Monograph:
Norman, PD (2016) Georeferencing socio-demographic information: postcodes, addresses, land parcels and geographical information systems. Report.

10.13140/RG.2.1.1843.9282

Georeferencing socio-demographic information: postcodes, addresses, land parcels and geographical information systems

Paul Norman, School of Geography, University of Leeds, p.d.norman@leeds.ac.uk

The collection and dissemination of population related data requires that all database records are georeferenced. Individual level microdata with records on people or buildings are usually georeferenced as a point in space using a National Grid Reference (NGR) as an Easting (x) and Northing (y). Countries with full cadastral mapping such as Finland (Tuomaala and Uimonen, 1998) and Israel (Benenson and Omer, 2000) have address coordinates for each property. Population level aggregate information for geographical areas is usually linked to a discrete zone which may be referred to by name and/or by an alphanumeric code. The boundaries of a geographic zone are located in space as a polygon. A database which contains georeferenced records can be regarded as a 'geographical information system'. Computer applications which combine a georeferenced database with the ability to graphically display the information as a map is also a geographical information system and referred to here as GIS software. It is the broad view of a GI system which is most relevant here but some operations within GIS software will be noted (but not specific software).

- Note that Raper et al. (1992: 9-28) very usefully discuss geographical entities in 'Postcodes and places' differentiating between various conceptions and definitions of places. Martin (1996: 1-4) discusses definitions of 'geographical information systems' in relation to socioeconomic data.

First we consider household residential populations in Great Britain before noting some differences in Northern Ireland and for institutional populations. We then outline the situation for mobility, both migration and commuting and in relation to the latter, place of work. Ultimately, all these circumstances are geographically inter-related so should be considered as part of a whole statistical system.

Georeferencing microdata

When individuals respond to a census or survey or when administrative data on individuals are compiled, an address may be recorded. This may include a reference to a particular building (e.g. the person's dwelling) and street name but most likely will include a postcode. The UK has had a full coverage of postcodes since 1974 (Raper et al., 1992) and the postcode has become a vital piece of geographically-related information for the collection and processing of population related data.

Postcode Address File (PAF) directories are compiled by Royal Mail which include information about the location (e.g. number of delivery points) and a point grid reference. Three aspects should be noted. First, the postcode system has been devised for the sorting and delivery of mail and its alphanumeric codes have specific meanings in this context. Second, whilst a postcode is allocated a point grid

reference, in reality it is a list of house numbers (so can be perceived as a line along the relevant street or a polygon drawn around all the properties). Finally, postcodes can change over time in their existence and in their location. The point location is usually referred to as the postcode centroid which is the address count weighted centre of the set of properties which comprise the postcode. Individual records from a census, survey or administrative dataset can be assigned the grid reference associated with the postcode. Both pieces of information can be used in a variety of ways (see below).

The postcode is not the most detailed point information that exists or has potential for use. The 1971 Census was the first to record NGRs for every address enumerated in GB (Clarke et al., 1980) and the product which has developed since is OS Address Point (Harris et al., 2002; OS, 2010). Ideally, the grid reference of the dwelling in which the person lives would be attached to their record in a microdatabase. To date though, the use of postcodes predominates since most users do not have access to building level resolution.

A georeferenced microdatabase is fully flexible within the limits of the sociodemographic detail recorded and the precision of the geocoding.

Georeferencing area data

Administrative processes (relating to national and local government and to electoral systems) have driven the collection of population-related data and thereby affect decisions about the geographies for which data are disseminated. These geographies dominate census geographies but for the dissemination of the decennial census data at sub-local government level, statistical geographies are defined whereby units are defined which are large enough (in terms of population / household count) to help preserve the confidentiality of people's personal information. Business marketing analysts often prefer postal geographies. Most of the geographies noted here are hierarchical with smaller units nesting into those which are larger. Alphanumeric coding systems used to identify areas often reflect the hierarchies (see Martin, 1996: 74 & 77).

In the digital GIS (software) era (post-1971 Census), the boundary of a geographic unit is defined as a polygon. A polygon is digitised as a series of nodes with lines in between and which closes off to contain the space within. Digital boundaries are available to users with an alphanumeric code attached and the name of the area, where relevant.

Area data is an aggregate of microdata. The individual records are summed to the area in which those people live. Decisions are made by data suppliers (often in consultation with users) on the cross-classifications of sociodemographic attributes which are available and whether there are categorisations into attribute groupings (e.g. single year of age information into five year age-groups).

Alphanumeric codes attached to tables of area data can be used to join the information to the attribute database of GIS software to enable choropleth / thematic mapping of variable distributions.

Linking microdata and areas

Following the 1991 Census a step change away from clerically assigned linkages between microdata and geographical areas occurred with the release of the Postcode-Enumeration District Directory for England and Wales and an equivalent in Scotland. Subsequent postcode directories (which are enhancements of Royal Mail products) include the All Field Postcode Directory (Simpson, 2002), the 2001 Postcode Headcount file (Rees et al., 2005), the National Statistics Postcode Directory (NSPD) and the ONS Postcode Directory (ONSPD).

These directories have gradually improved over time in terms of the build up of the historic record of postcode existence (including dates of introduction and termination), the resolution of the NGR which thereby increases the spatially precision and the number and variety of the area geographies with which each postcode is linked. Linkages between postcode and area are achieved using GIS software and the point in polygon function.

Using postcode directories in practice

Microdata records with the postcode of the person's residential address can therefore be linked to geographical areas by matching the individual records to the postcode directory. This represents a major resource to users who have a microdatabase whether national or local government officers, health professionals, academic researchers or business analysts who wish to geocode the records to points within the following types of applications:

- The snapshot collection of population data (such as a census) has individual records geocoded for subsequent aggregation into geographical zones for data release;
- The ongoing recording of administrative records (such as Vital Statistics on births and deaths or Benefits data) for later aggregation into areas for (say, annual or quarterly) data release;
- The addition of area data (from census, for example) to individual records on a large scale survey (such as the Health Survey for England) to thereby add value for research purposes;
- The linkage of a customer database to areas to identify catchment areas;
- The analysis of point distributions to identify clusters of disease or crime, for example.

A further application is to use the postcode distribution and linkages to areas to be able to convert data between different geographies (Simpson 2002; Norman et al., 2003). There are various reasons why this might be necessary:

- The geography of data availability may not be the geography relevant to a particular application;

- A time-series of data are required to that trends can be analysed but boundary changes have occurred;
- To develop a custom geography which best represents the phenomenon of interest or which has sufficiently large zones to protect confidentiality.

To convert sociodemographic data between different geographical systems Simpson (2002) and Norman et al. (2003) use postcode locations obtained from the national directories to link the 'source' polygons in which the data pre-exist, to the 'target' polygons, the geographical zone system for which the data are required. Within a 'geographic conversion table', address count-weighted postcode point distributions are used to calculate intersection weights between overlapping source and target polygons. This conversion method is well-established and produces reliable estimates for the target geographies. A similar approach has been adopted by the Office for National Statistics (ONS) to provide mid-year estimates for non-standard areas (Bates, 2008) but with no citation of the techniques which preceded it. The approach established by Simpson (2002) and Norman et al. (2003) has similarities to both areal interpolation (Gregory, 2002; Schroeder, 2007), as aggregate data are spread out to points, and also to dasymetric mapping (Mennis & Hultgren, 2006) since the presence or absence of a postcode point represents whether or not population exist at a location.

Differences in Northern Ireland

Whilst the NSPDs and ONSPDs have a UK coverage, users should be aware that the grid references attached to postcodes in Northern Ireland (NI) are for the Irish National Grid. This grid system has an origin which is different to the GB National Grid so the grid references are not compatible between GB and NI. The other principles described above will apply.

In addition to administrative and electoral geographies, population analysis in Northern Ireland has used a different geographical entity whereby data are available linked to grid squares which in 2001 were 1km by 1km cells. GIS software which incorporates points and polygons as described above is known as 'vector' GIS. Software which incorporates gridded cells is known as 'raster' GIS. The majority of population-related GIS analysis is carried out using vector GIS but raster GIS allows for the incorporation of environmental data (which has traditionally been cell base) and readily enables time-series analysis for the same geography. This data type has been developed and exploited in NI (Power & Shuttleworth, 1997; Shuttleworth & Lloyd, 2007) with Martin (1989; 1996) ably demonstrating how population data released for vector geographies can be converted into raster data.

Georeferencing institutional populations

The location of institutions such as care homes, student halls of residence, prisons, armed forces bases, boarding schools, etc, present problems for geocoding with inconsistent methods used over time and space. Currently, the decennial census is the only survey which covers both residential and

institutional populations. Variations in the population sizes of institution means that they can pass the minimum population threshold of the smallest geographic areas of data dissemination so have been defined as a non-geographic Enumeration District, for example. A grid reference has been provided in census lookup tables to enable the institution to be mapped (but as a point not as a polygon).

On their postcode directories, the Royal Mail inconsistently defines institutions as 'small' or 'large' users. The expectation is that large users are businesses but an institution can be regarded as both a residence and a business. The same is true of Higher Education institutions. These situations are further complicated because there can be non-geographic postcodes where a user has a PO Box number. Where this occurs, the grid reference assigned to the record may be of the nearest Royal Mail sorting office rather than the establishment itself. Mismatches between the postcode and the institution can also occur when the postcode is for a delivery point which may be for an office located at some distance from the accommodation.

In terms of data collection, the residents of an institution represent microdata, but the variation in scale of the institution itself and style of organisation make it hard to define whether the 'dwelling' part is a point within the polygon of the land the institution occupies. How these entities relate to administrative and census geographies is also hard to define.

Georeferencing mobility

Residential migration. On census forms, respondents are asked to provide their address one year ago if different to the address on the form. In terms of subnational migration, the postcodes of previous and current address allow one year transition data to be computed as interaction data and disseminated as origin-destination matrixes in the Special Migration Statistics. The distance of move can also be computed as a straight line distance between the eastings and northings of the postcode points. This information is available in the Sample of Anonymised Records (Norman & Boyle, 2010).

Outside census years, estimates of migration between areas have been underpinned by health service data whereby patients who have changed address inform their existing or a new GP. The geographical units for which interaction data have been available were based on health service geographies up to the late 1990s (Boden et al., 1992; Champion et al., 1998) and local authority districts thereafter. Underpinning the georeferencing process is the postcode of patients, whether or not they have changed from one time-point to the next and aggregations of postcodes at origins and destinations to produce interaction matrixes. Since the geographical unit of data recording is the postcode, the potential is for much finer grained interactions to be published for total moves even if age-sex data provision was not feasible.

We noted above that data can be converted between geographies using postcodes to link geographies and to apportion data between overlapping polygon units. Converting interaction data to an alternative geography using this approach can be inappropriate because when a boundary changes this might redefine whether or not a migration event has occurred (Norman et al., 2003). Boyle and Feng (2002) develop a method to reallocate flows between areas when boundary systems change using polygon centroids to associate overlapping geographies. Norman & Riva (2011) demonstrate how population weighted centroids can be calculated to thereby estimate population locations within small area polygons and link geographies over time.

Work place and commuting. On census forms, respondents are also asked to provide the postcode of their place of work. In combination with their home postcode, this underpins the provision of matrixes of origin-destination flows in the special commuting statistics and the distance of commute in the Sample of Anonymised Records. The census data on this topic is a unique source of geographic information.

Further developments using populations georeferenced in different ways would be to estimate non-residential, 'day time' populations (as opposed to residential 'night time' populations). The new 2011 question asking whether people stay at an address in addition to their enumerated location (e.g. armed forces base, a work address, student, another parent or guardian, holiday home) and this information will also underpin the possibility of more geographically versatile estimates being produced.

Recent developments and the near future

The unit postcode is key to georeferencing microdata whether for individual level analyses or for linking to geographical zones for the provision of aggregate data. Postcodes are geolocated as a point in space and zones as polygons. In reality, as geographical entities postcodes are not points. Other anomalies include institutions which may be georeferenced as points or polygons and may have their postcodes classified as residential or as business addresses.

The design of the small area geography of the 2001 Census saw more detailed address level geocoding used and postcodes being synthetically digitised as polygons (Martin, 2002). The resulting Output Areas (OAs) were defined as a statistical geography whereby the zones were small enough to allow detailed geographical research but large enough (in population and household numbers) for the safe release of cross-tabulated sociodemographic data; grouped where necessary. Soon after the 2001 Census further statistical geographies were defined as aggregates of OAs (Super Output Areas in England, Wales and Northern Ireland; Datazones in Scotland) and these have become the main geographical zones for which post-2001 administrative data are released.

The collection of the 2001 Census was hampered by the lack of a definitive address register. A high quality, comprehensive list of addresses was subsequently considered fundamental for the 2011 Census (Calder, 2009). The development of an address register first involved the negotiation of data sharing agreements with national suppliers of address lists to set out how data could be shared. The sources for the register are the Local Government Information House's National Land and Property Gazetteer (NLPG), Royal Mail's Postcode Address File (PAF) and Ordnance Survey's Address Layer 2 (AL2) products. Once these files were obtained by ONS, address matching occurred with anomaly addresses mis-matches being verified with data suppliers, local authorities and via ONS field checks. A different approach was taken regarding addresses for 'communal establishments' as these have different characteristics from residential addresses (Calder, 2009).

The data sharing agreements represents a step change in organisational attitudes and a large shift in the resolution of microdata georeferencing potential from postcode to building. However, Calder (2009: 36) states, "It is important to realise that the register currently being developed is solely for use in the 2011 Census. The address register being developed could have a wider role as a rolling address or population register, but this is not the focus of the current work. There are currently no plans to make the register publicly available and there are significant commercial, licensing and confidentiality obstacles that would have to be cleared before this would be possible."

The experience of building an address register (both the negotiations and agreements between organisations and the technical achievements) has served to underpin the creation of GeoPlace and a National Address Gazetteer (GeoPlace 2011). Jointly owned by the Local Government Group and Ordnance Survey, GeoPlace brings together from 1st April 2001, local government address and street gazetteers; the NLPG and National Street Gazetteer (NSG), with Ordnance Survey's addressing products to create a National Address Gazetteer for England and Wales. Primary keys in the database are a unique property reference number (UPRN) and a unique street reference number (USRN). As a single definitive source of accurate publicly-owned spatial address data, the national address gazetteer has potential to deliver significant cost savings across the public sector by eliminating the need for users to undertake data matching of different spatial address datasets. Local government in Scotland and Scottish Government are supportive of joining the National Address Gazetteer and are exploring options for achieving this goal. However, currently there are no plans to include Northern Ireland's Pointer (2011) data.

The taking of the 2011 Census was the motivation for the creation a definitive address register. The taking of censuses in general has been the opportunity to count and define the extent of small area geographies for the safe release of population data. Without a census other deadlines would need to be set to motivate appraisals of microlevel geographical entities. Current small area thresholds relate to

census counts of persons and/or households. A shift to dwelling count thresholds would merit being researched given that local government has records in relation to the collection of council tax and access to housing completion data from the planning system.

Ideally, if the UK took a population register this would include, along with the person's personal and household information, the point location of their main (and secondary) residence(s) and their place(s) of work. In addition their residential history would also be included with the point location of previous addresses (along with the date of house move). Much of this information exists in patient records so validation of the NHS patient register could provide a georeferenced microdatabase. Whilst an address register has the potential for address level geocoding, there is some merit in keeping this detail suppressed. Releasing postcode level information in itself may risk breaching confidentiality but a degree of geographical fuzziness exists. Whilst administrative data (such as patient records) might be used in the construction of a quasi population register, no source collects data on commuting. GPS and mobile phones could allow population mobility to be tracked and for day and night time population estimates to be made.

References

Bates, A. (2008) The development of a postcode best fit methodology for producing population estimates for different geographies. Population Trends 133: 28-34.

Benenson, I. and Omer, I. (2000) Studies of the unique GIS of the Israeli Population Census: high-resolution urban patterns and individual residential segregation. Proceedings of the 3rd AGILE Conference on Geographic Information Science: Helsinki/Espoo

Boden P, Stillwell J. and Rees P (1992) How good are the NHSCR data? In Migration Processes and Patterns Volume 2: Population Redistribution in the United Kingdom edited by Stillwell J, Rees P and Boden P. Belhaven Press: London; 13-27

Boyle P and Feng Z (200q) A method for integrating the 1981 and 1991 British Census interaction data. Computers, Environment and Urban Systems, 26: 241-56

Calder, A. (2009) Building the address register for the 2011 Census. Population Trends, 138: 22-26.

Champion T, Fotheringham S, Rees P, Boyle P and Stillwell J (1998) The Determinants of Migration Flows in England: a Review of Existing Data and Evidence, a report prepared for the Department of the Environment, Transport and the Regions. Department of Geography, University of Newcastle upon Tyne: Newcastle

Clarke, J., Dewdney, J. C., Evans I. S., Rhind, D. W. & Visvalingam, M. (1980) People in Britain: a Census Atlas. OPCS / HMSO: London.

GeoPlace (2011) GeoPlace Spatial Address and Location Data. Available online via: http://www.geoplace.co.uk/

Gregory, I. N. (2002). The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. Computers, Environment and Urban Systems 26: 293-314.

Harris, J., Dorling, D., Owen, D., Coombes, M. and Wilson, T. (2002) Lookup Tables and New Area Statistics for the 1971, 1981 and 1991 Censuses. In: Rees, P., Martin, D. and Williamson, P. (eds.) The Census Data System. John Wiley & Sons, Chichester: 67-82.

Martin, D. (1989) Mapping population data from zone centroid locations. Transaction of the Institute of British Geographers NS, 14: 90-97.

Martin, D. (1996) Geographical Information Systems: socioeconomic applications. Routledge: London

Martin, D. (1996). An assessment of surface and zonal models of population. International Journal of Geographical Information Systems 10: 973-989.

Martin, D. (2002), Geography for the 2001 Census in England and Wales, Population Trends, 108: 7-15.

Mennis, J., and Hultgren, T., (2006) Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. Cartography and Geographic Information Science, 33: 179-194.

Norman P & Boyle P (2010) Using Migration Microdata from the Samples of Anonymised Records and the Longitudinal Studies. In Technologies for Migration and Population Analysis: Spatial Interaction Data Applications (eds.) John Stillwell, Adam Dennett. IGI Global: Hershey, New York: 133-151

Norman P & Riva M (2012) Population health across space and time: the geographical harmonisation of the ONS Longitudinal Study for England and Wales. Population, Space & Place 18: 483-502 DOI: 10.1002/psp.1705

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

OS. (2010) Address-Point user guide and technical information. Ordnance Survey: Southampton

Pointer (2011) Address Data and Gazetteers: Pointer. Available online via: http://maps.osni.gov.uk/CMSPages/moreinfo_address_data.aspx

Power, J. & Shuttleworth, I. (1997). Intercensal population change in the Belfast Urban Area 1971-91: The correlates of population increase and decrease in a divided Society. International Journal of Population Geography 3: 91-108.

Raper J F, Rind D W, Shepherd J W. (1992). Postcodes: the New Geography. Longman Scientific and Technical: Harlow

Rees P, Parsons J & Norman P (2005) Making an estimate of the number of people & households for Output Areas in the 2001 Census. Population Trends 122: 27-34

Schroeder, J. P. (2007). Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data. Geographical Analysis 39 (3), 311–335.

Shuttleworth, I. & Lloyd, C. D. (2007). Linking Northern Ireland Census of Population Data, 1971-2001. Economic and Social Research Council Grant Number: RES-000-23-0478. Available online via UK Data Archive Study Number 5672: http://www.data-archive.ac.uk/.

Simpson, L. (2002). Geography conversion tables: a framework for conversion of data between geographical units. International Journal of Population Geography 8: 69-82.

Tuomaala, J., and Uimonen, M. (1998) Introducing the New Object-Oriented Cadastral Information System (JAKO) of Finland. FIG congress proceedings, Brighton