ACHIEVING DATA COMPATIBILITY OVER SPACE AND TIME: CREATING CONSISTENT GEOGRAPHICAL ZONES

Paul Norman [1]

Philip Rees [1]

Paul Boyle [2]

[1] School of Geography
University of Leeds, Leeds, LS2 9JT, UK
e-mail p.norman@geography.leeds.ac.uk
e-mail p.rees@geography.leeds.ac.uk

[2] School of Geography and Geosciences
University of St Andrews, St Andrews, KY16 9ST, UK
e-mail p.boyle@st-andrews.ac.uk

October 2001

CONTENTS

ABSTRACT iii
LIST OF TABLES iv
LIST OF FIGURES iv

1. INTRODUCTION 1

2. WARD BOUNDARY CHANGE SCENARIOS 3

3. APPROACHES TO CREATING CONSISTENT GEOGRAPHIES 5
3.1 Freeze history 5
3.2 Update to contemporary zones 5
3.3 Construct designer zones 6
3.4 Geocode individual/household-level data 7
3.5 Data availability for establishing a consistent ward geography in Eastern Region 8

4. GEOGRAPHICAL DATA CONVERSION USING LOOKUP TABLES: GENERAL PRINCIPLES 11
4.1 Geographical data conversion: background 11
4.2 Lookup table concepts 12
4.3 Defining geographical data conversion lookup table frameworks 13
4.4 Lookup tables for geographical data conversion: worked example 18
4.5 Choice of GCT or GML framework 19
4.6 Deriving intersection weights for imperfect (dis)aggregations 20

5. CREATING A CONSISTENT GEOGRAPHY USING GEOGRAPHICAL CONVERSION TABLES 21
5.1 Postcode locations for geographical data conversion 21
5.2 ED Building-Brick approach to data conversion 26
5.3 Postcode-Point approach to data conversion 46

7. A CONSISTENT GEOGRAPHY FOR WELWYN-HATFIELD: WORKED EXAMPLE 56

8. CONCLUSIONS 60

ACKNOWLEDGEMENTS 64

REFERENCES 66

ABSTRACT

Geographers have long-recognised the importance of boundary specification and the problems of using arbitrarily defined areas for the collection and dissemination of socioeconomic data. The focus has tended to be on the modifiable areal unit problem and on custom zone design with the problems created by temporal inconsistencies in zonal boundaries having less consideration. This is surprising as alongside occasional major structural reorganisations the UK experiences frequent administrative boundary changes causing difficulties in producing comparable statistics over time. Unless a consistent geographical approach with time-series data is taken it cannot be known whether changes are real or an artefact of boundary changes.

The late 1990s has seen initiatives from ONS to promote harmonisation of geographical information and the Update UK Area Masterfiles (UUKAM) project which allows the conversion of data between 1991 census and various late-1990s geographies. For studies which predate the ONS initiative and exceed the data conversions possible through the UUKAM project, a method must be devised to establish a data time-series on a consistent geographical basis otherwise the data quality will be compromised and analyses cannot objectively be compared over time. After illustrating the nature of the boundary change problems to be overcome, this paper describes and appraises methods by which data can be adjusted to an appropriate geography. The paper concludes with a list of advised checks for researchers carrying out similar work.

Keywords: temporal consistency, boundary changes, data conversion lookup tables

LIST OF TABLES

1. Joining boundary and population tables using a common geographical reference 11
2a. GCT data conversion from ‘many’ source zones to ‘many’ target zones 18
2b. GML data conversion from ‘many’ source zones to ‘many’ target zones 18
3. Availability of relevant lookup table information 26
4. Postcode-ward-ED linked lookup table variables 32
5. Ward numbers in the 1989-98 ward pedigree 34
6. Ward of year-postcode-ward 1998 linked lookup table variables 49

LIST OF FIGURES

1. Ward boundaries 1991 and 1998, Welwyn-Hatfield district 4
2. Types of links between sets of geographical zones 14
3. One to many relationships (disaggregation) 15
4. Many to one relationships (aggregation) 16
5. Many to many relationships (disaggregation-reaggregation) 17
6. 1991 postcode point and ED population distributions in Eastern Region 23
7. Constituent postcodes of wards in 1991 and 1998 (Haldens/Panshanger) 24
8. ED Building-Brick data disaggregation-reaggregation approach 27
9. Algorithm to calculate ‘freeze history’ source-target GCT1 weights 35
10. Algorithm to calculate ‘update to contemporary zones’ source-target GCT2 weights 36
11. Algorithm to use GCTs to carry out data disaggregation-reaggregation process 36
12. Adjustment of hypothetical populations to 1998 geography, ED version 1 38
13. Proportional change of ward populations 1997-98 38
14. Adjustment of hypothetical populations to 1998 geography, ED version 2 40
15. Ward boundaries 1991 and 1998, Thurrock and Peterborough Unitary Authorities 41
16. Estimated and original ED and ward populations 42
17. Estimated and original ED and ward populations 44
18. Postcode and household counts for each ED 45
19. Postcode-Point data disaggregation-reaggregation approach 47

20. Algorithm to calculate ward of year-postcode-ward 1998 GCT weights and adjust input data to target geography 50

21. Adjustment of hypothetical populations to 1998 geography, Postcode-Point household count approach 51

22. Adjustment of hypothetical populations to 1998 geography, Postcode-Point postcode count approach 52

23. Percentage differences between the use of household and postcode counts 52

24. Relationships between ward-level postcode and household counts with population 54

25. Comparison of postcode point and UUKAM estimated SAS populations 55

26. Crude death rates in Welwyn-Hatfield using non-consistent geographical boundaries 58

27. Crude death rates in Welwyn-Hatfield using consistent geographical boundaries 59

ACHIEVING DATA COMPATIBILITY OVER SPACE AND TIME: CREATING CONSISTENT GEOGRAPHICAL ZONES

1. INTRODUCTION

Geographers have long-recognised the importance of boundary specification and since the development of high-performance computing and Geographical Information Systems (GIS) much progress has been made in addressing problems posed by the use of arbitrarily defined areas for the collection and dissemination of socioeconomic data (Blake *et al.* 2000). Observations that different statistical results are obtained when different geographical boundaries and area subdivisions are used (Openshaw 1991) has led to a focus on the modifiable areal unit problem (MAUP) (Openshaw and Taylor 1981) and on custom zone design (Openshaw and Rao 1995). However, the problems created by temporal inconsistencies in zonal boundaries have had less consideration.

Alongside occasional major structural reorganisations the UK is subject to more administrative boundary changes over time than the rest of Europe put together (ONS 2000). For example, ward boundaries are regularly adjusted in response to population change to ensure that each local authority has similar councillor/elector ratios (UKSGB 2000) and the local government structure of the UK was substantially revised in the late 1990s with the creation of new unitary authorities in some areas (Wilson and Rees 1999). Furthermore, even if boundaries do not change, area names and reference codes often vary with different versions and spellings used across years and between different data suppliers. These issues are compounded by the large number of different geographies for which data are available and technical difficulties because the commonly used administrative and postal geographies do not align.

This degree of change and complexity causes difficulties in the production of comparable statistics over time. Unless a consistent geographical approach with time-series data is taken it cannot be known whether changes are real or an artefact of a boundary change. The broader context of this research is to estimate ward-level populations and calculate Standardised Mortality Rates (SMRs) for each mid-year 1990-1998 for the National Health Service (NHS) Eastern Region Public Health Observatory (ERPHO). However, the ward boundaries in Eastern Region have been potentially undergoing both small incremental changes throughout the 1990s as well as more substantial changes in the number and names of the wards with local government restructuring. In the former case, incremental changes are very difficult to determine, in the latter, the fact that structural changes have occurred is readily identifiable but details about the boundary changes are not.

ONS (2000) recognise it is critical to adopt a coordinated approach to geography and a need to be consistent in the use of names, codes and references. As a result, mechanisms are now in place to promote harmonisation within National Statistics including the ONS Geographic Referencing Strategy and the establishment of Gridlink, a core set of postcode location data. Addressing some of these issues, the Updated UK Area Masterfiles (UUKAM) project allows the conversion of data between 1991 census and various late-1990s postal and administrative geographies (Simpson 2001).

Since the study period predates the ONS initiative and exceeds the data conversions possible through the UUKAM project, a method must be devised to establish a data time-series on a consistent geographical basis otherwise the data quality will be compromised and the population estimates and SMRs cannot objectively be compared over time. After an illustration of the nature of the boundary change problems that need to be overcome, this paper describes and appraises methods by which data can be adjusted to the most appropriate geography for the application. An example of data presented in both non-consistent and consistent geographical bases is given and the paper concludes with a list of advised checks that researchers should carry out if attempting similar work.

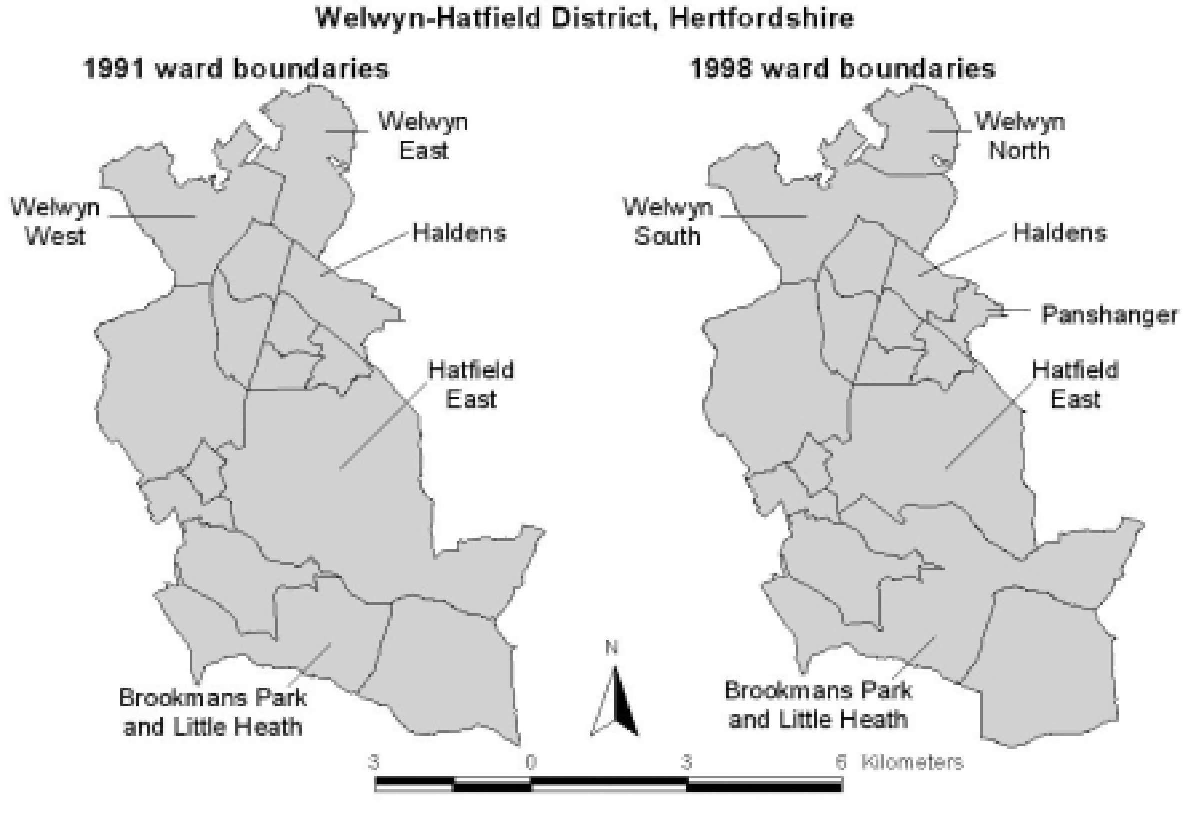
2. WARD BOUNDARY CHANGE SCENARIOS

To carry out research over time at the electoral ward scale, three boundary change scenarios may need to be overcome. These are described below and illustrated in Figure 1 which shows the 1991 Census wards and 1998 electoral wards for Welwyn-Hatfield, a local authority district in Hertfordshire.

1. In the north of the district, the 1991 wards Welwyn West and Welwyn East have changed to Welwyn South and Welwyn North by 1998. The new wards occupy the same combined area as previously but the boundary between the two has changed.
2. To the south of these in 1991 is Haldens ward which by 1998 had been subdivided with the creation of an additional ward, Panshanger.
3. In the central and southern parts of the district in 1991 are the wards of Hatfield East and Brookmans Park and Little Heath. The 1998 boundaries show that a substantial part of Hatfield East has been transferred to Brookmans Park and Little Heath.

It is straightforward to identify that wards have changed name (and/or reference code) and that new wards have been created. However, when ward names stay the same but their areal extents alter this is impossible to identify without detailed maps. Moreover, without a time-series of digital boundaries or regular surveillance of Hansard it is difficult to know when boundary changes take place. In fact the alterations to the wards in Welwyn-Hatfield were in May 1991 (Baker 1991) just after the Census thereby creating difficulties in the interpretation of data relevant to the post-Census geography.

Figure 1: Ward boundaries 1991 and 1998, Welwyn-Hatfield district



For sources see Acknowledgements sections b and g

3. APPROACHES TO CREATING CONSISTENT GEOGRAPHIES

Various strategies can be taken to standardise spatial systems to address the problem of establishing time-series data on a consistent geographical basis when zone boundaries change over time. These strategies include ‘freezing geographical history’, ‘updating to contemporary zones’, ‘constructing designer zones’ and ‘geocoding individual/household data’.

3.1 Freeze history

The freeze history approach involves fixing the zone system at a point in time and systematically tracking boundary changes so that later observations can be adjusted back to the original boundaries. This approach is adopted by EUROSTAT for monitoring trends at the most detailed NUTS5 regional level (Blake *et al.* 2000). It is also possible to use areal interpolation, a technique that involves the apportionment of data from one set of zones to another. In its simplest form the size of the source and target zones is used to weight the source zone values by the area of overlap with the target zone (Flowerdew and Green 1994). This is the method used by White *et al.* (1998) to fit data from other time periods to 1911 Registration Districts for the analysis of long-term change. A disadvantage of freezing geographical history is that, over time, the chosen zones will become less appropriate to current applications.

3.2 Update to contemporary zones

An alternative approach is to update data from previous spatial systems to a set of contemporary zones. This is also achievable by areal interpolation but can involve the use of lookup tables (LUTs) to link building-bricks from a previous geography to the current system thereby enabling the conversion of earlier data (Wilson and Rees 1998, 1999). For example, ONS geography (ONS 2000) track post-1991 ward boundary amendments and can supply LUTs of changes in ED/ward relationships. This approach is applicable for ward geographies but needs continual attention to note the boundary changes. A difficulty of using LUTs for data conversion is that direct measures of the weights to be applied to convert from one geography to another are rarely available so that surrogate measures must be used. In the ONS LUTs data conversion weights can be derived since the number of persons involved in the boundary change is noted. For other applications weights can usefully be derived using postcode/household counts from a source such as the UK Enumeration District (ED) to postcode directories. Since postcode locations can be used as small geographical data conversion building-bricks flexibility in subsequent aggregation is possible. Penhale *et al.* (2000) use GIS intersections of 1991 EDs and 1998 wards with postcode counts of delivery points as a proxy for population distribution to apportion data where EDs are split.

3.3 Construct designer zones

Another solution is to construct designer zones from smaller building-bricks to harmonise the zonal system based on boundaries that are common in different years. This approach is complex to apply and the harmonised boundary solutions may not match current geography. Moreover, to carry out research on relatively large study regions the necessary GIS digital boundary sets must be available for all time periods and even when they are there can be no guarantee that digitising discrepancies (through either error or changes in areal definitions) do not suggest boundary changes when none have occurred. For an analysis of inter-regional migration in Australia, Blake *et al.* (2000) attempted the assembly of small building-bricks to generate designer zones with coincident outer boundaries across a number of censuses. Unfortunately, differences in digitising boundaries undermined the development of an automated solution, so successive overlays of boundaries had to be checked manually followed by estimation of the area and size of population involved and a search for the nearest consistent boundary. Apart from being laborious, this approach became a compromise between the competing goals of temporal consistency and the maintenance of the functional integrity of the statistical division zones.

A further designer zone approach includes the creation of a new set of areas to represent the data in a more realistic way than existing administrative geographies (Openshaw and Rao 1995), though what may be deemed realistic at one time may be inappropriate at another. Developing an earlier approach by Norris and Mounsey (1983), Bracken and Martin's (1995) surface models of 1981 and 1991 census data enable moves from conventional geographies with the raster cell values used as building-bricks to be aggregated into user-defined zones. Despite this utility it is only during a census year when detailed and nationally consistent small area sociodemographic data become available that a UK population surface can be generated. Moreover, the complexity of linking data precludes their use as a means of establishing a standard geography for a year by year data time series.

3.4 Geocode individual/household-level data

The geocoding of individual/household-level data to discrete addresses avoids the difficulties noted above. Data can be assembled for the most appropriate set of boundaries to the problem being investigated ensuring temporal consistency and application relevance. The main constraints relate to cost of establishment and the need in many applications to guarantee confidentiality. This approach is being employed for the output of the 2001 Census (ONS 1999).

3.5 Data availability for establishing a consistent ward geography in Eastern Region

In practice, the choice of method is largely dictated by the nature, availability and cost of the input data rather than necessarily being application-driven. In the research being reported here datasets will only be used that are nationally consistent, widely available and inexpensive to academics, government organisations, health authorities and other health professionals. Given a need for a recent geography the aim is to standardise the ward input data needed for 1990-98 annual population estimates and SMRs to the 1998 ward geography, the latest year for which all input data are readily available. The ward-level data sets to be adjusted comprise i) 1991 Census populations and migrant counts ii) the electorate for each year 1990-98 and iii) the Vital Statistics births and deaths data for each year 1989-98. The most straightforward approach by which the ward data could be updated to the 1998 ward geography would be to intersect polygon GIS coverages for each year with 1998 and, assuming equal population density, to adjust the inputs based on area of overlap using simple areal interpolation. Unfortunately, the only appropriate GIS coverages freely available are 1991 ward and Enumeration District (ED) boundaries and the 1998 Unitary Authority boundaries so that GIS intersection approaches are impossible. (N.B. 1998 ward boundaries are available to this project through a special arrangement with OS, see Acknowledgements section g, but will only be used for illustrative purposes and not for GIS geometrical techniques.) The designer zone approach referred to above is unnecessary since it is a ward geography that is required. This is not to say that electoral wards are necessarily the ‘correct’ geography for health applications since their boundary specifications are arbitrary and modifiable (Barr 1993), but they are a convenient and frequently-used administrative geography of relatively small size for which sociodemographic data are regularly collected and disseminated.

Household-level geocoded data are outside the data-acquisition scope of this research but the potential still exists to create a consistent frozen geography if data are available for units that are small enough to be able to construct zones at one point in time from those at another. Shaw *et al.* (1998) point out this is feasible in post-1981 mortality studies since the postcode of last residence is attached to computerised files. In the UK there are about 1.7 million postcodes covering approximately 25 million addresses. Postcodes are created and maintained by Royal Mail to enable the efficient delivery of mail (Raper *et al.* 1992). The postcode has evolved since the early 1960s as key data to provide a spatial reference (UKSGB 2001) with the utility of the postcode system as a geographical referencing system acknowledged in the Chorley Report on the Handling of Geographic Information (DoE 1987). Pertinently the report notes that data are collected for numerous incompatible geographical regions which do not nest into each other and do not have boundaries consistent over time. Heywood (1997) notes some drawbacks of the postcode as a spatial reference but believes Chorley’s endorsement of the unit postcode changed the way most socioeconomic data are managed with their widespread adoption for spatial referencing by many types of organisations (Raper *et al.* 1992). In 1991 postcodes were recorded for the first time on individual census forms and provided the basis for the creation of the Postcode-Enumeration District Directory (PCED) by the then Office of Population Censuses and Surveys (OPCS). Postcode locations allow detailed data modelling without the necessity for digitised boundary information; an advantage being their small size at unit level (typically 14 residential addresses) which offers versatility for aggregation into other areal units (Martin 1992).

The UUKAM project (Simpson 2001) demonstrates data conversion using LUTs to link 1991 geographies to 1998 zones using postcode counts to derive apportionment weights between the geographies but links from other years are not available. Various postcode-based LUTs are available to the academic community via Manchester Information and Associated Services (MIMAS) that enable linkage between geographical areas and between time periods (MIMAS 1999). Before reporting the detailed method by which ward-based data for all the years in the study period can be adjusted to be consistent with the 1998 ward geography, it is useful to examine the general principles of using LUTs for geographical data conversion.

4. GEOGRAPHICAL DATA CONVERSION USING LOOKUP TABLES: GENERAL PRINCIPLES

4.1 Geographical data conversion: background

Lookup tables (LUTs) are database devices whereby sets of entities in different files can be linked using reference items that are common between each file. In GIS packages this type of database operation is carried out when tables of ward-based population data are 'joined' to an attribute table of vector ward boundaries using an item, in this case the ward reference code, that is present in both tables (see Table 1).

Table 1: Joining boundary and population tables using a common geographical reference

Entity	Area	Perimeter	Ward-Name	*Ward-Ref*		*Ward-Ref*	Population
Polygon	1495850	6270	Biscot	*10DJFA*	***Common linking item***	*10DJFA*	11644
Polygon	2767397	8252	Bramingham	*10DJFB*		*10DJFB*	12706
Polygon	2472973	8282	Challney	*10DJFC*		*10DJFC*	9995
Polygon	6274450	12006	Crawley	*10DJFD*		*10DJFD*	10665
Polygon	2247373	7964	Dallow	*10DJFE*		*10DJFE*	10399

A frequent problem in many applications is that the areal units for which data are available are not necessarily the units that are required (Flowerdew and Green 1994). A specialised subset of database LUT operations is their use for the conversion of data from one geographical zonal system to another with the LUT entities being sets of discrete geographical spatial units. The main purposes of geographical data conversion LUTs are to enable (Simpson 2001):

- The transformation of 1991 Census data from standard output (Enumeration Districts, wards, districts, etc.) to: recent postal geography; revised local government or electoral areas; special planning areas (enterprise zones, national parks, travel to work areas, functional zones) or; different geometrical areas (e.g. grid squares for linking with environmental data).

- The aggregation of data to units sufficiently large to provide reliable results (e.g. from postcoded events to local authority district).
- Analysing and presenting results for areas familiar to the audience of the research (e.g. new health areas such as Primary Care Groups).
- Merging of data sets drawn from different sources (e.g. for neighbourhood profiles containing both census and postal geography data).
- Estimating a time-series on a consistent geographical basis (e.g. Vital Statistics from wards before and after boundary changes).

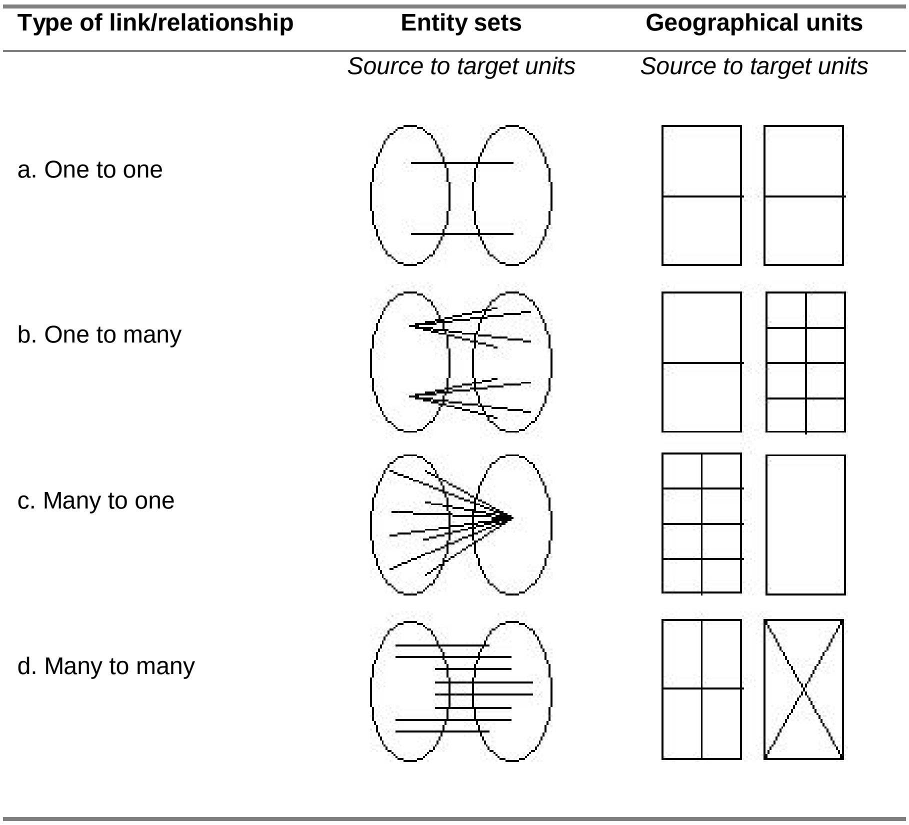
4.2 Lookup table concepts

Various types of database links can exist between LUT entities. Wilson and Rees (1998) conceptualise the relationships between sets of zones that are relevant to the handling of geographical data. Figure 2a illustrates simple ‘one to one’ links which, in terms of geographical zonal systems, occurs when zone names change, but boundaries do not. ‘One to many’ relationships are shown in Figure 2b whereby a source area may be disaggregated into a subset of smaller target areas, for example, the situation existing between Census wards and their constituent Enumeration Districts (EDs). The reverse situation illustrated by Figure 2c shows ‘many to one’ links whereby the source geography perfectly aggregates into a larger target geography. Essentially a nested geographical hierarchy, this circumstance often exists at a single point in time (e.g. Census EDs aggregate to wards which aggregate to local authority districts etc.; unit postcodes aggregate to postal sectors which aggregate to postal districts etc.) but boundaries are not necessarily consistent between time points. The ‘many to many’ relationships shown by Figure 2d exist when a zone in the source geography intersects with several zones in the target geography and similarly a zone in the target geography receives contributions from several source zones. This is typified by the lack of coterminous boundaries between census and postal geographies (Martin 1996). The data conversion necessary in this situation is a disaggregation of the source geography followed by a reaggregation into the target geography.

4.3 Defining geographical data conversion lookup table frameworks

The first three relationships described above may be readily defined for standard geographical zonal systems but ‘many to many’ links and non-standard aggregations and disaggregations require LUTs to be devised for data conversion from source to target geographies. The fundamental information necessary within data conversion LUTs are reference codes to both the source geography in which the data pre-exist and the target geography into which the data are to be converted together with an indication of the extent of overlap, a weight, between each zone in the source geography and each zone in the target geography. These weights, taking a value of more than zero but less than or equal to one, will sum to one across intersections by source area (Simpson 2001; Wilson and Rees 1998). Weights are not needed if the link is a one to one change of name relationship or if the links are one to many or many to one calculated on a simple ‘best-fit’ basis; the latter being a list of source zones paired up with the target zone in which the majority of the source zone lies. Whilst crude, this is often used if information needed to apportion source to target zones is not available (Wilson and Rees 1998).

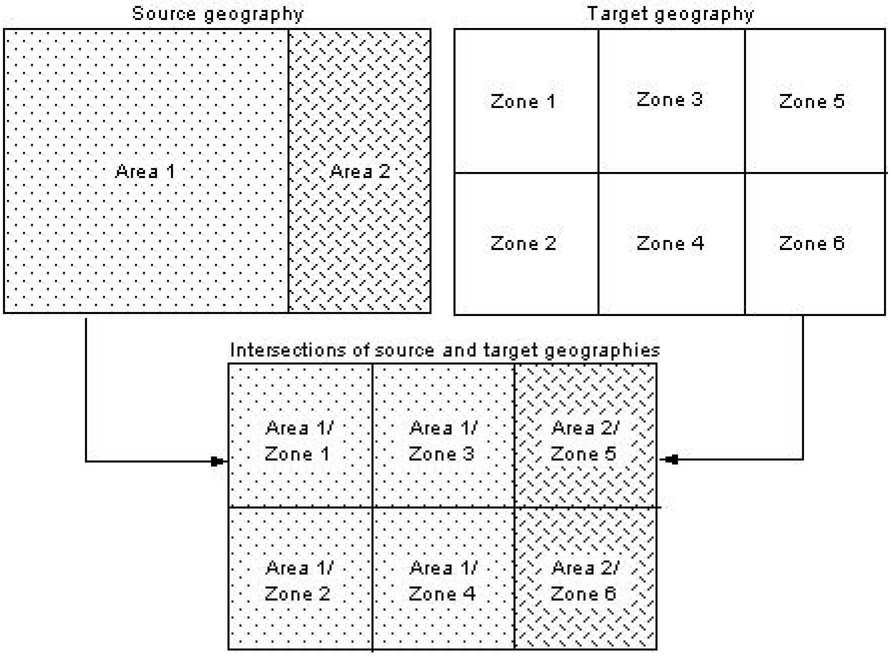
Figure 2: Types of links between sets of geographical zones



Source: after Wilson and Rees, 1998: 2

To support what is essentially a ‘building-brick’ approach, the information in data conversion LUTs can be organised in two different frameworks both of which have the ability to support the relationships illustrated in Figure 2. The first LUT framework is a ‘Geographical Conversion Table’ (GCT) and the second a ‘Geographical Membership List’ (GML). These frameworks are described and differentiated below.

Figure 3: One to many relationships (disaggregation)



Geographical Conversion Table

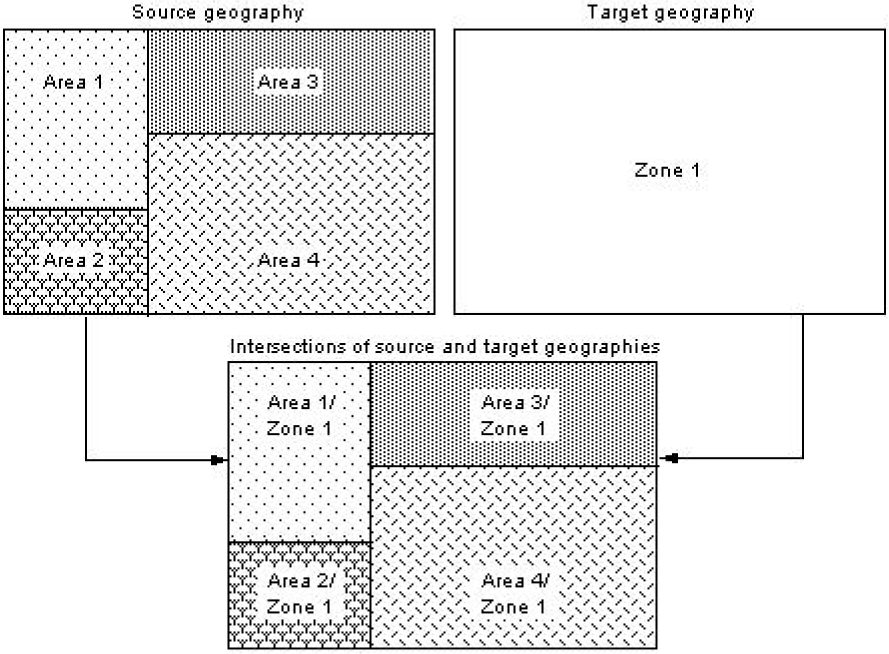
Source units	*Weight*	*Target units*
Area 1	0.25	Zone 1
Area 1	0.25	Zone 2
Area 1	0.25	Zone 3
Area 1	0.25	Zone 4
Area 2	0.50	Zone 5
Area 2	0.50	Zone 6

Geographical Membership List

Target units	*Weighted membership list*
Zone 1	Area 1 * 0.25
Zone 2	Area 1 * 0.25
Zone 3	Area 1 * 0.25
Zone 4	Area 1 * 0.25
Zone 5	Area 2 * 0.50
Zone 6	Area 2 * 0.50

Diagrammatic maps of source and target geographies and their intersections in Figure 3 illustrate the one to many relationships conceptualised in Figure 2b showing how Area 1 of the source geography is disaggregated into Zones 1 to 4 of the target geography and Area 2 is disaggregated into Zones 5 and 6. Each line of the GCT framework has a reference to a source unit, the target unit that the source unit is associated with and a weight to indicate the amount of overlap between source and target units; the first line showing that 0.25 of Area 1 overlaps with Zone 1. In the GMT framework the same data is arranged differently so that each line has a reference to a complete unit of the target geography and a reference to its constituent building-brick source unit. The weights indicate the proportion of the source units used for the building-brick.

Figure 4: Many to one relationships (aggregation)



Geographical Conversion Table

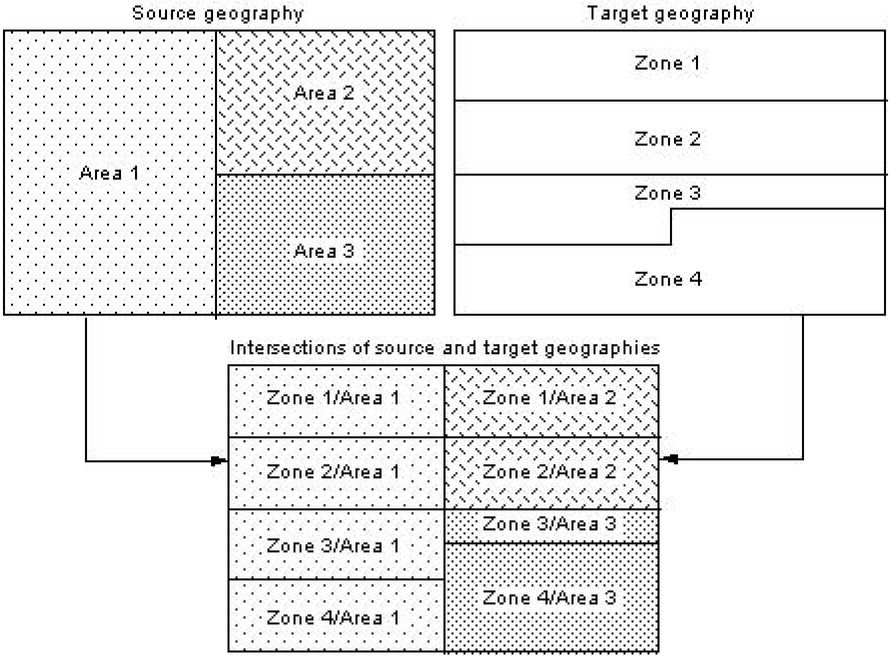
Source units	*Weight*	*Target units*
Area 1	1.00	Zone 1
Area 2	1.00	Zone 1
Area 3	1.00	Zone 1
Area 4	1.00	Zone 1

Geographical Membership List

Target units	*Weighted membership list*
Zone 1	(Area 1 * 1.00)+(Area 2 * 1.00)+ (Area 3 * 1.00)+(Area 4 * 1.00)

Figure 4 illustrates the many to one relationships shown in Figure 1c that are the reverse of the previous situation. The maps show that Areas 1 to 4 of the source geography wholly nest into Zone 1 of the target geography. Thus the GCT shows the link between each source unit and associated target unit each with a weight of 1.00 since all of every source unit overlaps the target. The GML arrangement of the data shows that Areas 1 to 4 aggregate perfectly into Zone 1.

Figure 5: Many to many relationships (disaggregation-reaggregation)



Geographical Conversion Table

Source units	*Weight*	*Target units*
Area 1	0.25	Zone 1
Area 1	0.25	Zone 2
Area 1	0.25	Zone 3
Area 1	0.25	Zone 4
Area 2	0.50	Zone 1
Area 2	0.50	Zone 2
Area 3	0.25	Zone 3
Area 3	0.75	Zone 4

Geographical Membership List

Target units	*Weighted membership list*
Zone 1	(Area 1 * 0.25)+(Area 2 * 0.50)
Zone 2	(Area 1 * 0.25)+(Area 2 * 0.50)
Zone 3	(Area 1 * 0.25)+(Area 3 * 0.25)
Zone 4	(Area 1 * 0.25)+(Area 3 * 0.75)

The more complex many to many relationships previously shown in Figure 1d are illustrated in Figure 5. The diagrammatic maps show that each source unit contributes to more than one target unit and that each target unit receives contributions from more than one source unit. The GCT defines the weight by which the source unit should be disaggregated to its associated target zone with the first line showing that 0.25 of Area 1 overlaps with Zone 1. The GML shows a weighted list of the building-bricks of each target unit. For example, Zone 1 of the target geography consists of 0.25 of Area 1 and 0.50 of Area 2.

Table 2a: GCT data conversion from 'many' source zones to 'many' target zones

1. Source unit populations		**2. Geographical Conversion Table**			**3. Estimates for intersections**	**4. Estimates for target units**	
Source ref.	*Population*	*Source ref.*	*Weight*	*Target ref.*	*Disaggregated data*	*Reaggregated data*	
Area 1	1500	Area 1	0.25	Zone 1	1500*0.25=375	Zone 1	375+500=875
Area 2	1000	Area 1	0.25	Zone 2	1500*0.25=375	Zone 2	375+500=875
Area 3	400	Area 1	0.25	Zone 3	1500*0.25=375	Zone 3	375+100=475
		Area 1	0.25	Zone 4	1500*0.25=375	Zone 4	375+300=675
		Area 2	0.50	Zone 1	1000*0.50=500		
		Area 2	0.50	Zone 2	1000*0.50=500		
		Area 3	0.25	Zone 3	400*0.25=100		
		Area 3	0.75	Zone 4	400*0.75=300		

Table 2b: GML data conversion from 'many' source zones to 'many' target zones

1. Source unit populations		**2. Geographical Membership List**		**3. Estimates for target units**	
Source ref.	*Population*	*Target ref.*	*Weighted membership list*	*Converted data*	
Area 1	1500	Zone 1	(Area 1*0.25)+(Area 2*0.50)	Zone 1	375+500=875
Area 2	1000	Zone 2	(Area 1*0.25)+(Area 2*0.50)	Zone 2	375+500=875
Area 3	400	Zone 3	(Area 1*0.25)+(Area 3*0.25)	Zone 3	375+100=475
		Zone 4	(Area 1*0.25)+(Area 3*0.75)	Zone 4	375+300=675

4.4 Lookup tables for geographical data conversion: worked example

To illustrate data conversion when many to many LUT relationships exist, the weights given in Figure 5 have been applied to hypothetical populations. In Table 2a the population counts (column 1) are converted from source to target geographies using the GCT approach. This is achieved by applying the weights in the GCT (column 2) to the source unit populations to estimate population counts in the source/target intersections (column 3). These intersection population estimates are then reaggregated into the relevant target units (column 4). In Table 2b the GML approach is adopted. The source unit populations (column 1) are apportioned into building-bricks using the weights in the GML (column 2). The membership lists of building-bricks are then aggregated into the target units (column 3).

4.5 Choice of GCT or GML framework

Although the GCT and GML frameworks basically contain the same information there can be advantages to either approach. In terms of the literature it should be noted that various data conversion LUT terms exist: a GCT is equivalent to a SASPAC 'gazetteer' file (SASPAC 1992: 7/57-7/62) and a GML is equivalent to a SASPAC 'new zone … using areas' procedure (SASPAC 1992: 7/63-7/65) and is analogous to 'area constitutions' in ONS parlance (ONS 2000). Wilson and Rees (1998: 20) compare SASPAC 'gazetteer' and 'new zone … using areas' procedures. For converting census data an advantage of the latter is that subtractions as well as additions are feasible so that a target zone can be defined as comprising (Ward01) + (Ward02 – ED03); thereby avoiding error propagation due to data blurring if just a list of EDs were used. In this way Wilson and Rees (1998: 46) utilise a GML approach to define a LUT of 1998 local authority areas in terms of 1991 Census areas that enables conversion of 1991 Census data into the 1998 geography.

The GCT framework in comparison with the GML approach is more straightforward to write generic computer programs for and potentially more versatile. This is exemplified by the UUKAM project (Simpson and Yu 2000; Yu and Simpson 2000; Simpson 2001) and the use of the All Fields Postcode Directory to convert data between pairs of administrative, electoral, census and postal geographies. Ultimately, the choice of GCT or GML framework for converting data from a set of source to target units will depend on a variety of factors:

- The form in which a LUT is supplied by a third party.
- The application in which geographical conversion is required.

- The availability of software to use the lookup tables or user programming preference.
- Whether the user tends to conceptualise in terms of source to target links or in terms of a target geography and its constituent parts.

4.6 Deriving intersection weights for imperfect (dis)aggregations

For both GCT and GML approaches, since it is unlikely that direct measures of the extent of the overlap between source and target zonal systems will be available, surrogate measures about the distribution of the population within each intersection must be derived. The weighting criterion may be based on:

- Physical size. The overlap extents may calculated through GIS boundary intersections, approximations based on paper maps or, to reduce the affect of assuming equal population density, identifying intersections of residential areas using land-use maps. Unless digital boundaries are available this approach may be impractical for national/regional studies.
- Population counts. ONS geography (ONS 2000) track post-1991 ward boundary amendments and can supply (at cost) LUTs of changes in ED/ward relationships. Weights can be derived since the number of persons involved in the boundary change is noted in the file.
- Point counts. In the area of overlap these may be derived from counts of postcodes ideally in combination with counts of households, addresses or electoral register-derived information. Point counts have the advantage of national coverage, timeliness and versatility especially if OS Address-Point data are available.

- Grid-based counts. Surface models of population counts can be used to estimate populations in source/target intersections with the grid cells used as building-blocks (Bracken and Martin 1995).

5. CREATING A CONSISTENT GEOGRAPHY USING GEOGRAPHICAL CONVERSION TABLES

The aim then is to develop a method through which to standardise the input data needed for population estimates and SMRs for each year 1990-98 to be consistent with the 1998 ward geography for NHS Eastern Region. In this section the assumptions that need to be made about postcode locations and their use in geographical data conversion will be examined. Two methods will then be described that are underpinned by GCT framework LUTs with postcode-derived counts used to estimate the intersection weights between source and target geographies. Firstly, since it directly relates to their original purpose, an ‘ED Building-Brick’ method has been developed that uses the postcode LUTs to link data from later years back to the 1991 ED geography. Links are then needed between the EDs and the 1998 ward geography to fulfil the requirements of a relatively up to date geography. Secondly, to alleviate some problems found with the ED Building-Brick method, a ‘Postcode-Point’ method has been developed whereby the postcode locations are used. This approach builds on the UUKAM project’s work whereby data from more years than just 1991 can be directly adjusted to the 1998 wards.

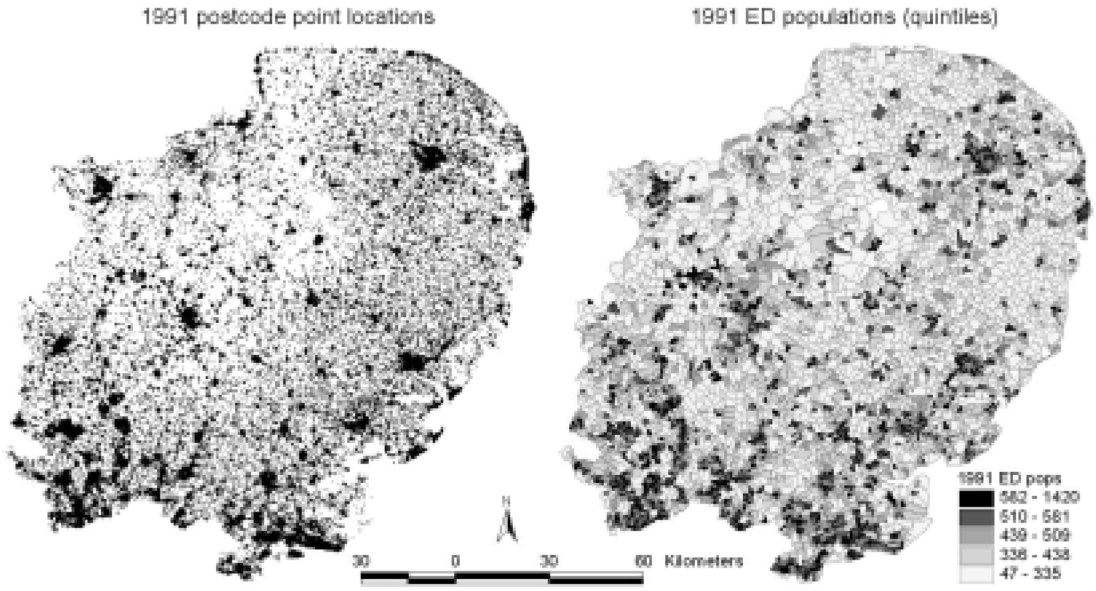
5.1 Postcode locations for geographical data conversion

As noted in the previous section, postcode counts can be used to derive measures of the intersection sizes between each source unit and target unit through which the original data may be converted from the ward geography for each year of the study period into the 1998 ward geography. As a geographical device postcodes are taken to be located in space as a point entity even though a postcode is applicable to a number of addresses (or a large multiple-user building) that theoretically could be digitised as a polygon. Postcodes are assigned the Ordnance Survey (OS) National Grid Reference (NGR) of the ‘first’ address in each postcode (ONS 2000) in the form of an Easting and a Northing (x and y coordinates) mainly to 100 metre resolution (MIMAS 1999). If postcode point locations and any information associated with them in supplied LUTs are to be used to link source and target geographies and to derive intersection weights, various assumptions must be made:

- Postcode distribution is a proxy for population distribution
- At a point in time, a set of postcodes can be taken to constitute a ward
- The termination and introduction of postcodes is a proxy for population change

Figure 6 shows a point map of the distribution of postcodes derived from the 1991 PCED, (MIMAS 1999) together with a map of EDs in Eastern Region showing 1991 Census populations. The more populous EDs are paralleled by the areas with the denser postcode distributions suggesting that postcode distribution is a sufficient proxy for population distribution.

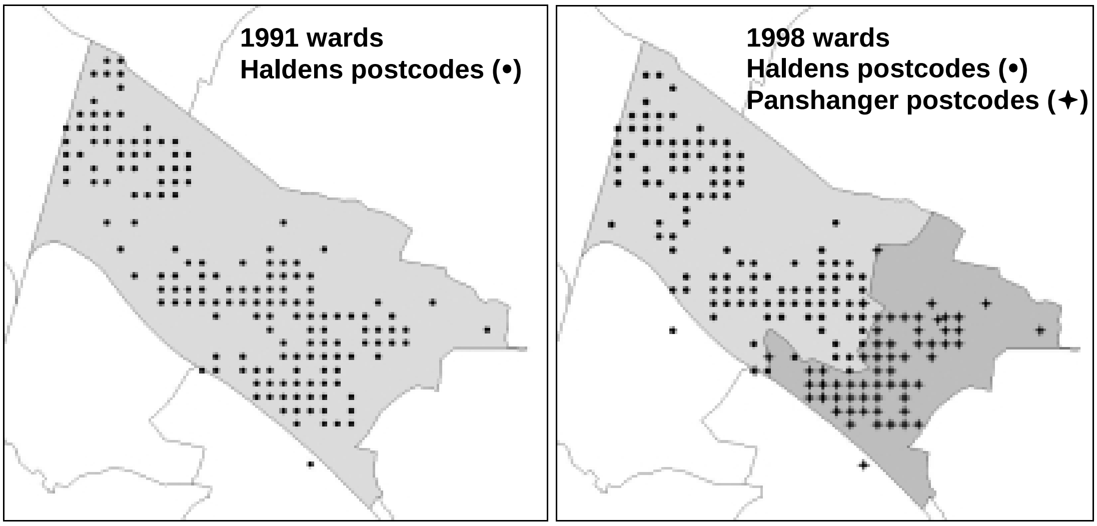
Figure 6: 1991 postcode point and ED population distributions in Eastern Region



For sources see Acknowledgements sections b, c and f

As noted in Figure 1, Haldens ward in Welwyn-Hatfield reduced in size between 1991 and 1998 with a new ward, Panshanger, being created. To illustrate the fit of a ward's constituent postcodes to the digital ward boundaries Figure 7 shows all the postcodes for Haldens ward in 1991 with the 1991 ward boundary and the postcodes for Haldens and Panshanger in 1998. The constituent postcodes for both wards in each year are, in the main, consistent with the digital ward boundaries. However, two issues arise relating to the 100m resolution of the postcode NGRs. Firstly a small number of postcodes are located outside the correct ward boundary and secondly, many postcodes which lie close together are given the same easting and northing coordinates so that as point locations they are stacked.

Figure 7: Constituent postcodes of wards in 1991 and 1998 (Haldens/Panshanger)



For sources see Acknowledgements sections b, f and g

Each quarter the Royal Mail's latest Postal Address File (PAF) is made available to ONS and provides the date of termination of postcodes no longer in use and the date of introduction of new postcodes. Two to three thousand existing postcodes are terminated each month and four to five thousand new postcodes are added (Yu and Simpson 2000). Although the majority of change is due to business 'large users', the termination and introduction of residential 'small user' postcodes can be taken as a proxy for population change at a sub-ward level because when housing demolitions occur postcodes are terminated, when newly built housing estates are constructed new postcodes are allocated and when infill development occurs, the number of delivery points is altered on the PAF (Penhale *et al.* 2000).

As previously noted, postcode-based LUTs are available to the academic community via MIMAS that enable linkage between one geographical area and one time and another (MIMAS 1999) and health professionals have access to similar information through the NHS Postcode Directory, maintained on behalf of the Department of

Health by ONS (ONS 2000). It should also be noted that many commercial organisations supply postcode location information (see Martin and Higgs 1997). The LUTs and the information relevant to the application are described below.

The Postcode-Enumeration District Directory (PCED) file provides a match between postcodes and EDs in England and Wales. For every unit postcode there are details of the EDs each postcode falls into, the number of resident households in each postcode-ED intersection and the Ordnance Survey (OS) National Grid Reference (NGR) (easting/northing). The directory was originally compiled with 1991 postcodes but has been updated for 1995, 1996 and 1997 (Martin 1992; MIMAS 1999b). The Central Postcode Directory (CPD) or Postzon file, originally created by the Department of Transport, consists of a single data record for each UK postcode. The record contains the NGR and local government and ward code for the first address in each postcode. Additional information includes the date of introduction or termination and the postcode user type (small or large). The CPD is available for 1991, 1993 and 1995 (Gatrell *et al.* 1991; MIMAS 1999b). The All Fields Postcode Directory (AFPD) is the most recent addition to the LUTs at MIMAS and is the most comprehensive and versatile. The AFPD is produced by ONS and combines data from the PCED and various other LUTs produced by ONS to link postcodes to a wide range of geographical units (MIMAS 1999b; Simpson and Yu 2000; Yu and Simpson 2000; Simpson 2001).

Reviewing the documentation shows the files contain similar but not consistent information. PCED and CPD files are both available in 1991 and 1995 but LUTs are unavailable for four years of the study period. All the LUTs are laid out in the GCT format as described previously. The PCED supports many to many links since postcodes can be linked to more than one 1991 ED with an indication of weighting given by the household count. The postcode-ward links given in the CPDs are on a one to one best-fit basis for the years the files are available and AFPD postcode links are also on a best-fit basis to the other geographies in the file. Table 3 summarises the information availability relevant to establishing a consistent geography on a ward basis for the study period.

Table 3: Availability of relevant lookup table information

Year	File	Postcode	NGR xy	Resolution	Ward of year	ED link	Hshld/Addr
1989							
1990							
1991	PCED			100m			
1991	CPD			100m		×	×
1992	×	×	×	×	×	×	×
1993	CPD			100m		×	×
1994	×	×	×	×	×	×	×
1995	CPD			100m		×	×
1995	PCED			100m	×		
1996	PCED			100m	×		
1997	PCED			100m	×		
1998	AFPD			100m			

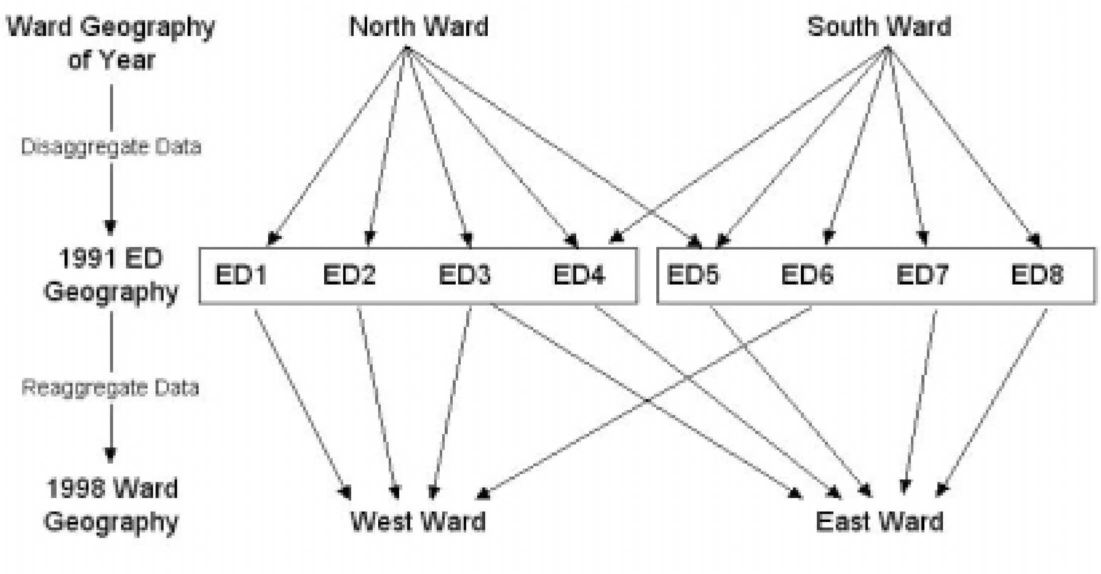
For sources see Acknowledgements section f

5.2 ED Building-Brick approach to data conversion

The PCEDs for 1991 and more recent years and the AFPD for 1998 link postcodes of the year of the file (the ‘year of interest’) back to the 1991 ED geography. The AFPD also links the EDs forward to the 1998 ward geography. These LUTs can therefore be used to develop a hybrid between the ‘freeze history’ and the ‘update to contemporary zones’ approaches to establishing a consistent geography. The principle of the approach being to use the constituent postcodes of a ward in a particular year and disaggregate ward data to the 1991 ED geography using a set of postcode-ED linked

GCTs and then to rebuild the data using another set of GCTs to the 1998 ward geography. This is illustrated in Figure 8 whereby in a particular year two hypothetical wards, North and South, are linked to various 1991 EDs. Whilst each ward entirely overlaps three EDs, ED4 and ED5 are intersected by both. By 1998 East and West wards have been created each overlapping various whole EDs but both intersecting ED3. Data from North and South wards would be disaggregated into the ED geography using postcode-derived weights to estimate each area of overlap and then reaggregated from EDs to the 1998 geography in a similar manner.

Figure 8: ED Building-Brick data disaggregation-reaggregation approach



It is necessary then to obtain data from the three sets of national LUTs available at MIMAS that will indicate the number of postcodes in each ward of the year of interest, a linkage to the 1991 ED geography and a count of households at each postcode. Since the population estimation and SMR input files relate to electoral wards, the consistency of the names, codes and file order of the wards from each data source must be checked and links with the 1991 Census EDs established. As noted in

Table 3, the three sets of national LUTs available at MIMAS contain similar but not consistent information. Data relating to the study region must be extracted from the national LUTs, area linkages must be established and any missing information estimated. The stages necessary to achieve this are discussed below.

Stage 1. Obtain national postcode LUTs. As previously noted, the PCEDs, CPDs and AFPD are available to the academic community via MIMAS. These were downloaded from the server once necessary permissions were set up (see MIMAS 1999 and Acknowledgements section f). The files held as a result are the PCED for 1991, 95, 96 and 97, the CPD for 1991, 93 and 95 and the AFPD for 1998 and will be referred to below using a prefix of the LUT name and a two digit suffix for the year (e.g. CPD95)

Stage 2. Obtain digital boundary data for the study region. ED polygon boundaries and boundaries for the six counties that comprise the region (Bedfordshire, Cambridgeshire, Essex, Hertfordshire, Norfolk and Suffolk) were downloaded from EDINA's UKBORDERS website (see UKBORDERS 2001 and Acknowledgements section b). The internal county boundaries were dissolved in the GIS package ArcView to create an external digital boundary for the whole of Eastern Region.

Stage 3. Create GIS point covers for all national LUTs. For each file, the OS NGRs for each postcode were used to locate each as a point in ArcView. The NGR ten digit easting/northing coordinates given for the postcodes in the national LUTs do not match the map unit measurements of the GIS region and ED polygon coverages. This is corrected by multiplying the postcode xy coordinates by 10.

Stage 4. Abstract LUT data for Eastern Region. Two approaches can be taken to select data for the study region from the national files. One is to select data records that match a given criteria and the other is to select point locations that fall within the external boundary of the region; both are possible using ArcView. The CPD and AFPD files include a reference to the county each postcode is associated with in the postcode point cover's attribute table (the GIS database of fields and records) and the relevant records can be selected from the table using logical queries. This approach is not possible with the PCED files as the geographical codes only relate to 1991 and not to later years. The postcode points that are located within the study region digital boundary can be selected from the national file using a GIS 'point in polygon' operation (a GIS geometrical calculation, see Martin 1996). GIS point in polygon does not give the same results as a database logical query as the postcode NGR 100m resolution locates some study region points outside and some unwanted postcodes inside the study region polygon (as previously noted in Figure 7). A comparison of the county lookups and point in polygon approaches carried out on the CPD95 file shows this applies to only a small number of points (112 out of 133271 postcodes) but is a potential source of error.

Stage 5. Delete invalid postcodes. Postcodes are deemed invalid to the application in various circumstances and are deleted from the files using logical queries on the attribute tables. Postcode records are deleted where the postcode is terminated prior to the year of interest and if the postcode is not yet introduced by the year (new postcodes for 1999 are given in the AFPD98). Postcodes are also deleted when designated as 'large users' since these are assumed to be business premises and it is the distribution of residential postcodes that is required.

Stage 6. Establish postcode to ward of the year of interest links for all LUTs. This information is needed to determine the postcode constituents of each ward in any year. On the PCED91, the CPD for all years and the AFPD98 each postcode has a link to the ward of the year of interest. However, the equivalent links are not given on the PCEDs for 1995, 96 and 97 so there is a need to estimate this information. For the PCED95 information can be obtained from the CPD95 but for the 1996 and 1997 PCEDs data from other years must be used. To do this the assumption needs to be made that if changes in sub-district ward structure have not occurred between years then the postcode-ward attribute information will be applicable. On this basis since the ward structure for 1995 is the same as in 1996, information from the CPD95 can be assigned to the PCED96 and similarly as 1997 has the same wards as 1998 information for the PCED97 can be obtained from the AFPD98.

This is achieved in two steps. Firstly, where postcodes are valid across the years and present in both files the ward codes are assigned where a match can be made. Secondly, to estimate the information not assignable through postcode record matching, GIS ‘point to nearest point’ spatial joins are used to transfer the ward attributes of the postcode points in the LUT of one year to the postcodes in another. This is achieved through a radial search from the coordinates of the postcodes in one GIS point cover to the postcode points in another. The attribute data is transferred from the nearest postcode point to the origin point of the search. The postcode record matching was carried out between the PCED95 and CPD95 and for the LUTs of adjacent years indicated above. Unmatched postcodes in the PCEDs were then allocated a ward code using the point to nearest point technique. The furthest distance that this process took place over was between postcode points 141 metres apart suggesting that large changes are unlikely to have occurred between the years where the ward structure remained the same and that, although minor boundary changes may be missed, the assumption of applicability of information between years is reasonable.

Stage 7. Establish postcode to ED links for all LUTs. This information is needed so that the postcodes from the LUTs for each year can be related back to the 1991 ED geography. ED lookups are given for all postcodes on the PCED and AFPD files but in the CPD93 and CPD95, links to the 1991 ED geography are not given. GIS point in polygon is used then to assign ED codes to the CPD postcode points. Since point in polygon can give different results due to the NGR resolution for consistency this was also carried out for all the PCED and AFPD files. On each file around twelve points were not allocated to an ED using point in polygon because the points were located in wide rivers. The original ED code was allocated for these where available, but for the CPD93 and 95 these points were manually relocated into the nearest ED polygon. The GIS-assigned ED codes were subsequently used for all postcode-ED linkages.

As a result of carrying out stages 1-7, the postcode-ward-ED links now contained in the LUTs for Eastern Region are summarised below in Table 4.

Table 4: Postcode-ward-ED linked lookup table variables

File	Postcode	Ward of year	1991 ED-link	Hshlds/Addr
PCED91				
CPD91				✗
CPD93				✗
CPD95				✗
PCED95				
PCED96				
PCED97				
AFPD98				

Based on sources referenced in Acknowledgements sections b, c, d, e & f

A data shortage not addressed above is the unavailability of postcode LUTs at MIMAS for 1989, 90, 92 and 94. In terms of the ward structure of the region, the numbers and names of wards are i) the same in 1989 and 1990 as in 1991, ii) the same structure in 1992 as in 1993 and iii) the same in 1994 as in 1995. Data from the years that are available will be used for the years where information is missing. Other unresolved inconsistencies between the LUTs are that i) for each postcode on the CPD files there are no counts of households or addresses and ii) that household counts (derived by ONS and included in the PCEDs) are a different entity to the residential address counts (supplied by the Royal Mail and included on the AFPD) since an address can contain multiple households. Whilst the household counts could be assigned using postcode matching and the point to nearest point technique described above it is a fine level of detail that would need to be assumed applicable across years, and given the different definitions for households and addresses, at this stage the decision was made to calculate the source to target geography intersection weights using just the postcode counts rather than postcode counts qualified by household/address counts.

5.2.1 Ward and ED lookup files

To enable matching the postcode LUT records to wards, a year by year ‘pedigree’ of the number, names and reference codes of each ward must be constructed. The ward names and codes from the 1991 Census populations files and the electoral and Vital Statistics (VS) files 1989-1998 have been combined along with the ward LUTs supplied with the CPD and AFPD so that comparisons can be made (for sources of these files see Acknowledgements sections c to f). This indicates when changes in numbers of wards occur during the study period and reveals that the different input data sources have inconsistent names and reference codes despite the definitive information being available in Statutory Instruments (HMSO 1987-2001) when wards are created or have boundary alterations.

Discrepancies in the numbers of wards exist between the electorates and VS. These have been made consistent after consultation with the local authorities concerned. In 1991 the census populations show 1182 wards whereas the 1991 electorates and VS have 1184. This relates to the Census Local Base Statistics (LBS) and data being suppressed in small wards for confidentiality reasons (Cole 1994). For the population estimates, rather than the Census LBS being used as a base population, the EWCPOP 1991 mid-year estimates will be used and the EWCPOPs are also two wards short of the other data sources. Investigation of all the ward files reveals a wide variety of codings, conventions and spellings confirming the ONS (2000) view that there is a need to be consistent in the use of names, codes and references. Creating a ward pedigree for the study period is a time-consuming clerical task. A summary of the ward numbers is given in Table 5.

Table 5: Ward numbers in the 1989-98 ward pedigree

Year	1989 to 1991	1992 to 1996	1997 to 1998
Number of wards	1184	1185	1192

For sources see Acknowledgements sections c, d & e

In addition to ward pedigree files, an ED lookup file is needed. Since the point-in-polygon technique has been used, a polygon-based ED list is needed and this can be obtained from the attribute table of the GIS ED coverage. A list of ED codes derived in this way shows 11155 polygons with 11082 unique codes attached. There are multiple polygons as some EDs are split by rivers and some include islands along the region's coastline. The GIS attribute table forms the basis of an ED lookup table for matching records to the postcode LUTs.

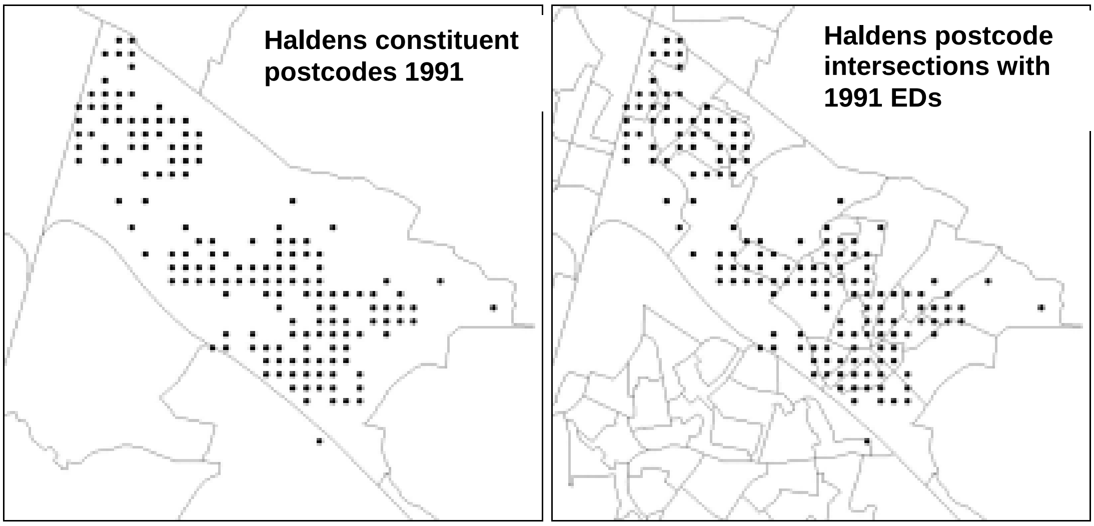
5.2.2 Calculating GCT weights for data disaggregation-reaggregation

The information in the postcode-ward-ED LUTs is to be used to calculate overlap weights for the intersections between the source and target geographies so that GCTs can be used to adjust the ward input data for the years 1989-98 to the 1998 geography. The hybrid freeze history and update to contemporary zones approach means that effectively the source to target adjustments are carried out twice. First, the source geography is the ward geography of the year of interest and the 1991 EDs are an interim target geography and second, the EDs become the source geography and the 1998 wards the target.

Three algorithms are outlined in Figures 9 to 11. The first algorithm shows how the weights are calculated for the intersections between the source geography wards for the year of interest and the target 1991 ED geography. The second algorithm illustrates how weights are calculated to indicate the intersections between the source 1991 EDs with the target 1998 ward geography is essence reverse engineering the first algorithm. The third algorithm outlines how GCTs will be used to disaggregate the source data for each year of interest to create estimates in the 1991 ED geography and then how a single GCT will be applied to rebuild the ED estimates for each year into the 1998 ward geography.

Figure 9: Algorithm to calculate 'freeze history' source-target GCT1 weights

Step	Procedure
1	Upload the LUT for Eastern Region for the year of interest (e.g. PCED91)
2	Count the number of constituent postcodes in each source ward in the year of interest (e.g. postcodes in Haldens ward 1991, below left)
3	Count postcode intersections with 1991 ED target geography (e.g. Haldens postcodes 1991 with 1991 EDs, below right)
4	Divide postcode-ED intersection count by ward constituent postcode count
5	The result is the source-target geography disaggregation weight, the proportion of the source geography ward of the year of interest intersecting with each 1991 ED, the interim target geography


Haldens constituent postcodes 1991
Haldens postcode intersections with 1991 EDs

For sources see Acknowledgements sections b, c, d, e & f

Figure 10: Algorithm to calculate 'update to contemporary zones' source-target GCT2 weights

Step	Procedure
1	Apportion sample data from the target 1998 geography to 1991 ED geography using GCT disaggregation weights calculated above
2	Sum the disaggregated target geography estimated in 1991 ED geography
3	Divide the GCT disaggregated target ward-source ED geography intersections by the total ED estimate
4	The result is the proportion of each source ED intersecting with a target 1998 ward (which will take a value of more than zero but less than or equal to one)

Figure 11: Algorithm to use GCTs to carry out data disaggregation-reaggregation process

Step	Procedure
1	Upload source geography ward data for year of interest (e.g. deaths data for 1991 Haldens ward)
2	Use GCT1 to estimate source to 1991 ED interim target geography intersections by multiplying ward data by GCT weights (e.g. deaths data for 1991 Haldens ward disaggregated into intersection counts)
3	Aggregate intersection counts into interim target ED totals (e.g. deaths data for 1991 Haldens ward estimated for 1991 ED geography)
4	Use GCT2 to apportion ED geography to target 1998 geography using source-target weights (e.g. deaths data for 1991 Haldens ward estimated for 1991 ED geography is reaggregated to 1998 geography Haldens and Panshanger wards)

5.2.3 ED Building-Brick derived consistent geography: assessing the output

The algorithms as outlined above have been programmed in FORTRAN 90 to produce GCTs for 1991, 1993 and 1995 to 1998, the years for which the postcode-ward-ED LUTs have been assembled. To test the programs and assess their output hypothetical ward populations have been used as input data. The total study region population has been set at 1,192,000 divided equally in the source geographies between the 1,184 wards in 1990 and 1991, the 1,185 wards in 1992-96 and the 1,192 wards in 1997 and 1998, giving 1,000 persons per ward in the target geography.

The hypothetical ward data have been disaggregated and reaggregated from source to target geographies to be consistent with 1998 geography using the GCTs. Since there are PCED and CPD files duplicating 1991 and 1995 and some years missing (1989, 90, 92 and 94), there are eight output files relating to six out of the nine years of the study period. For the study region as a whole in every year (i.e. the years the procedure has been programmed) the total population is disaggregated and reaggregated back to within a few decimal places so that no whole persons are lost from the system. The ward geography for 1998 disaggregates and reaggregates almost perfectly but this is to be expected as the reaggregation weights are based on its own disaggregation weights.

The program outputs can be assessed through graphs to show the size of each ward population once adjusted to 1998 geography and choropleth maps to identify the locations where large differences have been found to occur for the interim ED geography and for the 1998 target geography. Any differences identified between source to target GCT adjusted data can be due to genuine changes in ward size and structure, incorrect LUT links between geographical areas or incorrect assumptions, method or programming.

Figure 12 shows the output of each of the programs with the hypothetical populations adjusted to the 1998 geography. If the GCTs are estimating the geographical changes correctly then any counts of less than 1,000 on the graph represent wards that have reduced in size by 1998 and wards that have increased for counts of over 1,000. Whilst there is much clustering around 1,000 persons per ward there are many large deviations that are well in excess of differences likely to be due to boundary changes.

Figure 12: Adjustment of hypothetical populations to 1998 geography, ED version 1

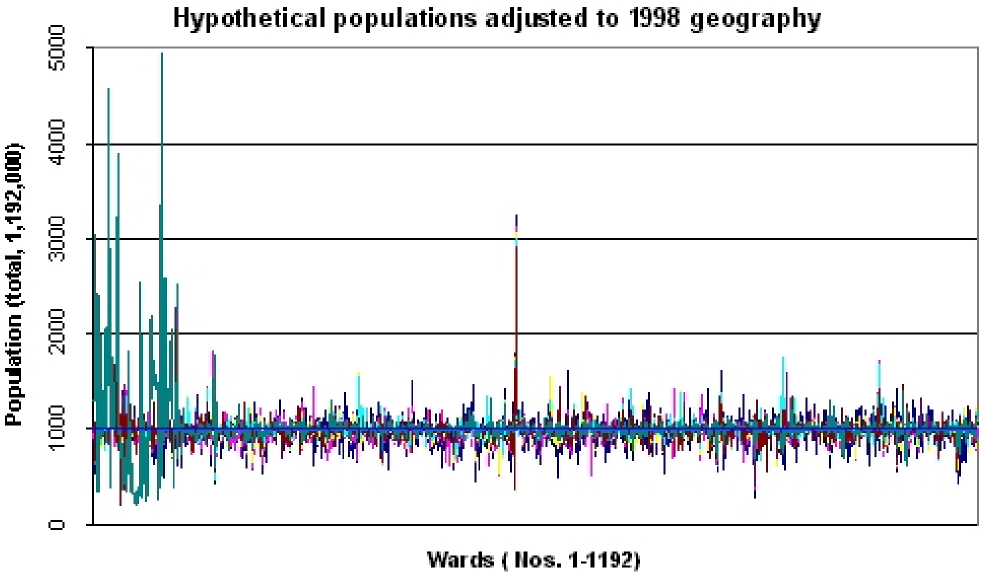
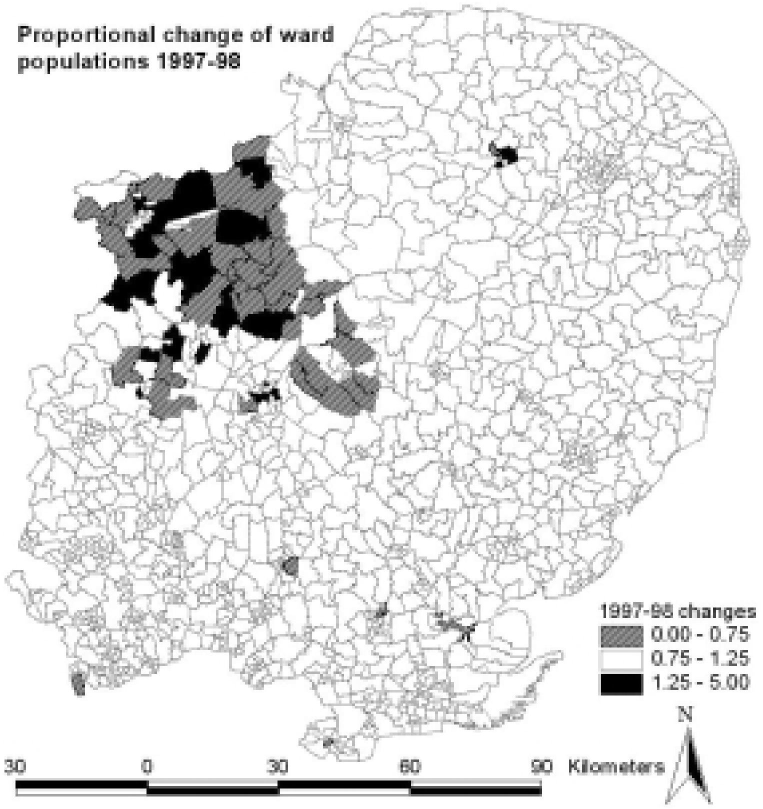


Figure 13: Proportional change of ward populations 1997-98

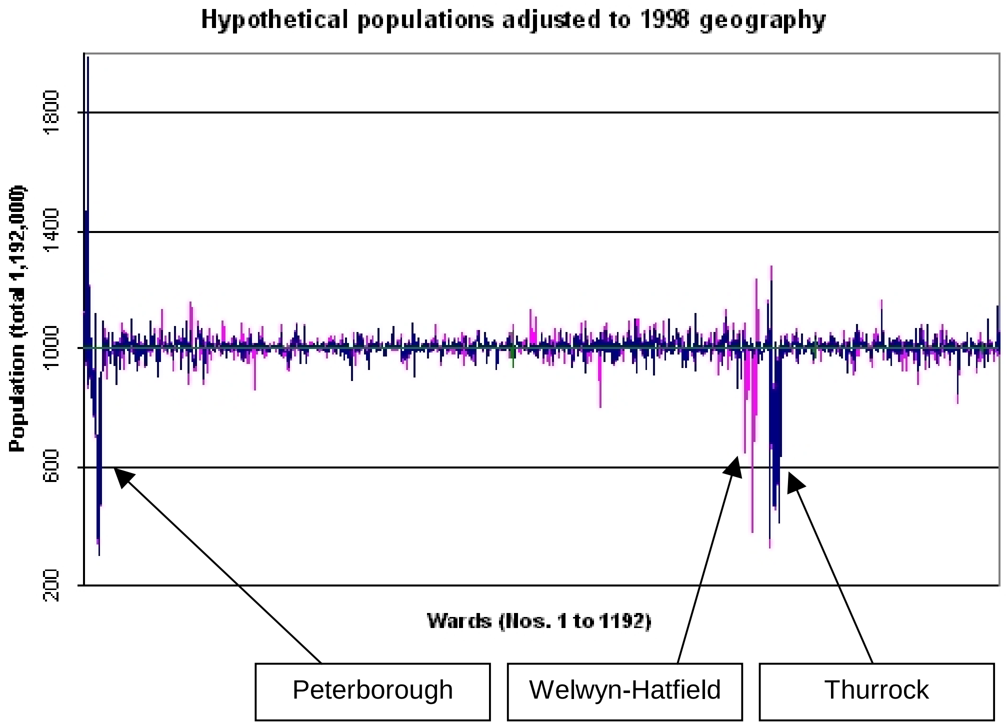


For sources see Acknowledgements section g

To identify locations where excessive year by year adjustments have occurred, the population counts adjusted to 1998 geography in one year were divided by the count for the next. The results of these calculations were choropleth mapped using the 1998 ward boundaries with the largest differences observed between 1997 and 1998 especially in the north-west of the region (see Figure 13). This is unexpected because the ward structure did not change between these years suggesting, together with other large year to year changes, that there are some problems in the geographical links in the GCTs.

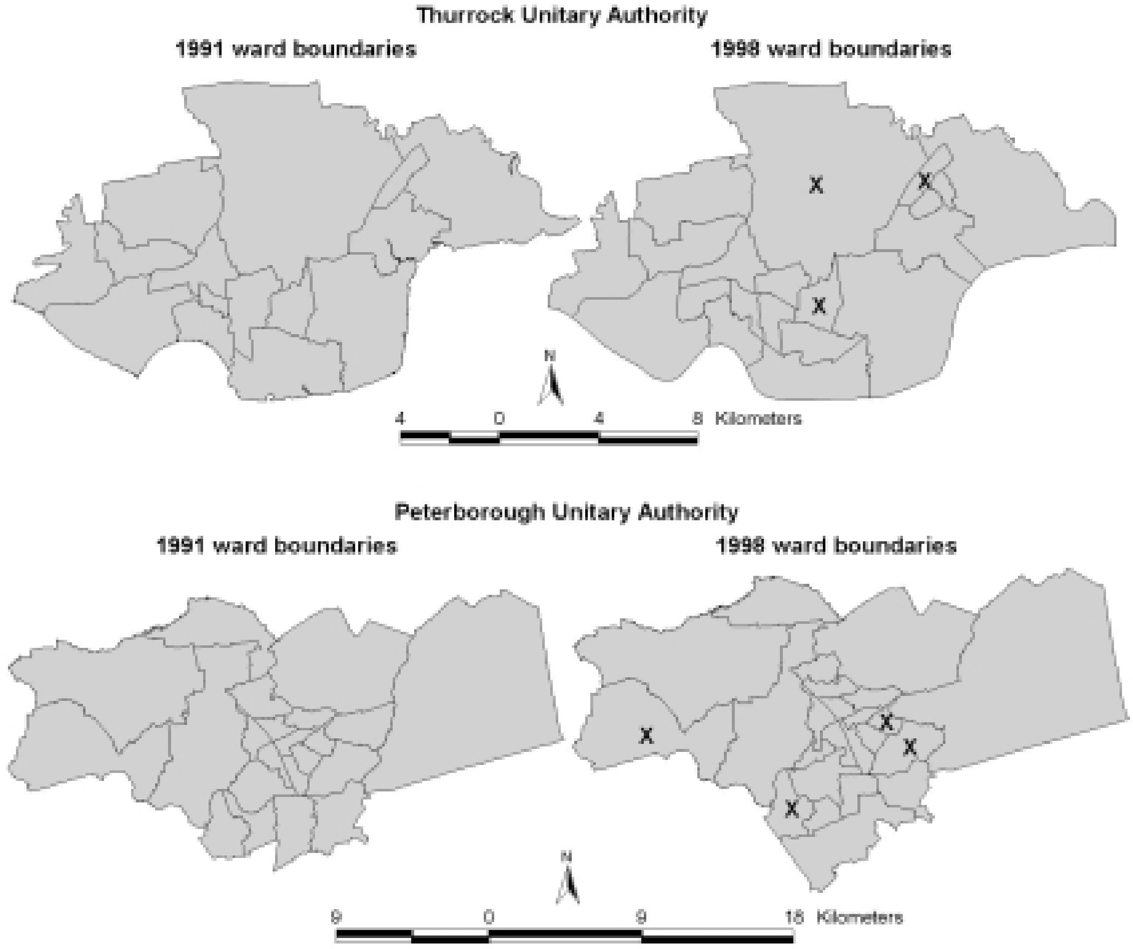
Investigation of the GCTs and input files showed some differences in ward code conventions that had not previously been identified. The 1998 ward codes in the AFPD vary between six and four figure alphanumeric codes whereas the elector and VS files use six figure throughout. These were rationalised to give consistency between the postcode-ward-ED LUTs and the ward data input files. The programs to calculate the GCTs were re-run and the hypothetical population counts adjusted to the 1998 geography using the new GCTs. For 1991 and 1995, since LUT information is available from both the PCED and CPD an average of the data estimated to 1998 geography is used.

Figure 14: Adjustment of hypothetical populations to 1998 geography, ED version 2



The output in Figure 14 has a more marked clustering for each ward around 1,000 persons than was previously shown. There are still wards with populations well above and below 1,000 and these should represent wards that have respectively increased or decreased in size between source and target ward geographies. The locations of the largest adjustments have been identified to determine whether these represent substantial changes in the ward boundaries over time or error. For wards in the local authority district Welwyn-Hatfield, previously noted as experiencing boundary changes during 1991 (see Figure 1), bars on the graph above the 1000 person level include Welwyn South and Brookmans Park and Little Heath, both of which have larger areal extents than the previous boundary definitions. Similarly wards which reduced in size over time, Welwyn North, Haldens and Hatfield East are represented on the graph by bars below the 1000 person level. Other locations with large changes in hypothetical populations are Peterborough and Thurrock which both became Unitary Authorities by 1998 experiencing substantial changes to their constituent wards. These are illustrated in Figure 15, wards with no boundary changes marked with an X.

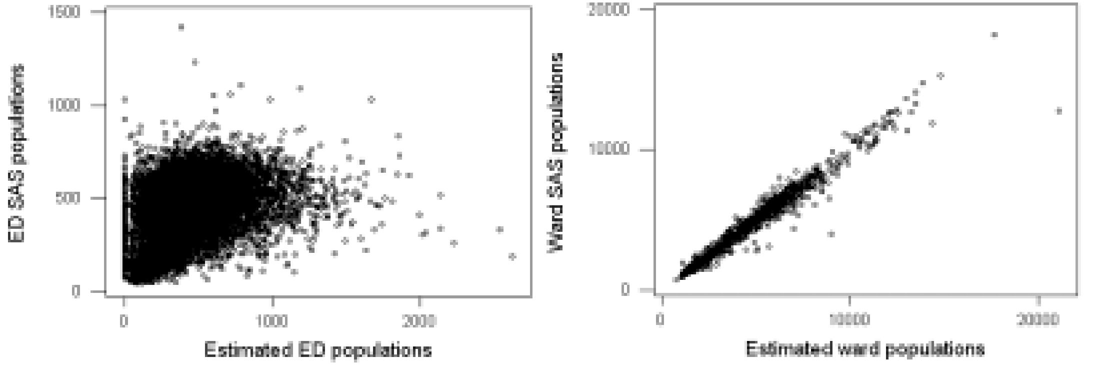
Figure 15: Ward boundaries 1991 and 1998, Thurrock and Peterborough Unitary Authorities



For sources see Acknowledgements section b & g

As previously noted, many of the larger year by year variations identified are in locations where ward boundaries known to have changed. However, variation will be due to aspects of the postcode-ward-ED linkage approach that relate to the postcode NGR resolution and point in polygon-derived links. Further checks on the method can be carried out at the interim target geography stage to assess how well ED populations are estimated. To do this, ED populations from the 1991 Census SAS can be aggregated into the wards those EDs comprise. These ward populations can then be adjusted to the 1998 geography with the interim ED estimates output so that they can be checked against the original SAS ED populations.

Figure 16 Estimated and original ED and ward populations



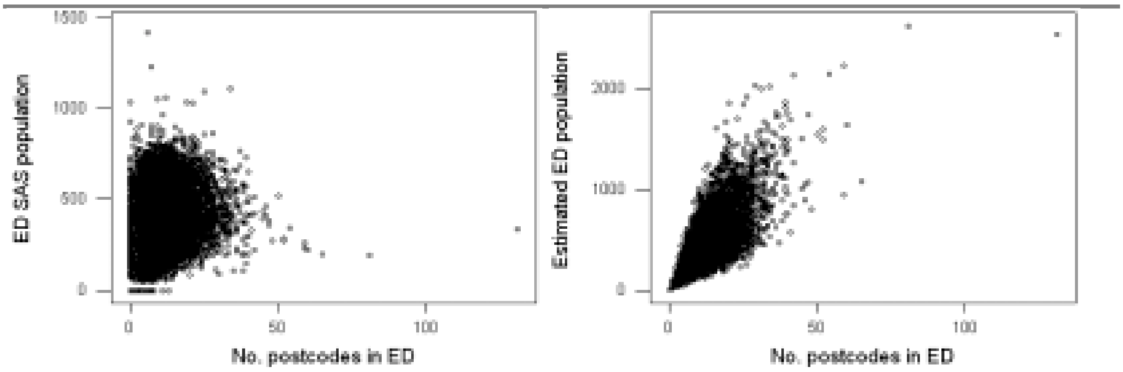
The scatterplot on the left in Figure 16 shows the estimated ED populations plotted against the original ED populations from the SAS. Although statistically significant (p-value 0.000), there is a correlation of just 0.347 between the estimated and original populations. When these estimated ED populations are aggregated into 1991 wards the result is a closer match between the estimated and original ward populations with a correlation of 0.982 (p-value 0.000). This suggests that whilst the method is not a good estimator of ED populations, ward-level populations are estimated well. However, if populations are incorrectly allocated to EDs, difficulties may occur when part EDs are apportioned to the 1998 geography.

Various problems relating to the postcode resolution and point in polygon technique will lead to poor estimates of ED populations. As previously noted a postcode can be allocated to an incorrect ED because the NGR locates it the wrong side of a boundary. The stacking of postcode coordinates can lead to some EDs having an over-allocation of population and some EDs are completely missed having no postcode intersections and therefore no estimated population. The left-hand scatterplot in Figure 16 showing the latter situation as a vertical line of points above zero on the x-axis.

Highly populous urban wards have more EDs, each being relatively small in areal extent whereas less populous rural wards have fewer EDs, each being larger in size. This means that if a misallocation of a postcode occurs through intersecting with the wrong ED, stacking in an ED or missing EDs completely, this is more likely to happen in urban areas as there is more margin for error in rural areas. The greatest amount of population change will occur in urban areas and since population change leads changes in electoral ward boundaries, boundary changes are likely to be more frequent. Incorrect postcode-ED point in polygon-derived links will lead to poor urban ED and 1998 ward estimates.

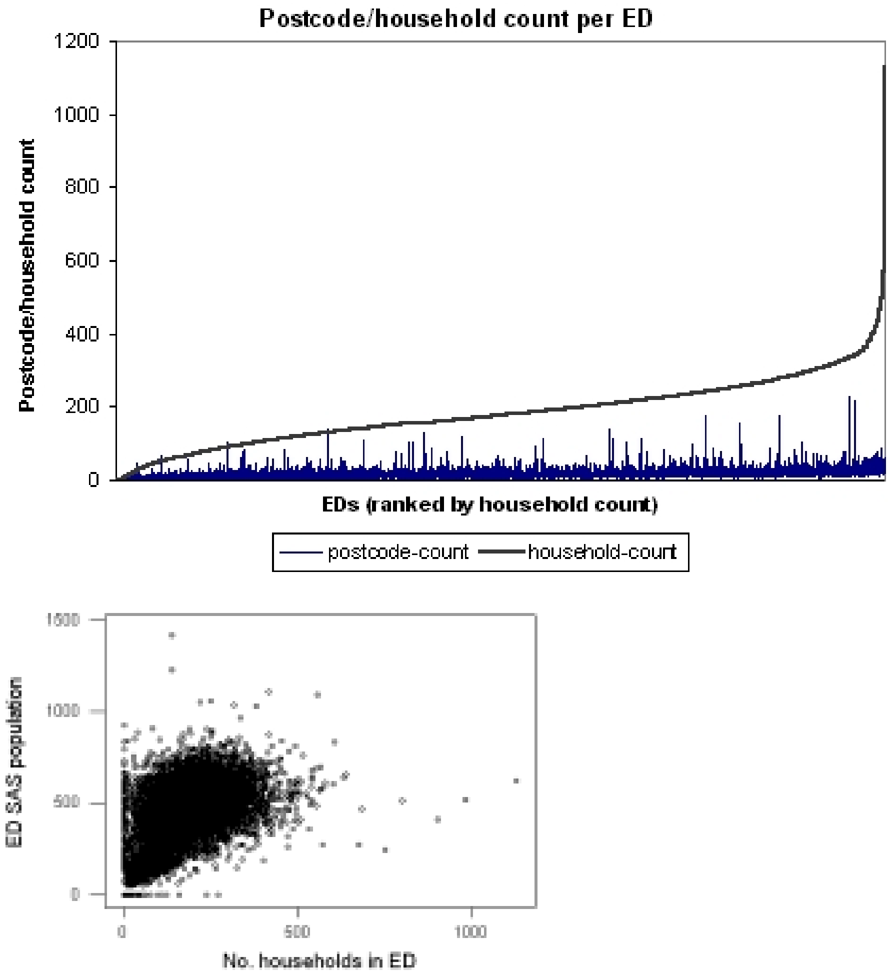
The distribution of postcodes is taken to be a proxy for population distribution. Whilst the greatest density of postcodes occurs in urban areas, since postcodes are being associated with EDs, the urban-rural gradient of the areal extent and number of EDs per ward may affect the urban-rural density gradient of postcode locations. Figure 17 shows a scatterplot of the number of postcodes per ED plotted against the ED population which, with a correlation of 0.055 (p-value 0.000), suggests that there are very similar numbers of persons per ED and postcodes per ED whether in urban or rural areas. A second scatterplot shows the number of postcodes in each ED and the estimated ED populations which with a correlation of 0.752 (p-value 0.000) demonstrates that the method is artificially creating a strong positive relationship between number of postcodes in an ED and the size of the population that does not exist in the original data.

Figure 17: Estimated and original ED populations and postcode counts



When a population is estimated for an ED using the number of postcode intersections, equal population density across that ED is assumed, a common issue for the interpretation of choropleth maps (Monmonier 1991). In reality, new housing estates and infill developments will tend to be concentrated in one part of an ED. Any new postcodes created will correctly lead to an increased proportion of the input ward data being allocated to that ED (assuming a correct point in polygon link) but the data will be assigned to the whole ED. This will not be a problem if the entire ED ‘belongs’ to a 1998 ward, but if an ED is divided with the creation of a new ward poor estimates will result for the wards that receive data from the ED. This problem is as applicable to rural and edge of town locations as for urban areas since any new housing development will tend to be concentrated in one location and any assumption of equal population density across the relatively large extents of sparsely populated rural EDs will be erroneous.

Figure 18: Postcode and household counts for each ED



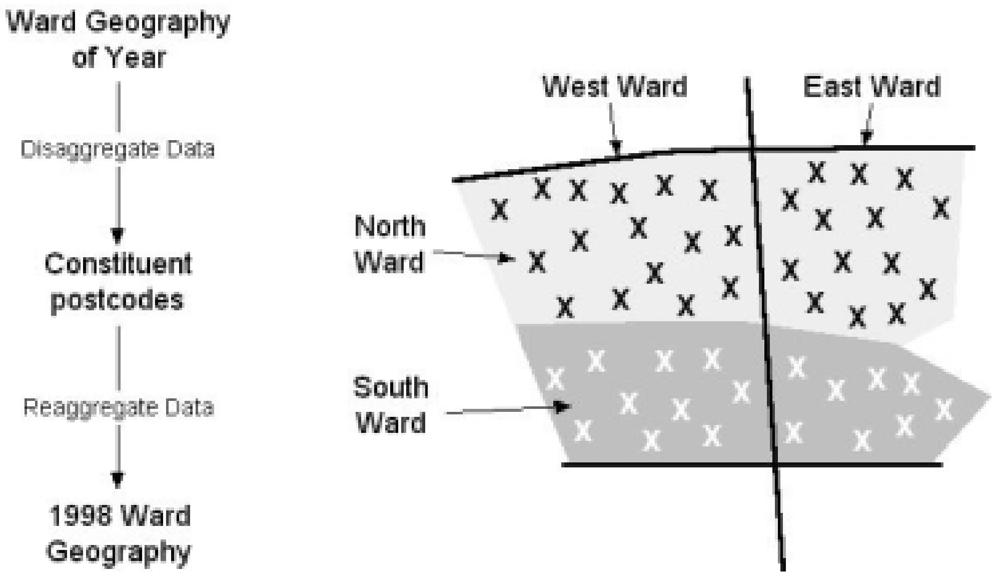
The greater frequency of postcodes (and of stacked postcodes) occurs in urban areas thereby concentrating the estimated populations in the more urban EDs. However, if urban postcodes contain more households than those in rural areas, the omission in this method of household counts for each postcode could mean that population is under-estimated in some EDs. Figure 19 first shows a comparison of postcode and household counts in each ED As previously noted there is not a decrease in the number of postcodes per ED with increased rurality, but the number of households per postcode increases in the more urban EDs. The scatterplot in Figure 18 shows a stronger relationship (correlation 0.468, p-value 0.000) with the number of households per ED and the size of the SAS population compared with the relationship between number of postcodes and ED populations noted above. Without the inclusion of a count of households to qualify the postcode derived source to target geography intersection weights, population in effect will have been spread too evenly across a ward's composite EDs.

A further problem is caused by an inconsistent use of techniques through which the linkages between geographical areas have been established. The freeze-history ward to ED links have been made using the GIS point in polygon technique but the links to the 1998 wards use the lookups given in the AFPD. The two approaches will be subject to different sources of error, the point in polygon through the problems discussed above and the AFPD through incorrect allocation of the original file information (for a discussion of AFPD validation see Yu and Simpson 2000). It would be more consistent either: i) to have used the postcode to ED links given in the original PCEDs and AFPD and to derive the missing ED links for the CPDs from these or; ii) to have used point in polygon for both the postcode-ED links and also to establish postcode to 1998 ward links using point in polygon with the 1998 digital ward boundaries although this use of the 1998 boundaries is outside the aims of this project.

5.3 Postcode-Point approach to data conversion

To alleviate the issues noted above relating to postcode-ED density relationships, the mixed technique two-step disaggregation-reaggregation approach and the lack of the use of household counts, a 'Postcode-Point' approach has been developed. The principle of this approach is that for each ward in the year of interest, data are disaggregated using GCTs into the constituent postcodes of the ward and then rebuilt to the contemporary zones using the constituent postcodes of the 1998 wards. Figure 20 illustrates this with data for North ward (black crosses) and South ward (white crosses) disaggregated to their constituent postcodes. The data are then reaggregated into the target geography using the constituent postcodes of West ward and of East ward (postcodes to the left and right of the diagram respectively). To aid clarity, boundaries are shown in Figure 19 but are not used in the method. This approach is feasible since postcode LUTs at MIMAS all give OS NGR eastings and northings for each postcode and the majority of postcodes are current throughout the study period and have the same NGRs so that geographical information for each year of interest can be related to 1998 using the same postcode. Postcodes that are introduced or terminated can have locational information assigned from the nearest postcode in space from the file of one year to file of another.

Figure 19: Postcode-Point data disaggregation-reaggregation approach



Underpinning the method are the same LUTs used for the ED Building-Brick approach and as before there is a need to estimate missing information. The aim is to construct GCTs that contain a postcode link between source ward geography for the year of interest and the target 1998 ward geography using the count of households at every residential postcode to estimate the area of overlap between source and target geographies. To construct GCTs the method is the same as for the ED Building-Brick approach up to *Stage 6* with data for Eastern Region abstracted from the national LUTs, invalid postcodes deleted and the information about the constituent postcodes of wards in each year of interest established.

Replacing *Stage 7,* to create a direct link between the postcode of the year of interest to the 1998 ward geography a combination of postcode matching and the GIS point to nearest point technique are used. For each of the PCEDs and CPDs where postcodes are valid between the year of the file and 1998, the ward codes from the AFPD are assigned to the PCED or CPD postcode wherever a match can be made. To estimate the remainder of the links point to nearest point is used in ArcView whereby for any postcode on the PCEDs or CPDs without a ward code this is assigned from the nearest postcode on the AFPD.

Household counts were not used in the ED Building-Brick method. These are included on the PCEDs for 1991 and 1995 to 1997 and address counts are given in the AFPD, the only year without this information is 1993 (where LUTs are available). Household counts are estimated for the CPD93 using the PCED95, the closest LUT in time with the same ward structure. As above, postcode record matching is first used where these are valid across the years and where matches cannot be made, the household counts are assigned using GIS point to nearest point from PCED95 to CPD93. The CPDs for 1991 and 1995 do not contain household counts, but as they pre-exist on the PCEDs, the CPDs for these years will not be used. Any issues of difference in entity definition between households and addresses are irrelevant as address counts are only given in the target year. The information necessary to calculate GCT source–target overlap weights for the Postcode-Point method in Eastern Region is summarised in Table 6.

Table 6: Ward of year-postcode-ward 1998 linked lookup table variables

File	Ward of year	Postcode	Ward 1998	Hshlds/Addr
PCED91				
CPD93				
PCED95				
PCED96				
PCED97				
AFPD98				

Based on sources referenced in Acknowledgements sections b, c, d, e & f

5.3.1 Calculating GCT weights for data disaggregation-reaggregation

The information in the ward of year-postcode-ward 98 LUTs is to be used to calculate overlap weights for the intersections between the source and target geographies. The principle of this approach is that for each ward in the year of interest data are disaggregated using GCTs into the constituent postcodes of the ward weighted by the household count for each postcode and then rebuilt to the contemporary zones using the constituent postcodes of the 1998 wards. The algorithm to achieve this is given in Figure 20.

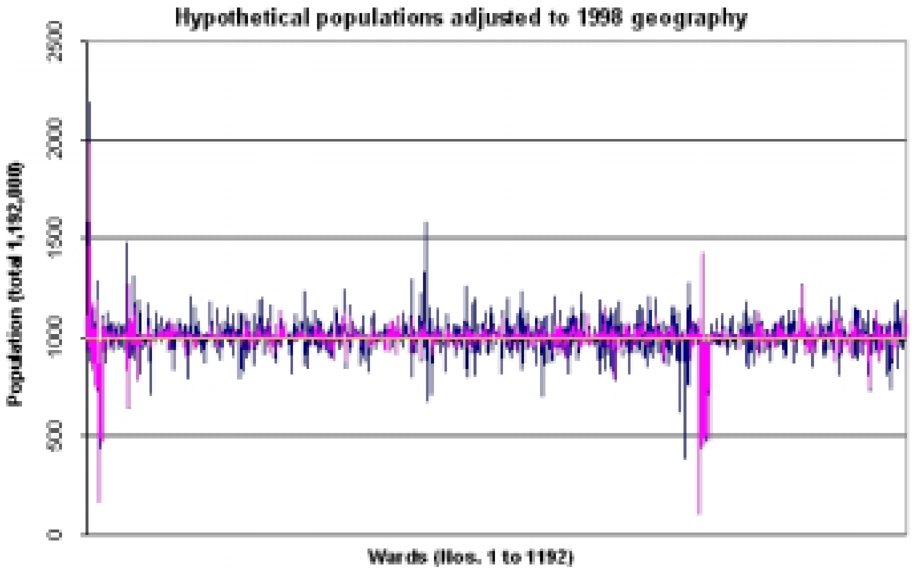
Figure 20: Algorithm to calculate ward of year-postcode-ward 1998 GCT weights and adjust input data to target geography

Step	Procedure
1	Read in LUT with postcode, household count, source reference and target reference
2	Sum the postcode household counts for each source zone
3	Calculate each postcode's disaggregation weight = (household count for each postcode / sum of households for the source zone)
4	Read in source zone population data
5	Disaggregate source zone population data = (source zone data * each postcode's disaggregation weight)
6	Sum disaggregated data to target zones using postcode link

5.3.2 Consistent geography derived using the Postcode-Point approach: assessing the output

The algorithm as outlined above has been programmed in FORTRAN 90 and as before a hypothetical population of 1,192,000 equally divided amongst the source geography wards has been adjusted to the 1998 geography. Figure 21 shows the program output for the years that the LUTs are available for. There is a strong clustering around 1,000 persons per ward, with large differences in areas where substantial boundary changes and the creation and abolition of wards is known to have taken place.

Figure 21: Adjustment of hypothetical populations to 1998 geography, Postcode-Point household count approach



Compared with the ED Building-Brick version the inclusion of the household counts in the disaggregation weights concentrates populations in the more urban areas. To investigate the effect of the inclusion of household counts further the Postcode-Point program was re-run with the hypothetical populations adjusted to 1998 wards using just postcode counts to calculate the disaggregation weights. The program output show somewhat less variation above and below 1,000 persons per ward (see Figure 22) with the household count approach allocating a larger proportion of the populations to the more urban wards and the postcode count method allocating more population to larger suburban and edge of town wards. Figure 23 shows that, whilst for most wards the differences in estimates are relatively small, the largest differences occur in locations which have experienced substantial ward boundary changes, Peterborough and Thurrock This underlines the need to include all available information to indicate the location and extent of change.

Figure 22: Adjustment of hypothetical populations to 1998 geography, Postcode-Point postcode count approach

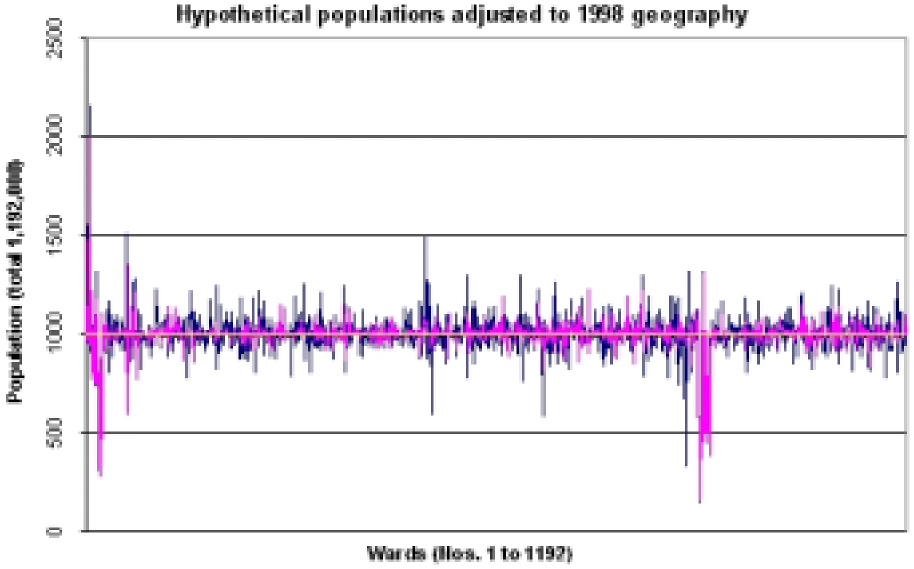
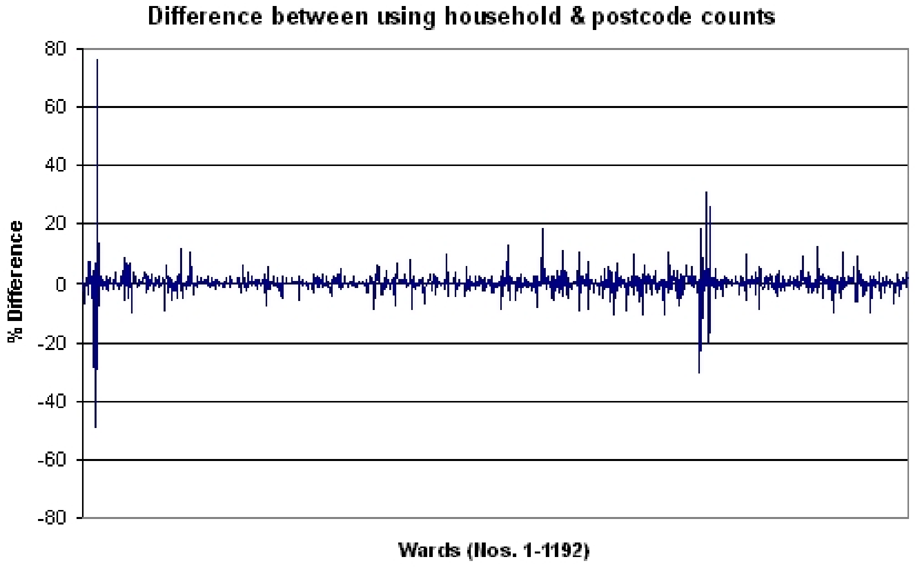
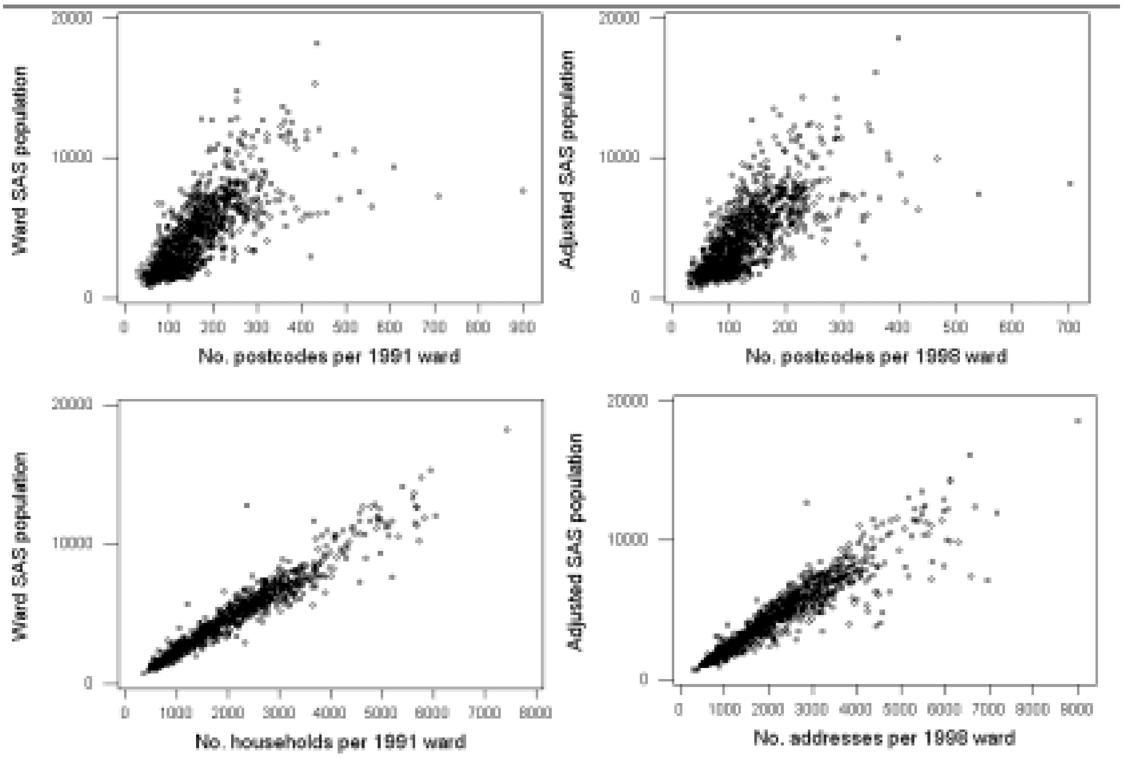


Figure 23: Percentage differences between the use of household and postcode counts



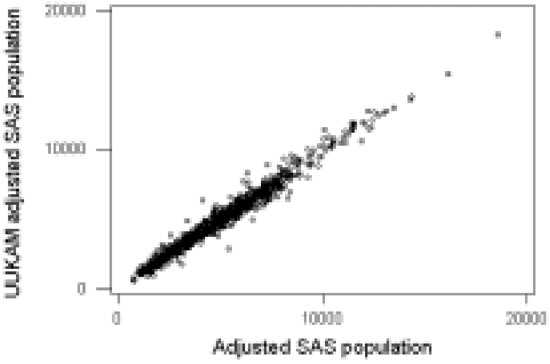
A problem with the ED Building-Brick method was an artificially created density relationship between EDs, postcode locations and the estimated ED populations. For the Postcode-Point approach it is possible to check whether the relationships between postcode and household counts with the source geography are maintained when data are adjusted to the 1998 geography. Figure 24 shows scatterplots of postcode counts for each 1991 ward against the ward SAS populations together with postcode counts for each 1998 ward against the SAS populations adjusted to the 1998 geography. The consistency of the plots and correlations of 0.741 for 1991 geography and 0.720 for 1998 geography (p-value 0.000 for both) suggests that the relationship between postcode counts and the ward-level SAS population is maintained. Similarly, Figure 24 shows scatterplots of household counts for 1991 wards and the SAS populations and the address counts for 1998 and the SAS populations estimated for the 1998 wards. A much stronger relationship is illustrated in the plots between both household and address counts with the original and adjusted SAS populations and this is confirmed through correlations of 0.971 and 0.943 for 1991 and 1998 respectively (p-value 0.000 for both) demonstrating that the relationships between ward-level household and address counts and populations are not artificially altered.

Figure 24: Relationships between ward-level postcode and household counts with population



The interim target geography was used to check the performance of the ED Building-Brick approach. This is not possible with the Postcode-Point approach but the program output can be compared with another source as the UUKAM offers online conversion between various pairs of geographical zones. Since these include conversion from 1991 Census wards to 1998 wards the 1991 Census SAS populations for wards in Eastern Region were converted using the project’s website (http://convert.mimas.ac.uk/). Figure 25 shows the output from Postcode-Point approach plotted against the output from the UUKAM project which with a correlation of 0.990 (p-value 0.000) confirms a high degree of match.

Figure 25: Comparison of postcode point and UUKAM estimated SAS populations



This similarity is to be expected since both the Postcode-Point and UUKAM approaches use LUT information from the AFPD and a similar methodology to derive the source-target intersection weights in the GCTs. However, variation between the outputs is inevitable for several reasons:

- The Postcode-Point approach uses household counts from the year of interest, in this case the PCED91, to calculate the intersection weights whereas the UUKAM project uses 1998 residential address counts from the AFPD (Simpson 2001).
- The PCED91 provides the number of resident households in each intersection of postcodes and EDs and can apportion data to different wards where a postcode crosses a ward boundary but the AFPD allocates postcodes to wards on a best fit basis. However, error due to the use of whole rather than part postcodes in constructing GCTs has been found to exist but is not great (Simpson 2001).
- The difference in definition between a household and an address since an address can contain multiple households.
- The links between the source and target geographies in the Postcode-Point approach have been estimated and allocating attribute data from the nearest postcode may not always be correct.

- The region of interest for the Postcode-Point approach is fixed for Eastern Region whereas the UUKAM project uses the national AFPD. Whilst no persons are lost during the data conversion process, i.e. the GCTs in both are "exhaustive" (Simpson 2001 p. 4), a small number of persons (0.03%) are allocated by the UUKAM to 1998 wards outside the study region.

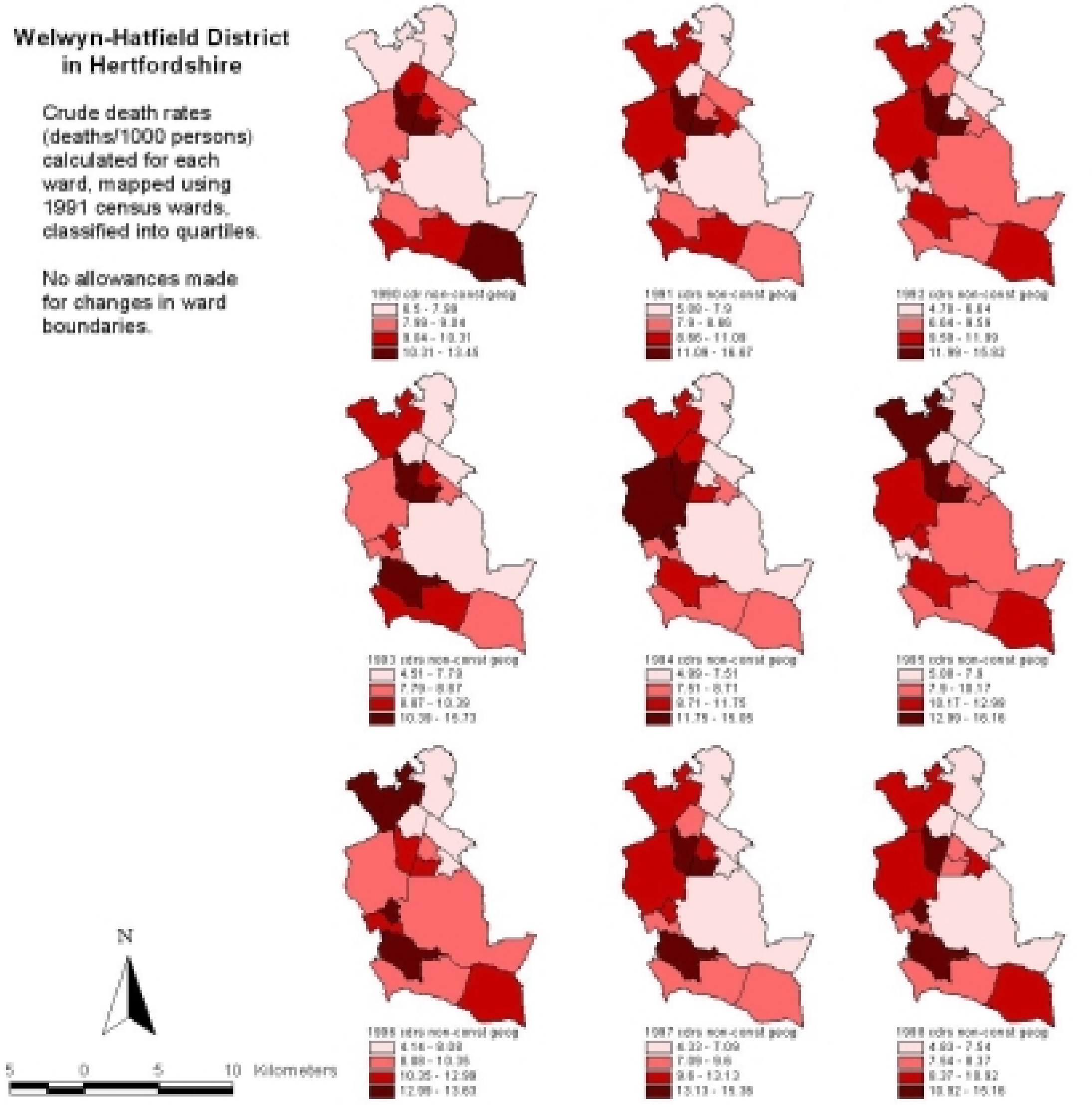
Compared with the ED Building-Brick method issues about the 100m NGR resolution location of the postcodes are less critical because no error is introduced by incorrect point in polygon associations with the ED digital boundaries and the use of postcode derived weights that are independent of the ED geography will be more versatile as postcodes are a smaller building-brick. Problems of mixed methods through GIS point in polygon and given LUT geographical area linkages are negated by just using the given linkages. Validation checks are less easy than those carried out on the ED Building-Brick method, but the Postcode-Point approach performs well in comparison with the UUKAM project and justifies its use as the method by which to adjust all the input data for the population estimates and SMRs to be consistent with the 1998 ward geography.

7. A CONSISTENT GEOGRAPHY FOR WELWYN-HATFIELD: WORKED EXAMPLE

Academic, local authority and health authority researchers have free access to digital boundaries for the 1991 Census wards and can attach population and other variables to those boundaries using GIS join item procedures (see Table 1). If, for example, a researcher wished to calculate crude death rates (acknowledging this is not an ideal measure) for NHS Eastern Region to investigate mortality for the period 1990-98 the input data needed are the deaths for each year obtainable from the Vital Statistics and the populations at risk for which the 1991 Census ward SAS can be used since ward-based population estimates are not available for each year. These datasets can be joined to the digital ward boundaries with crude death rate per 1000 persons calculated (deaths / population at risk * 1000). If no allowances are made for the consistency of the geographical zones over time any spatial data analysis and choropleth mapping will be compromised since boundary changes that have occurred will alter the relationship between the population at risk, the ward and the number of deaths so that biases in the crude death rates will result. To illustrate the problems that can occur, crude death rates have been calculated for wards in Eastern Region with the results classified in quartiles and choropleth mapped for Welwyn-Hatfield district in Figure 26.

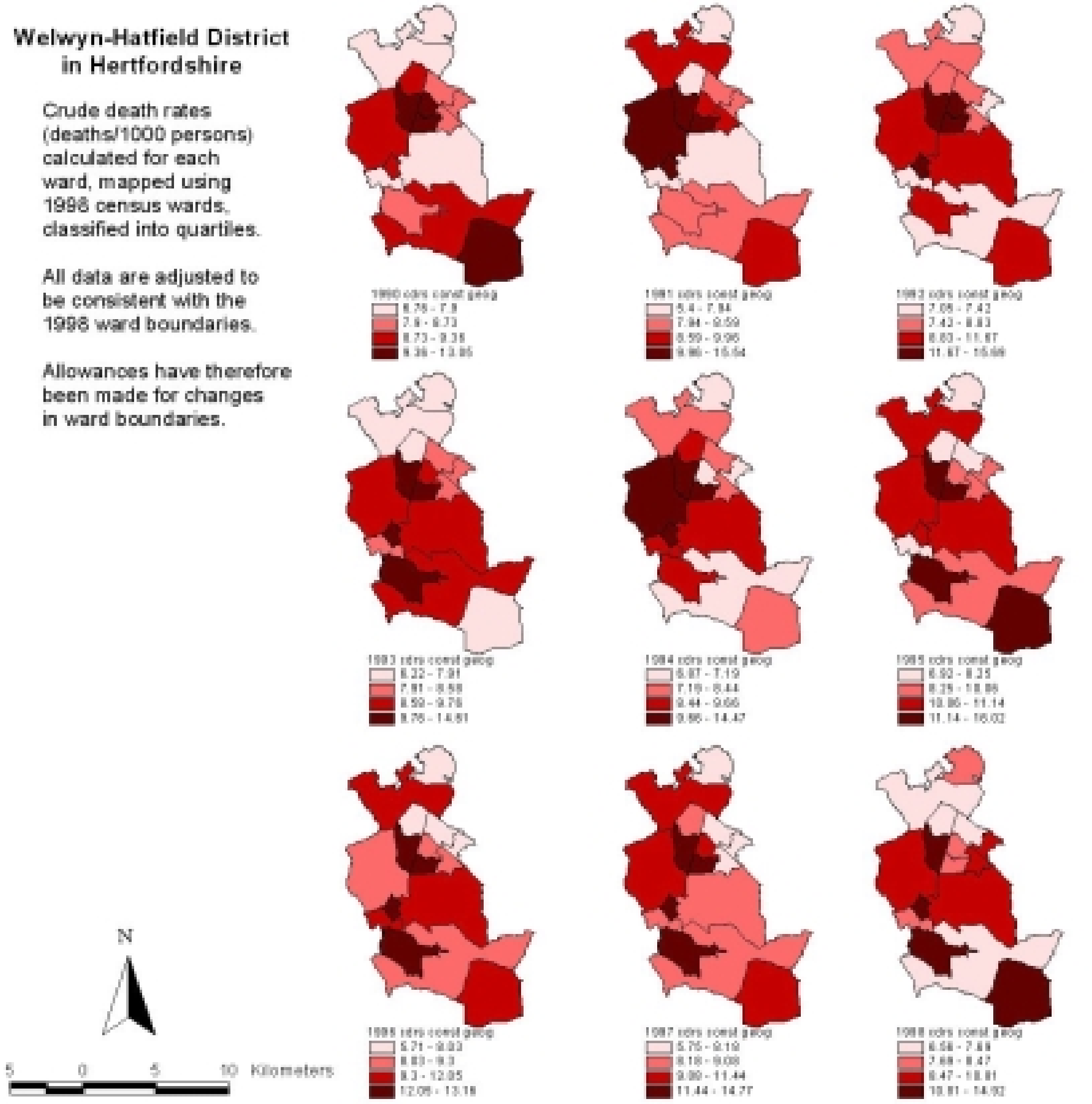
Welwyn-Hatfield district in Hertfordshire experienced various changes to its ward geography just after the Census in 1991 (see Figure 1). Three wards changed size but not their names after the 1991 boundaries were defined and a naïve analysis of the crude death rates will lead to a misinterpretation of the geography of mortality. Haldens, for example, shows an improvement in death rate after 1991 appearing to have one of the lowest rates in the district. Hatfield East, whilst having a varied mortality experience compared to the district as a whole, has crude death rates at least as good as Brookmans Park and Little Heath. Since the number of wards in Eastern Region increased through the study period, some of the data will not be allocated to the digital boundaries and information for Panshanger, a ward created during 1991, is lost since this ward code does not exist in the Census ward boundaries. GIS software does not warn the user of this.

Figure 26: Crude death rates in Welwyn-Hatfield using non-consistent geographical boundaries



For sources see Acknowledgements sections b, c and d

Figure 27: Crude death rates in Welwyn-Hatfield using consistent geographical boundaries



For sources see Acknowledgements sections c, d and g

To allow an analysis of mortality in Eastern Region on a consistent geographical basis the deaths data 1990-98 and 1991 Census ward SAS populations have been adjusted to the 1998 ward geography using the Postcode-Point GCT approach. Crude death rates were calculated, classified into quartiles and mapped using the 1998 digital ward boundaries for Welwyn-Hatfield district (see Figure 27). Mortality in Haldens shows a gradual improvement through the study period but the ward no longer has relatively low rates in every year because the population at risk taken from the 1991 Census has been adjusted to take account of the creation of Panshanger ward. Prior to the use of consistent boundaries Haldens' population at risk was artificially high thereby suppressing the death rate. Comparing the crude death rates calculated on a consistent geographical basis for Hatfield East and Brookmans Park and Little Heath shows that after the first two years of the study period, Brookmans Park and Little Heath ward has at least as good a mortality experience as Hatfield East. This is different from the rates previously calculated and due to the adjustments made to the deaths data and Census SAS populations to account for the boundary changes.

8. CONCLUSIONS

From the late 1990s the establishment of the ONS Geographic Referencing Strategy and Gridlink, the annual availability of the All Fields Postcode Directory and the expected dissemination of the 2001 census in postcode-related outputs should enable researchers to establish consistent geographies and link data between different geographies and time periods. However, investigations of time-series data for periods preceding these innovations are hampered by a lack of comprehensive information by which to tackle the problems posed by boundaries that change over time. Given that the necessary data either do not exist or are not generally affordable to academic researchers, this paper has explored how the widely available information may be used to establish a consistent ward geography for the period 1990 to 1998.

Reviewing the approaches that address the lack of temporal consistency of boundaries has shown that postcodes locations allow detailed data modelling without the necessity for digitised boundary information since their small size offers versatility for aggregation into other areal units and that LUTs can link different geographies with postcode counts used to derive apportionment weights between the geographies. In this paper, the principles of using specialised LUTs to convert data from one geographical zone system to another have been examined along with the assumptions that need to be made about postcode locations and their use as building-bricks for geographical data.

Two methods have been investigated to establish a data time-series on a consistent geographical basis. An ED Building-Brick approach has been developed that is a hybrid of the freeze history and update to contemporary zones approaches. Using Geographic Conversion Tables, data for the year of interest can be disaggregated to 1991 ED geography and then reaggregated using a further GCT to be consistent with 1998 ward geography, the latest year for which all the input data required for the application are available. The organisation of the information for all the lookup tables to indicate the number of postcodes in each ward and a linkage to the 1991 ED geography requires considerable data preparation and some assumptions to be made. The compilation of further LUTs for record matching and data input files requires care due to a lack of harmonisation of ward names, codes and references. Once some geographical linkage errors were corrected, relatively large adjustments to hypothetical populations were identified as being in locations where boundary changes were known to have occurred between 1991 and 1998. There are however methodological problems concerning density effects that relate to the resolution of the postcode grid references and the size differential in areal extent of EDs between urban and rural areas. Since many EDs contain multiple stacked postcodes and some have no postcodes intersections this can lead to an over- and under-allocation of data and the lack of the use of household counts will under-estimate the population in the most urban EDs. Moreover, the inconsistent use of techniques through which links between geographical areas have been established can lead to error propagation.

Subsequently, a Postcode-Point approach has been developed whereby data for wards in each year are disaggregated using GCTs into the constituent postcodes of the ward and then directly rebuilt to contemporary zones using the constituent postcodes of the 1998 wards. This approach alleviates issues identified for the ED Building-Brick approach relating to the postcode-ED density relationships, the lack of the use of household counts and the mixed technique two-step disaggregation-reaggregation approach. The program output is consistent with areas where boundary changes have occurred and with 1991 ward data adjusted to the 1998 ward geography by the UUKAM project. The main advantages to using the Postcode-Point method are that the postcodes themselves are used as the building-bricks being smaller and thus more versatile with quality checks showing the results to be sound. A ward geography has been used in this research but the approach can be used whenever the constituent postcodes of each area of interest are known. The method complements the work of the UUKAM project and is compatible with subsequent releases of the AFPD for later years. Moreover, the technique will continue to be applicable when accurately geocoded individual/household-level data become widely available to the academic community and health professionals.

The experience of investigating a method to establish time-series data on a consistent geographical basis suggests that researchers carrying out similar work would be advised to be aware of the following:

- Until proved otherwise, assume that files containing ward data are incompatible with other ward data through mismatches of file order, ward reference codes, ward name spellings, date of collection and boundary definition even if the files are from the same source for different years and from different sources for the same year. Suppliers of the data are unlikely to volunteer the compatibility provenance of their data and may well not know it.
- 1991 Census data and digital ward boundaries are specific to the date of the Census and to the Census wards. Alternative sources of ward data for 1991 are not necessarily compatible with Census ward data and associated digital boundaries.
- Digital boundaries for wards for other years are time specific. Digital boundaries for other geographies and times are specific to those geographies and times. Assume that there have been boundary changes from one time period to another until proved otherwise. Even if no administrative or legal boundary changes have occurred over time, assume that digital boundaries will be different due to digitising error, changes in precision or generalisation until proved otherwise.
- Data, even from the most reputable sources, may contain inaccuracies.

ACKNOWLEDGEMENTS

a. The authors wish to acknowledge the support of the Economic and Social Research Council (ESRC), CASE Award S00429937028 and CASE partner, the Eastern Region Public Health Observatory (ERPHO), previously the Anglia Clinical Audit and Effectiveness Team (ACET)

b. Digital maps of Eastern Region are based on data provided by the United Kingdom Boundary Outline and Reference Database for Education and Research Study (UKBORDERS) via Edinburgh University Data Library (EDINA) with the support of ESRC and the Joint Information Systems Committee of Higher Education Funding Councils (JISC) and boundary material which is Copyright of the Crown, the Post Office and the ED-LINE consortium.

c. The 1991 Census statistics used in the research are Crown Copyright; the EWCPOP mid-1991 small area estimates are copyright of the Estimating with Confidence project. Both are made available by the Census Dissemination Unit through the Manchester Information and Associated Services (MIMAS) of Manchester Computing, University of Manchester. The 1991 Census data have been purchased for academic research purposes by ESRC and JISC.

d. Vital Statistics (VS4) data for wards in Eastern Region have been extracted from national files made available by the Office for Population Censuses and Surveys (OPCS) and Office for National Statistics (ONS). VS4 for 1989-92 have been obtained via MIMAS with permission from the UK Data Archive. VS4 for Eastern Region 1993-94 on paper files purchased direct from ONS with assistance from Jane Oakley. VS4 for 1995-98 on computer files direct from the Data Archive. The UK Data Archive is jointly funded by the University of Essex (where it is located), ESRC and JISC. Vital Statistics data are Crown Copyright made available through ONS and the Data Archive by permission of the Controller of HM Stationery Office.

e. Electoral data for wards in Eastern Region have been provided by ONS with assistance from Gerald Tessier of the Boundary Commission. Numerous local authority and unitary authority Electoral Registrations Units have answered queries about electoral data, their assistance is gratefully acknowledged.

f. Various lookup tables are made available by the Census Dissemination Unit through the MIMAS service of Manchester Computing, University of Manchester. The Postcode-Enumeration District Directory (PCED) was originally created by OPCS and has been updated by ONS. The Central Postcode Directory (CPD) (Postzon) is based on a file originally created by the Department of Transport and has been enhanced by the Post Office, OPCS, GRO(S), Ordnance Survey, the Welsh Office and various local authorities; permission to use the CPD has been given by the UK Data Archive. The All-Fields Postcode Directory (AFPD) is produced by ONS with information from ONS, GRO(S), NISRA, the Post Office and Department of Health. The Updated UK Area Masterfiles ESRC funded project (H507255164) has re-engineered the AFPD to link census geographies to other administrative geographies. Data in these lookup tables are Crown Copyright, ESRC purchase.

g. Boundary-Line 1998 digital ward boundaries for Eastern Region have been made available by special permission of Ordnance Survey (OS) with assistance from Sallie Payne of OS and Linda See, School of Geography, University of Leeds. Ordnance Survey is a registered trademark and Boundary-Line is a trademark of Ordnance Survey, the national mapping agency of Great Britain. OS Boundary-Line data are Crown Copyright.

REFERENCES

Baker K (1991) *Statutory Instrument 1991 No. 695 The District of Welwyn Hatfield (Electoral Arrangements) Order 1991*, Home Office, HMSO online www.hmso.gov.uk/si/si19910695_en_1.htm accessed 10/07/01

Barr R (1993) Mapping and spatial analysis. Chapter 9 in *The 1991 Census User's Guide* edited by Angela Dale and Cathie Marsh. HMSO, London, pp. 248-268

Blake M, Bell M and Rees P (2000) Creating a Temporally Consistent Spatial Framework for the Analysis of Inter-Regional Migration in Australia, *International Journal of Population Geography*, 6, pp. 155-174

Bracken I and Martin D (1995) Linkage of the 1981 and 1991 UK Censuses using surface modelling concepts, *Environment and Planning A*, Vol. 27, pp. 379-390

Cole K (1994) Data modification, data suppression, small populations and other features of the 1991 Small Area Statistics, *Area*, 26.1, pp. 69-78

Department of the Environment (1987) *Handling geographic information: the report of the Committee of Enquiry chaired by Lord Chorley (the Chorley Report)*, HMSO, London

Flowerdew R and Green M (1994) Areal interpolation and types of data. Chapter 7 in *Spatial analysis and GIS* edited by Stewart Fotheringham and Peter Rogerson, Taylor & Francis, London, pp. 121-145

Gatrell A C, Dunn C E and Boyle P J (1991) The relative utility of the Central Postcode Directory and Pinpoint Address Code in applications of geographical information systems, *Environment and Planning A*, Vol. 23, pp. 1447-1458

Hansard (1988-2001) *Hansard (House of Commons Daily Debates).* Online http://www.parliament.the-stationery-office.co.uk/pa/cm/cmhansrd.htm, accessed 27/08/01

Heywood I (1997) *Beyond Chorley: current geographic issues*, Association for Geographic Information, London

HMSO (1987-2001) *UK Statutory Instruments.* HMSO available online http://www.hmso.gov.uk/stat.htm accessed 10/07/01

Martin D (1992) Postcodes and the 1991 Census of Population: issues, problems and prospects, *Transactions of the Institute of British Geographer*, N.S. 17, pp. 350-357

Martin D (1996) *Geographic Information Systems: Socioeconomic applications*, Routledge, London

Martin D and Higgs G (1997) Population georeferencing in England and Wales: basic spatial units reconsidered, *Environment and Planning A*, Vol. 29, pp. 333-347

MIMAS (1999) *Lookup Tables: Central Postcode Directory, Postcode-Enumeration District Directory, All-Fields Postcode Directory*, Manchester Information and Associated Services online census.ac.uk/cdu/Datasets/Lookup_tables/ accessed 10/07/01

Norris P and Mounsey H M (1983) Analysing change through time. Chapter 9 in *A Census User's Handbook* edited by David Rhind, Methuen, London pp. 267-286

ONS (1997) *UK standard geographic base – concepts and implementation plans*, Working Paper No. 34, Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Brighton, 22-25 September 1997, online www.unece.org/stats/documents/1997/09/gis/34.e.html accessed 10/07/01

ONS (1999) *The 2001 Census of population*, White Paper Cm 4253, The Stationary Office, London

ONS (2000) *Geography in National Statistics*, Office for National Statistics, online www.statsbase.gov.uk/nsbase/methods_quality/geography/home.aps, accessed 21/11/00

Openshaw S (1991) Developing appropriate spatial analytical methods for GIS. Chapter 25 in *Geographical Information Systems: Principles and Applications*, edited by David J Maguire, Michael F Goodchild and David Rhind, Longman, Harlow, pp. 389-402

Openshaw S and Rao L (1995) Algorithms for reengineering 1991 Census geography, *Environment and Planning A*, Vol. 27, pp. 425-446

Openshaw S and Taylor P J (1981) The modifiable areal unit problem in *Quantitative Geography: A British View*, Wrigley N, Bennett R J (eds), Routledge and Kegan Paul, London, pp. 335-50

Penhale B, Noble M, Smith G and Wright G (2000) *Ward level population estimates for the* 1999 *Index of Local deprivation,* University of Oxford, Department of Applied Social Studies and Social Research, Oxford

Raper J, Rhind D W and Shepherd J (1992) *Postcodes: the New Geography*, Longman, London

SASPAC (1992) *SASPAC User Manual, Part 1,* CST 450, First Edition, April 1992, Manchester Computing Centre, Manchester

Shaw M, Dorling D and Brimblecombe N (1998) Changing the health map in Britain 1951-91, *Sociology of Health and Illness*, Vol. 20 No. 5, pp. 694-709

Simpson L (2001) *Updated UK area masterfiles: final report*, paper presented at the ESRC Census Development Programme workshop, 15th – 16th May 2001, School of Geography, University of Leeds. Online http://les1.man.ac.uk/ccsr/rschproj/lookup.htm accessed 10/07/01

Simpson L and Yu A (2000) *Data conversion and look up tables*, report of the Updated Area Masterfiles Project, ESRC Census Programme 3rd workshop, 1st November 2000, Royal Statistical Society, London. Online http://les1.man.ac.uk/ccsr/rschproj/lookup.htm accessed 10/07/01

UKBORDERS (2001) *Digital maps of Eastern Region*, obtained via Edinburgh University Data Library (EDINA), online http://edina.ac.uk/ukborders accessed 10/07/01

UKSGB (2000) Geographic Base Characteristics. Section 3 in *UK Standard Geographic Base, Best Practice Guidelines,* UKSGB online www.ngdf.org.uk/uksgb/homepage.html via GIraffe Data Integrator to Guidelines, accessed 12/06/01

UKSGB (2001) *United Kingdom Standard Geographic Base: Gridlink.* Online www. ngdf.org.uk/uksgb/homepage.htm, accessed 19/01/01

White B, Gregory I and Southall H (1998) *Analysing and Visualising Long-Term Change*, Department of Geography, Queen Mary and Westfield College, online www.geog.qmw.ac.uk/gbhgis/gisruk98/ accessed 22/11/00

Wilson T and Rees P (1998) Lookup tables to link 1991 population statistics to the 1998 local government areas *Working Paper 98/5*, School of Geography, University of Leeds, online http://www.geog.leeds.ac.uk/wpaper/wp98-5.pdf accessed 10/07/01

Wilson T and Rees P (1999) Linking 1991 population statistics to the 1998 local government geography of the United Kingdom, *Population Trends* 97 pp. 37-45

Yu A and Simpson L (2000) *The All Fields Postcode Directory (AFPD): validation for use as a lookup table*, progress report prepared for ESRC Census Development Programme workshop, 3rd–4th May 2000, School of Geography, University of Leeds, online http://les1.man.ac.uk/ccsr/rschproj/lookup.htm accessed 10/07/01