Institute of Education

Next Steps

Sweep 8 – Age 25 Survey

CLS Data Note / User Guide to the data (First Edition)

David Church

Contents

Table of Contents

Introduction

There has been an increasing awareness of the value of geographically linked data in social scientific research, especially since the ‘GIS revolution’ of the early 1990s (Longley and Batty, 1996). Spatial data can be approached from a number of directions. For example, “longitudinal studies are particularly valuable to geographers because they chart change, collect information across various domains and are spatially referenced” (Ekinsmyth, 1996: 364). On the other hand, economists, particularly those of a more heterodox bent, are beginning to appreciate the value of spatially referenced data, especially in research into the economics of education (e.g. Gibbons et al, 2013 who use the National Pupil Database to estimate the effects of neighbourhood composition on teenagers’ behavioural and educational outcomes in England). Epidemiology and its associated disciplines are perhaps most consistently associated with investigating the spatial effects of the type of data collected across the different longitudinal cohort studies. For example, Christakis and Fowler (2007) used data from the Framingham Heart Study in the US to examine the spread of obesity in a large social network over 32 years while Tunstall et al (2010) used data from the Millennium Cohort Study to analyse the health outcomes of pregnant women who moved house. Two particularly fruitful fields are, firstly, the investigation of so-called ‘neighbourhood effects’ across a number of socio-economic domains (e.g. Lupton and Kneale (2012) used data from the 1970 British Cohort Study to investigate neighbourhood influences on teenage parenthood) and, secondly, network-based analyses of particular issues such as obesogenic environments (e.g Burgoine et al, 2014), accessibility to health-promoting community resources (e.g. Wolch et al, 2011) and the impact of built environment (morphological) characteristics on health and well-being (e.g. Sarkar et al, 2014).

However, balancing the obvious advantages of incorporating the spatial dimension within longitudinal social scientific research, there are a number of important limitations to be borne in mind when dealing with this type of data. The principal consideration is protecting the identity of cohort members, particularly in the current era of ‘open data’ and increasing linkage of previously disparate administrative datasets. It is recognised that there is a cultural dimension to the issue of data confidentiality, with, for example, Scandinavian countries taking a more laissez-faire approach, adopting the perspective that data gathered using public funds should be available for public consumption. More socially conservative states like the US and UK, on the other hand, have tended to take a much more protectionist approach to personal data (Exeter et al, 2014). At present, the UK Data Archive takes the approach that access to geo-referenced data below Government Office Region (GOR) level should be subject to increasing access restrictions the more likely the data is to reveal the identity of cohort members. Other limitations include non-uniformity of geo-identifiers used across different sweeps of the various cohort studies and varying levels of accuracy in terms of the geo-identifiers collected (a particular problem of early sweeps before the standardisation of unit postcodes).

Data and methodology

In order to enable the process of spatial analysis of longitudinal cohort study data, unit postcodes are gathered from the addresses collected during interview, which are then validated by the CLS Cohort Maintenance Team using a range of specialist software products from AFD[1]. This postcode data is then used to generate point data, usually within a GIS. There are a number of licensed and open source GIS packages available (e.g. ArcGIS[2], MapInfo[3] and QGIS[4]). The primary data source for spatialising longitudinal cohort study data within this software is the ONS Postcode Directory, available from the UK Data Service Census Support website[5]. This dataset has been released quarterly since 2004 (every February, May, August and November) and contains Ordnance Survey eastings and northings for each unit postcode centroid. These eastings and northings are spatialised in GIS in the form of ‘x’, ‘y’ points, usually to an accuracy of 1 metre of the mean postcode centroid[6]. Areal (polygon) representations of unit postcodes are produced in the OS ‘Code-Point with Polygons’ product, available from Edina Digimap[7]. The current (February 2017) release of the Postcode Directory contains some 42 different geographies dating back to 1991, encompassing Census, health, administrative, environmental and educational domains. Once again, boundaries for the majority of these geographies are available from Edina Digimap.

The point-based longitudinal cohort study data is associated with these geographies by means of a ‘field join’ (based on postcode and ‘pcds’ field in the ONSPD) or by a ‘spatial join’ (i.e. based on location) within a GIS. One drawback of the use of mean unit postcode centroids in geo-referenced data is their wide variation in size (typically unit postcodes in rural areas will cover a much larger extent than their urban counterparts). This means that greater accuracy will be achieved in geo-referencing the address at interview of cohort members who live in more urbanised areas. Recent developments in geo-coding with products such as OS ‘AddressBase’[8] mean that it is possible to create point data based on the grid references of individual properties/locations (thus with an accuracy of metres rather than tens/hundreds of metres achieved using unit postcode centroids). This is particularly important for the types of network and accessibility based analyses alluded to above, as well as specific types of environmental analysis, such as the impact on residents of the electromagnetic fields of high voltage overhead power lines (e.g. Swanson et al, 2014).

[1] www.afd.co.uk/
[2] Licensed software, available from http://www.esriuk.com/
[3] Licensed software, available from http://www.mapinfo.com/
[4] Open-source software, available from http://www2.qgis.org/en/site/
[5] http://census.edina.ac.uk//pcluts.html
[6] There are, however, a range of ‘grid reference positional quality indicators’, ranging from 1 (‘within the building of the matched address closest to the postcode mean’ to 9 (‘no grid reference available’). Some 85% of all postcodes have a positional quality indicator of 1.
[7] http://digimap.edina.ac.uk/digimap/home
[8] http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-products.html

Extent and nature of the data

England and Wales use the same naming conventions across different geographies. Post-devolution, Scotland has adopted slightly different naming conventions. For example, in both 2001 and 2011 Census geography, what are known as 'Lower Super Output Areas' and 'Middle Super Output Areas' in England and Wales are called 'Data Zones' and 'Intermediate Geographies' respectively in Scotland and the mean populations used to create these areal units also varies between E&W and Scotland[9]. The projected coordinate system used to display geo-referenced data across Great Britain (i.e. England, Wales and Scotland) is the British National Grid[10]. The range of Ordnance Survey products (e.g. MasterMap, AddressBase, OpenData) is available for Great Britain (i.e. excluding Northern Ireland).

Northern Ireland uses its own equivalent to the ONS Postcode Directory, called the Central Postcode Directory[11], and the process of spatialising geo-referenced data works in exactly the same way as with Great Britain data. Northern Ireland uses a different projected coordinate system from the rest of the UK, the Irish National Grid[12].

Availability of Datasets

Because of the potentially disclosive nature of these datasets, geographic identifiers below Government Office Region (GOR) are being released under Secure Access[13]; these are listed at Appendix A.

Appendix B lists the contents of the Secure Access Datasets.

Appendix C lists the geographies available on the Datasets available under the standard End User Licence.

As you will note in Appendix C, there are two versions of Next Steps Sweep 8 geographically linked data available, one set based on 2001 Census boundaries and the other set based on 2011 Census boundaries. Due to the potential for identification of a small number of individual cohort members (i.e. those located in 'slivers' between those areal units whose boundaries changed between 2001 and 2011), users can choose either, but not both, of these datasets.

Further Information

If you have any queries, please contact us at clsfeedback@ioe.ac.uk or view our website at http://www.cls.ioe.ac.uk.

[9] For example, the mean population of a 2011 LSOA is 1,500 whereas the mean population of a 2011 DZ is 808.

[10] WKID 27700, Authority EPSG

[11] Available, upon application and receipt of password, from http://www.nisra.gov.uk/geography/postcode.htm

[12] WKID 29902, Authority EPSG

[13] http://ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab

References

Burgoine, T., N. G. Forouhi, et al. (2014). "Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: population based, cross sectional study." Bmj **348**.

Christakis, N. A. and J. H. Fowler (2007). "The spread of obesity in a large social network over 32 years." New England journal of medicine **357**(4): 370-379.

Ekinsmyth, C. (1996). "Large-scale longitudinal studies: their utility for geographic enquiry." Area: 358-372.

Exeter, D. J., S. Rodgers, et al. (2014). "“Whose data is it anyway?” The implications of putting small area-level health and social data online." Health Policy **114**(1): 88-96.

Gibbons, S., O. Silva, et al. (2013). "Everybody needs good neighbours? Evidence from students’ outcomes in England." The Economic Journal **123**(571): 831-874.

Longley, P. A. and M. Batty (1996). Spatial analysis: modelling in a GIS environment, John Wiley & Sons.

Lupton, R. and D. Kneale (2012). Theorising and measuring place in neighbourhood effects research: The example of teenage parenthood in England. Neighbourhood Effects Research: New Perspectives. M. van Ham, D. Manley, N. Bailey, L. Simpson and D. Maclennan. Berlin, Springer.

Sarkar, C., C. Webster, et al. (2014). Healthy Cities: Public Health Through Urban Planning, Edward Elgar Publishing.

Swanson, J., T. Vincent, et al. (2014). "Relative accuracy of grid references derived from postcode and address in UK epidemiological studies of overhead power lines." Journal of Radiological Protection **34**(4): N81.

Tunstall, H., K. Pickett, et al. (2010). "Residential mobility in the UK during pregnancy and infancy: Are pregnant women, new mothers and infants ‘unhealthy migrants’?" Social Science & Medicine **71**(4): 786-798.

Wolch, J., M. Jerrett, et al. (2011). "Childhood obesity and proximity to urban parks and recreational resources: A longitudinal cohort study." Health & place **17**(1): 207-214.

APPENDIX A: Data Sources for Geographical Identifiers available under Secure Access

(please see overleaf)

Dataset	Description	Boundary Data Source (UK Data Service unless stated)
Interview Wards	1998 Ward Boundaries	English Electoral Wards, 1998 Welsh Electoral Wards, 1998 Scottish Electoral Wards, 1998 Northern Ireland Electoral Wards, 2001[15]
	2001 Census Area Statistic Ward[14]	English Census Area Statistic Wards, 2001 Welsh Census Area Statistic Wards, 2001 Scottish Census Area Statistic Wards, 2001 Northern Ireland Electoral Wards, 2001[16]
Interview Output Area (OA)	2001 Output Area	English Output Areas, 2001 Welsh Output Areas, 2001 Scottish Output Areas, 2001 Northern Ireland Output Areas, 2001
	2011 Output Area	English Output Areas, 2011 Welsh Output Areas, 2011 Scottish Output Areas, 2011 Northern Ireland Small Areas, 2011

[14] see http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/electoral-wards-divisions/statistical-wards--cas-wards-and-st-wards/index.html for definition and http://census.edina.ac.uk//pcluts_download.html?data=pcluts_2014nov for latest ONS Postcode Directory code/name lookups

[15] Northern Ireland Electoral Wards remained unchanged between 1998 and 2001

[16] Northern Ireland does not have CAS wards but the electoral wards have alternative codes in Census outputs.

Dataset	Description	Boundary Data Source (UK Data Service unless stated)
Interview Lower Super Output Area (LSOA)	2001 Lower Super Output Area	English Lower Super Output Areas, 2001 Welsh Lower Super Output Areas, 2001 Scottish Datazones, 2001 Northern Ireland Lower Super Output Areas, 2001
	2011 Lower Super Output Area	English Lower Super Output Areas, 2011 Welsh Lower Super Output Areas, 2011 Scottish Lower Super Output Areas, 2011 Northern Ireland Super Output Areas, 2011[17]
Interview Middle Super Output Area (MSOA)	2001 Middle Super Output Area	English Lower Super Output Areas, 2001 Welsh Lower Super Output Areas, 2001 Scottish Intermediate Geographies, 2001
	2011 Middle Super Output Area	English Lower Super Output Areas, 2011 Welsh Lower Super Output Areas, 2011 Scottish Intermediate Geographies, 2011
Interview Local Authority District	2001 Local Authority District/Unitary Authority	English Administrative Districts, 2001 English Unitary Authorities, 2001 Welsh Unitary Authorities, 2001 Scottish Council Areas, 2001 Northern Ireland District Councils, 2001

[17] In both 2001 and 2011 Census Geography, Northern Ireland does not have Middle Super Output Areas

Dataset	Description	Boundary Data Source (UK Data Service unless stated)
	2011 Local Authority District/Unitary	Ordnance Survey Boundary-Line, November 2013[18]
Interview Westminster Parliamentary Constituency	2001 Westminster Parliamentary	English Westminster Parliamentary Constituencies, 2001 Welsh Westminster Parliamentary Constituencies, 2001 Scottish Westminster Parliamentary Constituencies, 2001 Northern Ireland Westminster Parliamentary Constituencies, 2001
	2011 Westminster Parliamentary Constituency (Sweep 5 only)	Ordnance Survey Boundary-Line, November 2011[19]

[18] Available from http://digimap.edina.ac.uk

[19] Available from http://digimap.edina.ac.uk

APPENDIX B: Variables available under Secure Access

Geographical Identifiers (based on address at interview), Next Steps Sweep 8 (2001)

Variable name	Description
NSID	Next Steps survey id
W8CTRY	Sweep 8 Country at Interview
W8GOR	Sweep 8 Former Government Office Region at Interview
W8WARD98	Sweep 8 Ward Code (1998 Boundaries)
W8CASWARD	Sweep 8 Census Statistic Ward Code 2001
W8OA01	Sweep 8 Output Area Code 2001
W8LSOA01	Sweep 8 Lower Super Output Area Code 2001
LSMSOA01	Sweep 8 Middle Super Output Area 2001
W8OSLAUA	Sweep 8 LA District/Unitary Authority 2001
W8PCON	Sweep 8 Westminster Parliamentary Constituencies 2005
W8LEA	Sweep 8 Local Education Authority 2009

Geographical Identifiers, (based on address at interview), Next Steps Sweep 8 (2011)

Variable name	Description
NSID	Next Steps survey id
W8CTRY	Sweep 8 Country at Interview
W8GOR	Sweep 8 Former Government Office Region at Interview
W8WARD98	Sweep 8 Ward Code (1998 Boundaries)
W8CASWARD	Sweep 8 Census Statistic Ward Code 2001
W8OA11	Sweep 8 Output Area Code 2011
W8LSOA11	Sweep 8 Lower Super Output Area Code 2011
LSMSOA11	Sweep 8 Middle Super Output Area 2011
W8OSLAUA	Sweep 8 LA District/Unitary Authority 2011
W8PCON	Sweep 8 Westminster Parliamentary Constituencies 2005
W8LEA	Sweep 8 Local Education Authority 2009