical zones. *International Journal of Population Geography*, *9*(5), 365–386.

ONS (2011). Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (December 2011) Lookup in England and Wales. https://geoportal.statistics.gov.uk/datasets/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-december-2011-lookup-in-england-and-wales/about

ONS (2016). Census geography: An overview of the various geographies used in the production of statistics collected via the UK census. https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

Openshaw, S. (1981). The modifiable areal unit problem. In N. Wrigley & R. J. Bennett (Eds). *Quantitative geography: A British view* (pp. 60–69). Routledge & Kegan Paul.

Pearce, J., Cherrie, M., Shortt, N., Deary, I., & Ward Thompson, C. (2018). Life course of place: A longitudinal study of mental health and place. *Transactions of the Institute of British Geographers*, *43*(4), 555–572.

Peat, J. (2002). *Health science research: A handbook of quantitative methods*. Sage.

Petrović, A., van Ham, M., & Manley, D. (2022). Where do neighborhood effects end? Moving to multiscale spatial contextual effects. *Annals of the American Association of Geographers, 112*(2), 581–601.

Raper, J. F., Rind, D. W., & Shepherd, J. W. (1992). *Postcodes: The new geography*. Longman Scientific and Technical.

Scheider, S., Wang, J., Mol, M., Schmitz, O., & Karssenberg, D. (2020). Obfuscating spatial point tracks with simulated crowding. *International Journal of Geographical Information Science, 34*(7), 1398–1427. https://doi.org/10.1080/13658816.2020.1712402

Schmutte, I. M., & Vilhuber, L. (2020). Balancing privacy and data usability: An overview of disclosure avoidance methods. In: Cole, Dhaliwal, Sautmann, and Vilhuber (Eds), *Handbook on using administrative data for research and evidence-based policy.* Online accessed 7 December 2021. https://admindatahandbook.mit.edu/book/v1.0-rc6/discavoid.html.

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*, 253–263.

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and false identification risk in geomasking techniques. *Geographical Analysis*, *50*(3), 280–297.

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 855–867.

Stats, N. Z. (2022). Integrated data infrastructure. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Terashima, M., & Kephart, G. (2016). Misclassification errors from postal code-based geocoding to assign census geography in Nova Scotia Canada. *Canadian Journal Public Health, 107*(4–5), e424–e430. https://doi.org/10.17269/CJPH.107.5459

UK Data Service (2021). What is the Five Safes framework? Online accessed 13 December 2021 https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

Zandbergen, P. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*. https://doi.org/10.1155/2014/567049

Zhang, P., & Kamel Boulos, M. N. (2022). Privacy-by-design environments for large-scale health research and federated learning from data. *International Journal of Environmental Research and Public Health, 19*(19), 11876. https://doi.org/10.3390/ijerph191911876

Authors and Affiliations

Paul Norman[1] · Jessie Colbert[2] · Daniel J. Exeter[2]

✉ Paul Norman
p.d.norman@leeds.ac.uk

Jessie Colbert
jessie.colbert@auckland.ac.nz

Daniel J. Exeter
d.exeter@auckland.ac.nz

[1] School of Geography, University of Leeds, Leeds LS2 9JT, UK

[2] School of Population Health, The University of Auckland, Auckland, New Zealand