

# Geography Conversion Tables: A Framework for Conversion of Data between Geographical Units

Ludi Simpson\*

*Cathie Marsh Centre for Census and Survey Research, University of Manchester, UK*

## ABSTRACT

**A conceptual framework for conversion of data between geographical units is developed. It is used to examine the construction of geographical conversion tables, which have taken many forms and have recently been stimulated in the UK by government emphasis on neighbourhood profiling. Where boundaries of different geographical systems overlap, then conversion of data from one system to another involves approximation. In this case non-hierarchical geography conversion tables are used and data conversion is equivalent to statistical synthetic estimation. Errors arise both in the construction of geographical conversion tables and in the approximation when converting data between overlapping geographies. A typology of errors when estimating data by geographical conversion is complemented by empirical measurement for a variety of UK examples. Data conversion using weighted sums of source unit data is shown to be more accurate than the allocation of data for whole source units to single target units. Measures of quality for estimates from data conversion are necessary and are proposed. Copyright © 2002 John Wiley & Sons, Ltd.**

*Received 19 June 2001; revised 13 October 2001; accepted 19 October 2001*

**Keywords:** geography; look-up tables; synthetic estimation; neighbourhood profiles

## INTRODUCTION

In the UK, estimation of social statistics for small areas has become a priority for central government (National Strategy for Neighbourhood Renewal, 2000). This priority is stimulating fruitful advances at the intersection of statistical science and computational geography, one of which is the transfer of data from one set of geographical units to another.

The main purposes of converting data from one geography to another are:

- To aggregate data to units sufficiently large to provide reliable results (for example, from postcoded records to local government areas)
- To present results for areas that are familiar to the audience for the research (for example, from small census units to current electoral areas)
- To estimate a time series on a consistent basis (for example, health administrative statistics before and after boundary changes)
- To merge data-sets drawn from different sources (for example, for neighbourhood profiles containing both census data and data based on postal geography).

As an illustration, welfare benefits data are

---

\* Correspondence to: L. Simpson, Cathie Marsh Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL, UK.  
Email: ludi.simpson@man.ac.uk

published in Britain referring to local ward boundaries current at the time of collection. The benefits data from two separate years when ward boundaries differed were converted to each refer to the same set of parliamentary constituency boundaries. The approximation involved where a ward did not precisely fit into a single constituency was not sufficient to invalidate the comparison over time, and the data could then be displayed for geographical areas that are more meaningful to the audience of political scientists than the original units (Dorling and Simpson, 2001).

A 'look-up table' has been the common but somewhat misleading term for a geography conversion table. Such tables are often not hierarchical; one unit of source geography may be shared between many units of target geography, making the resulting conversion as much a process of 'looking down' as 'looking up'. The 'Updated UK Area Master-files' project from which this paper arose has created 200 UK-wide conversion tables and an on-line facility for their use (<http://convert.mimas.ac.uk>, 2001; Simpson, 2001).

The only way to avoid all approximation in data conversion is to maintain source data-sets as individual records grid-referenced to points. These may be aggregated to any set of geographical units, and thus compared over time and to other such data-sets, with no approximation involved. However, for confidentiality as well as practical resource reasons, many data-sets will continue to be held as geographical aggregates, presenting analysts with the challenge of data conversion.

Data conversion is therefore central to social research on local conditions, to market research and targeting, and to the management of local governmental programmes. This paper provides a general framework for geographical conversion tables and data conversion between geographies. It investigates and quantifies the quality of the resulting statistical estimates.

The structure of the remainder of this paper is developed as follows. Different ways in which geography conversion tables are constructed are described through examples. A conceptual framework follows for geography conversion tables and their use to convert data from source units to target units. The errors

and approximations involved in data conversion are categorised and quantified. A number of challenges for the consistent maintenance of geography conversion tables, with weights relevant to the data that may be converted, are itemised. Finally, discussion returns to a summary of the costs and benefits of weighted conversion of source data to estimate statistics for target areas, relative to the aggregation of whole source units.

## EXAMPLES OF GEOGRAPHY CONVERSION TABLES

The examples described in this section have been chosen to include a wide variety of construction methods and of weighting criteria (the measure of the overlap between two geographical units). Measurement of the relationship between units of different geographies often involves maps. When maps are computerised, the functions of Geographical Information Systems (GISs) can be usefully employed. Maps are not used when conversion tables are derived from a database list of records, each referenced to more than one geography. Table 1 summarises the geography conversion tables described in this section. Where a map is sufficiently detailed to display buildings, it may be used to assess visually the extent of residential areas either side of boundaries, and thus manually write down a conversion table from one set of areas to another. In the UK, Census area boundaries in 1971, 1981 and 1991 were provided to local authorities overprinted on large-scale Ordnance Survey maps. These have been commonly used to link a local neighbourhood area to its standard census areas for which statistics are published. The conversion table thus written down can be used in the UK census extraction package SASPAC (London Research Centre, 1999) where it is termed a 'gazetteer file'. The method is laborious, dependent on an up-to-date map base, and approximate, but has successfully served census analysts over three decades.

To compare historical social data before and after the major boundary changes of 1933–1936 in the UK, a conversion table from 1931 to 1939 boundaries has been derived from governmental records of boundary changes. The

Table 1. Construction methods for geography conversion tables.

Method of construction	Example <sup>a</sup>	Source units	Target units	Weighting criterion
1. Visual examination of detailed maps	Construction of SASPAC gazetteer files	Census output areas	Service areas	Number of dwellings
2. Legal reports of boundary changes	Southall, the '31-'39 convert table	1931 Local Government Districts	1939 Local Government Districts	Population
3. Statistical reports of boundary changes	Wilson and Rees, District boundary changes	1991 Census EDs	1998 Local Government Districts	Population
4. GIS analysis of overlapping digitised boundaries	Noble <i>et al.</i> ward boundary changes	1991 Census EDs	1998 Electoral Wards	Area (hectares)
5. GIS analysis of overlapping digitised boundaries, with point content	Bradford Community Statistics Project	1991 Census EDs	Neighbourhoods	Number of addresses
6. GIS analysis of points within a single boundary set	Dorling and Atkins, Census comparison over time	1991 Census EDs	1981 Census wards	Population
7. Digitised points within digitised boundaries	US Census TIGER Map service	Zip codes	Census standard areas	See text
8. Derived from a Census database of households	England and Wales 'ED-postcode link'	Postal codes	Census geography	Number of residential households
9. Derived from a database of postcodes	Updated UK Area Masterfiles project	Eight postal, census and electoral geographies	25 postal, census, electoral, administrative and statistical geographies	Number of residential addresses

<sup>a</sup> For details of each example, see text.

Registrar General's Statistical Reviews list each change of boundary, and with it the 1931 Census population in each territory transferred. Thus the 1931 Census population is the weighting criterion for old areas that are shared between more than one new area. Further investigation was required when the same area was affected by more than one change during the 1930s. The resulting table is now used to convert 1931 Census data to 1939 boundaries, which then changed little until the next census in 1951 (Southall, 2001).

Wilson and Rees (1999) have recast the entire UK 1991 Census small area statistics, as well as

population estimates, to fit the new local government District boundaries, many of which had been subject to boundary changes by 1998, including the creation of new authorities. They constructed conversion tables from lists of census areas in each new District that were provided by the UK statistical offices.

Noble *et al.* (2000) created population estimates for 8000 electoral wards of England as established in 1998, using among other ingredients the 1991 Census data from 100,000 Census Enumeration Districts (EDs). GIS software overlaid digitised boundaries to identify the extent of surface area overlap measured in

hectares. The method provided an automated means of allocating Census data to a geography that had not existed at the time of the Census. However, large surface areas can contain small or no population. The weight calculated on the basis of surface area may correlate poorly or negatively with the distribution of Census data within EDs (and in general with any data based on settlements or industry) and thus add approximation to data conversion.

GIS can also provide a count of events within each overlap of two sets of digitised boundaries, using 'spatial join' and 'point in polygon' routines. In the UK, the Ordnance Survey product Addresspoint records the grid reference of each address, providing a weight for geography conversion tables based on the number of addresses rather than the surface area in each overlap of boundaries. Addresspoint does not distinguish residential households from others, but the existence of an organisation name identifies most non-residential addresses. GIS has been used in this way to create geography conversion tables with weights relevant to where people live (Thomasson, 2000). These have allowed a boundary-free on-line statistical service for the Bradford region (accessed from <http://www.bcsp-web.org>, 2001).

Dorling and Atkins (1995) employed a similar approach to allocate 1991 Census enumeration districts to 1981 Census wards. The weighting criterion was the population recorded in the Census. Manual validation was able to improve the automated procedure, before using the results to monitor changing subnational population densities between 1971 and 1991.

The US Census Bureau provides its Census data via a 'Look-up' service. While this does not involve data conversion as such, one facility does allow, like many other Internet applications, the user to enter a zip or postal code. A map with census boundaries is centred on a point labelled with the chosen zip-code as an aid to the user's navigation to the areas they are interested in. Separately, whole census block areas are aggregated to approximate 'ZIP Code Tabulation Areas' (<http://www.census.gov/geo/www/garm.html>, 2001).

A database of statistical or administrative

records may contain more than one geography, from which conversion tables can be derived without the use of maps or digitised boundaries. The UK Census database contains codes both for census geography and for postal geography. The Office for National Statistics (ONS) created from it a geography conversion table, the ED-postcode link, showing the number of households in the overlap of each postcode and Census ED, the smallest postal and census units (ONS Geography, 1995). The directory is updated each year as postcodes are changed. From the directory users can create further geography conversion tables between any census and postal classifications - for example, from postal districts to Census Districts. Similar directories are held for Scotland and Northern Ireland by their respective statistical agencies.

Another ONS directory, the All Fields Postcode Directory (AFPD), represents each postcode in the UK by a single record. The directory is compiled four times yearly by the Office for National Statistics from information supplied by the Ordnance Survey, the Royal Mail, Electoral Boundary Commissions, the General Registrar's Office for Scotland, the Northern Ireland Statistics and Research Agency and others (ONS Geography, 2000). Each record includes reference to the number of residential addresses in that postcode, and a code for each of several geographies. The code indicates the unit of that geography that the postcode mainly lies within.

The directory itself is a conversion table from postcodes to each of the geographies it records. It is by construction a hierarchical table in which each postcode is entirely allocated to a single target area. However, a new conversion table can be constructed from the postcode directory when any two of its other geography codes are chosen as source and target respectively. Where the AFPD indicates that all the postcodes in a source geography unit lie within one target geography unit, the conversion table has a single record for that source unit, with weight one. Where the source unit contains residential postcodes allocated to different target units, the AFPD provides a weight based on the number of residential addresses in the overlap of the source unit with target geography units.

Table 2. A geography conversion table: example

Source unit (Census ED) s	Target unit (Electoral ward) t	Weight $w_{st}$
EGFW14	JAMX	0.3095
EGFW14	JAMY	0.6905
EGFW15	JAMX	0.7667
EGFW15	JAMY	0.2333
EGFW16	JAMY	1.0000
EGFW17	JAMY	1.0000

Source: Updated UK Area Masterfiles project. Records are from the Peterborough area.

After cleaning the AFPD and adding Census geography codes for Scotland and Northern Ireland, it has generated 200 UK-wide conversion tables (Simpson and Yu, 2001; Simpson, 2001; <http://convert.mimas.ac.uk/>, 2001) for academic and public use.

The conversion tables discussed here have different formats and varied construction but can all be described within the framework outlined in the next section.

## CONCEPTUAL FRAMEWORK

A *geography* is a set of non-overlapping units of geography, and a *geography conversion table* is a list of records, each including three fields to identify:

- s A source geography unit – for example, a code for a 1991 census ED.
- t A target geography unit – for example, a code for a 1998 electoral ward.
- $w_{st}$  A weight, taking a value more than zero but less than or equal to one.

The weight shows the proportion of the source geography unit that lies in the target geography unit, according to some *weighting criterion*.

Table 2 shows an example of a geography conversion table, from the UK 1991 Census output areas – Enumeration Districts (EDs) – to UK 1998 electoral wards. The first two EDs overlap more than one ward and therefore have more than one record, while the next two EDs are entirely contained within a single ward. Geography conversion tables may in practice be arranged differently from Table 2,

and have extra fields to show, for example, the labels for each geography unit or the values from which the weight has been calculated, but their essential properties are contained in the three fields as defined above.

Documentation of the conversion table (the prefix ‘geography’ and sometimes also ‘conversion’ are dropped where the sense is not made ambiguous) defines the origin of geographical boundaries and the weighting criterion, including their reference dates.

The weighting criterion may be the surface area (for example, expressed in hectares), the population, the number of households or addresses, or another variable. The weights for records of the same source unit add up to a maximum of 1, since the unit can only be allocated once. If the sum always equals 1, the conversion table is *exhaustive*, and no data are lost on data conversion.

When the source geography units are much smaller than the target units, many of them will be represented by one record with weight equal to one. If all weights are one, the conversion table is *hierarchical* (as, for example, Census Output Areas are all contained within Census Districts). The percentage of source units with weight equal to one is the *degree of hierarchy*. This can be calculated as:

$$100 * \frac{\sum_{s,t} (w_{st} = 1)}{\sum_s (1)}$$

In Table 2 the degree of hierarchy is  $100 * 2 / 4 = 50\%$ .

Recognising that a source unit distributed half-and-half to two target units is a poorer fit to the target geography than one that is almost wholly within a single target unit, the *degree of fit* sums the maximum weight for each source unit, again expressed as a percentage of all source units. The degree of fit can be calculated as:

$$100 * \frac{\sum_s (\max w_{st})}{\sum_s (1)}$$

In Table 2 the degree of fit is  $100 * (0.6905 + 0.7667 + 1 + 1) / 4 = 86.4\%$ .

The degree of fit is used in the next section as an important measure of uncertainty when

Table 3. Data conversion: example.

Source unit s	Source data $D_s$	Target unit t	Converted target data $D_t$
EGFW14	20	JAMX	29
EGFW15	30	JAMY	36
EGFW16	5		
EGFW17	10		

Each source data value is multiplied by the appropriate weight from Table 2.

tables are used to convert data from source to target geography. The degree of fit may also be calculated for a single target unit. The calculation for a single unit can be simply the sum of the weights with target t divided by the number of source units with target t, or be extended to reflect that weights of 0.1 or 0.9 may be considered equally close a fit to a single target unit.

Data for source geography units ( $D_s$ ) are converted to data for target geography units ( $D_t$ ) using the weights recorded in the conversion table:

$$D_t = \sum_s w_{st} D_s$$

Table 3 shows data converted using the conversion table of Table 2.

When data are converted, unless the table is hierarchical (all weights equal to 1), uncertainty is added in the conversion, such that the results for target units are *estimates*. The degree of hierarchy and degree of fit express the amount of estimation involved in data conversion. The weighting criterion in data conversion is exactly equivalent to the ancillary variable in a synthetic estimator for small areas described in statistical texts (Skinner, 1993; Ghosh and Rao, 1994).

When an independent data-set represents the truth, the approximation involved in data conversion has been measured in the next section by the absolute error, also expressed as a percentage of the true value.

## DATA CONVERSION IN PRACTICE

Data conversion with no error is only possible

when source data are accurately coded to dwelling units or to other points, which can then be aggregated hierarchically to any target geography. In that case all weights have value one, and no synthetic estimation is involved. This is one goal of the government's current geo-referencing strategy (National Strategy for Neighbourhood Renewal, 2000: Annex G). The focus on geo-referenced point data does not address data that will continue to be held for geographical areas. These will be common to prevent data disclosure from individual records and to represent large microdata sets as condensed data cubes (Westlake, 2000).

Thus conversion of data between overlapping geographical areas will continue to require conversion tables with non-unity weights as described above. This section takes a critical look at such data conversion. It categorises and quantifies with examples the approximation that is an inevitable consequence of converting data from one set of geographical areas to another.

The conversion tables described in Table 1 stand apart from the data to be converted. The SASPAC software has already been mentioned as a tool to transfer specific census data-sets to new zones, using conversion tables provided by the user. Other users may use statistical software such as SAS or SPSS to match the data for source units to a conversion table, multiply each record of source data by the weight indicated in the conversion table, and then aggregate the weighted data values allocated to each target unit. More generally, any relational database software may be used to relate a data-set to a conversion table, and to group the weighted records for output of the target totals.

The Updated Area Masterfiles project created a website that at the time of writing makes use of 200 UK geography conversion tables. Users 'upload' their own data for conversion to their selected target geography, which they then 'download' together with a log of unmatched source data units (<http://convert.mimas.ac.uk>, 2001). Convenient data conversion between geographies is now a reality for non-technicians, with no need for programming and database skills.

None the less, users need to be aware of the

Table 4. Error when estimating from whole source units and from weighted source units.

Method of allocation	Target area size		Absolute error in estimated number of claimants in a target area		
	Mean number of claimants	Mean number of addresses	Maximum	Root mean square error Number	%
Weighted source units	726	2477	66.4	21.5	3.0
Whole source units	726	2477	245.0	73.5	10.1

Source units: 927 Census EDs in Bradford District.

Target units: 70 Neighbourhood areas.

Data: Number of benefit claimants.

levels of approximation arising from inaccurate construction of the conversion table and from the data conversion itself. In each case the reason for approximation can be usefully subcategorised, resulting in four sources of error in converted data:

- Errors from the construction of the conversion table
  1. Best-fit whole allocation of source units rather than weighted allocation.
  2. Incorrect calculation of the weights.
- Errors from data conversion
  3. The source units are not wholly contained within single target units.
  4. The weighting criterion is not correlated to the data within divided source units.

There is no general level of error that should be expected, since it will depend on the geographies involved, the construction of the conversion table, the weighting criterion and its correlation with the data. For this reason, the examples in this section must be illustrative rather than comprehensive. The final section will return to a general discussion of the errors in data conversion. For a specific conversion of data, the error will be less than the sum of these four types of error, for while they are independent they will often cancel out.

A variety of sampling and non-sampling errors also affect the source data before conversion. The additional errors from conversion tables and their use should be assessed in the context of those other errors, and may or may not be relatively insignificant.

### Approximation 1: Best-Fit Whole Allocation of Source Units Rather than Weighted Allocation

To avoid estimated statistics, or because no weighting criterion exists, conversion tables are sometimes constructed with weights taking only a value of one. Each source unit is then wholly allocated to the single target unit which it fits best by some stated criterion. Table 4 shows the loss of accuracy when using a best-fit approach to estimating neighbourhood statistics, as tested for the development of the online community statistics system for the region of Bradford, referred to above.

The location of welfare claimants is known within the Bradford region to Bradford Council which administers the benefits, but is not generally available for reasons of confidentiality. Counts of claimants were therefore first allocated to 927 non-confidential source census areas (EDs) for use in the public Community Statistics System. They were subsequently converted to 78 target Neighbourhoods in the region using weights based on the number of addresses in the overlaps of each geography (Thomasson, 2000; <http://www.bcsp-web.org>, 2001). Bradford Council constructed the Neighbourhood areas independently of Census areas for the purpose of local government consultation. The average error in the number of claimants in each Neighbourhood as estimated from weighted sums of source units, when compared with the true number of claimants known to the Council, was some 22 claimants out of an average 726 claimants, or less than 3% (Table 4). When allocating EDs

Table 5. Error when estimating census areas by aggregating whole postcodes.

Census unit (target)	Number of units	Median absolute error	Median absolute % error
ED	112,551	10	5.73
Ward	9,592	8	0.51
District	403	12	0.03

Source: ED-postcode link directory, England and Wales 1991. 1,744,476 records for 1,146,224 postcodes with at least one resident.

Source units: 1991 postcodes in England and Wales.

Target units: 1991 Census areas.

Data: number of resident households.

wholly, based on the location of the centroid of the ED, the average error was increased more than three-fold, to 74 claimants, or over 10%. The maximum error across the 70 Neighbourhood areas was also magnified three-fold when allocating source data as whole units.

The amount of error added from using best-fit allocation will be less when the source units are small relative to the target units. Then most source units will be wholly allocated to a single target unit under either strategy.

The errors from best-fit allocation can be explored further using the England and Wales ED-postcode link directory referred to in Table 1 (example 8). The use of postcodes to allocate survey and administrative data to other geographies is common and is usually achieved by allocation of each postcode wholly to a single target unit. The ED-postcode link directory records the number of enumerated households falling into each overlap of postcode and census geography. The recorded true number of households for each ED is compared in Table 5 with the number that results when allocating the census households in a postcode wholly to the census area in which *most* of the postcode's households lie.

The median absolute error from using whole postcodes in a best-fit conversion table is considerable when the targets are the smallest census unit (EDs), amounting on average to a little over 5% of the true recorded number of households. However, the percentage error reduces far below 1% for wards and for Districts in spite of their much longer boundaries over which wrong allocations can be made. The over-allocations and under-allocations tend to balance each other.

Different methods of allocating a whole

source unit to a target unit were used in these two examples. The location of the centroid of address grid references, and the location of the majority of addresses, would lead to different allocation in some cases. Thus while allocation of whole source units avoids estimated statistics, the results are none the less dependent on choices taken by the provider of statistics and hidden from the 'end user'.

### Approximation 2: Incorrect Allocation of the Weights

Regardless of the use of best-fit whole source units, or weighted source units, data conversion will create extra error if the source units are wrongly allocated to target units. Normally this is impossible to gauge quantitatively. However, two of the files referred to in Table 1 (examples 4 and 9) have been combined for England to provide a much improved allocation of UK 1991 census geography (EDs) to 1998 electoral geography (wards), as reported in Simpson (2001).

The main problem lay in an inconsistency in the construction of the postcode database from which the second file was derived. The errors

Table 6. Error induced in converted data by incorrect allocation of weight.

Electoral unit (target)	Median absolute error	Median absolute % error
Ward	32	1.82
District	58	0.13

Source units: 1991 Census EDs in England.

Target units: 1998 wards and Districts.

Data: number of resident households.

in the database were not clustered in a few areas; the general phenomenon presented itself as source census units overlapping with too many electoral wards. Because both ED and ward geographies are relatively small, errors in allocation are likely to have a relatively large impact.

The error shown in Table 6 is that which has been eliminated by the use of the improved equivalent table. The error is considerable, especially for wards, where the number of households is mis-estimated by the database conversion table on average by nearly 2% of the more accurate figure based on the improved file.

### Approximation 3: The Source Units are Not Wholly Contained within Single Target Units

Imagine there was no inaccuracy in the construction of a geography conversion table – the weight was an accurate reflection of the intended weighting criterion in every overlap between each pair of geographical units. There would still be approximation in the data conversion wherever a source unit does not lie wholly within a single target unit.

The extent to which the data conversion does not allocate whole units is measured simply by the degree of hierarchy and the degree of fit defined earlier and exemplified in Table 7. The degree of fit shows simply the proportion of source units that lie wholly within one target unit (weight = 1) according

to the conversion table. This can be quite low even for geographies that are approximately equal. Thus all but 10% of local government District boundaries had changed by 1998, although most of these changes were minor at the boundary edges, or amalgamations of more than one District into a single new authority.

The degree of fit shows more precisely the proportion of data that is not subject to estimation, by summing the maximum weight for each source unit and expressing it as a percentage of the number of source units. Thus if a source unit is shared equally between two target units, 50%–50%, the degree of fit is lower than if it is 99% in one target unit and 1% in another. The table makes the minor nature of the District changes clear, since the degree of fit is over 97%. Postal sector geography does not fit so well within electoral ward geography, with a degree of fit of under 70%, although the degree of hierarchy is similar at 10.6%.

Generally, the smaller the source unit relative to the target unit, the higher are the degrees of hierarchy and fit. The 100% fit of unit postcodes to census and electoral geography is a result of the creation of these conversion tables from a directory of whole postcodes as already described. Table 8 shows for England and Wales a more accurate degree of hierarchy and fit between postal geography and Census geography in 1991, based on the same part-postcode directory used in Table 5 above. 1991 Census EDs were not based on

Table 7. Selected UK geography conversion tables: degrees of hierarchy and fit.

Source geography (number of units)	Target geography (number of units)	Degree of hierarchy	Degree of fit
ED '91 (150,909)	Postal District '99 (2,780)	90.5%	98.4%
ED '91 (150,909)	Postal Sector '99 (9,232)	76.8%	95.4%
ED '91 (151,543)	Ward '98 (11,124)	95.8%	99.1%
Ward '91 (11,103)	Postal District '99 (2,780)	46.5%	91.4%
Ward '91 (11,103)	Parliamentary Const. '97 (659)	70.1%	99.1%
District '91 (459)	District '98 (434)	10.2%	97.1%
Postal Sector '99 (9,252)	Ward '98 (11,134)	10.6%	68.5%
Postcode '99 (2,131,286)	ED '91 (153,275)	100.0%	100.0%
Postcode '99 (2,131,286)	Ward '98 (11,171)	100.0%	100.0%

Notes: The degree of hierarchy and degree of fit are as defined in the text. The tables are derived by the Updated UK Area Masterfiles project.

Table 8. Four conversion tables between postal and 1991 Census geography, using part postcodes: degree of hierarchy and degree of fit.

Source unit	Target: ED		Target: Ward	
	Hierarchy	Fit	Hierarchy	Fit
Postcode unit	78%	94%	96%	99%
Postal sector	2%	15%	8%	68%

Notes: The degree of hierarchy and degree of fit are as defined in the text. The tables are derived by the Updated UK Area Masterfiles project.

postal geography (unlike in 2001). Only three-quarters of postcode units fit wholly within a census ED.

The measures of fit as used so far are properties of the entire conversion table. Importantly for applications, the same measures can be applied to each target geography unit within a conversion table. Some target units may be the aggregation only of whole source units; others may also include estimated parts of source units. The degree of fit to a single target unit measures an important aspect of the reliability of each value output from data conversion for that target unit.

#### Approximation 4: The Weighting Criterion is Not Correlated to the Data within Source Units (Synthetic Estimation)

Where the fit between a pair of geographies is not exact – and the previous paragraphs show that this is often the case – an accurate conversion table has more than one record for each source unit, with a value of the weight showing how it is shared between each target unit. Data conversion uses the weight to share data from the source unit to the target units. The synthetic estimate for each target unit carries error to the degree to which the weighting criterion within the source unit is not exactly correlated with the distribution of the data to be converted. One can rarely measure the correlation, as the data being converted are not usually available for the intersections of source and target units. Table 4 gave one instance where these data were available, showing an error of 3% when converting benefits data from Census EDs to areas approximately ten times their size.

Synthetic errors of this sort can be significant when social data include strong geographical concentrations of particular types of people. For example, if a source unit's benefit claimants were mainly resident at one hostel, the erroneous sharing of this area's claimants between more than one target area would result in significant errors for each target area.

One can closely simulate the heterogeneity of census variables within larger census units by the heterogeneity between EDs. The degrees of fit between Census EDs and postal districts, and Census wards and postal districts, are 98.4% and 91.4% respectively. 1991 Census EDs fit much more closely within postal districts than do Census wards. They are used in Table 9 to identify the error involved in conversion of census characteristics from Census wards to postal districts.

Table 9 compares the conversion of Census data from 1991 Census wards to 1999 postal districts with the more accurate conversion from the much smaller 1991 Census EDs. The differences are due to the imperfect geographical correlation between the census data and the weighting criterion (the number of residential addresses) within each ward.

The census number of households in a postal district is well estimated from data conversion from wards, because it is highly correlated to the current number of addresses on which the conversion weights are based. The median absolute percentage error is only 0.6%. The number of pensioner households of different types is geographically heterogeneous within census wards – not closely correlated to the number of addresses in each area. It is thus less well estimated for postal districts by data conversion from wards, with percentage errors of over 1%. Relatively rare events which are

Table 9. Error due to synthetic estimation.

Household characteristic	Mean no. of households in a postal district	Error when postal district characteristics are estimated from ward data		Standard deviation	
		Median absolute error	Median absolute % error	Correct (from EDs)	Estimated (from wards)
<i>(a) Target: 2142 postal districts in England and Wales</i>					
Resident households	9220	46	0.6%	5804	5792
With pensioners	2092	27	1.1%	1942	1932
Only pensioners	2293	23	1.3%	1462	1454
Lone pensioner	1385	16	1.4%	915	910
Lone parent	378	5	2.6%	354	349
Crowded households	45	1	6.8%	84	82
<i>(b) Target: 92 postal districts of West Midlands County</i>					
Resident households	9900	115	1.0%	5803	5761
With pensioners	3391	98	3.0%	2072	2037
Only pensioners	2427	90	3.5%	1509	1483
Lone pensioner	1508	52	2.9%	918	904
Lone parent	477	44	9.9%	350	334
Crowded households	55	6	18.4%	65	58

Source units: 1991 Census wards.

Target units: 1999 postal districts.

Data: households and their characteristics.

Notes: 'With pensioners': all households with pensioner(s) with or without others. 'Lone parent': households of one adult with one or more dependent children. 'Crowded': households with more than one person per room.

The benchmark from which the 'errors' are calculated is the allocation of ED data to postal districts. The ward data was created as the sum of ED data, so that none of the discrepancy is due to census data modification.

Source: 1991 Census Small Areas Statistics; 'ED 91 to Postal District 99' and 'Ward 91 to Postal District 99' conversions using tables based on the All Fields Postcode Directory.

also not evenly spread within wards, such as crowded households and lone parents, are still less well estimated by data conversion from wards to postal districts. The number of crowded households is mis-estimated by an average of over 5%. Whether these errors are acceptable depends on the purpose of the data conversion and the availability of alternative data.

These average errors across England and Wales should not allow users to feel that the errors are never very great. The same analysis when restricted to West Midlands County, with similar-sized postal districts, shows considerably more heterogeneity within its wards (which tend to be larger as in other urban areas) and thus considerably more error in the synthetic estimation of postal districts from ward data.

Also evident is the 'numbing' of the data, a reduction of variation between areas once estimated by data conversion. This is also an

expectation from the statistical literature of synthetic estimation already referenced. In the example, the reduction of variation is not great because the error in estimation for each target area is not great.

Clearly if data were available for postal districts directly or for EDs, then these should be used in preference to ward data. This section has shown the magnitude of the error that occurs due to synthetic estimation when only ward data are available (as is the case for many 1991 Census data in the UK, and for non-census demographic and other data).

## MAINTENANCE AND DEVELOPMENT

The practical examples of data conversion in the previous section clarify a number of issues that affect the development of geography conversion tables. Ideally the weight in a conversion table will be closely correlated to the data to be converted. Since there is a limit

neither to the data-sets that may be converted, nor to their various distributions within source units, a single weighting criterion will not be optimal. Were there to be a choice of weighting criteria, that choice would offer different statistics related to each of population, employment, unemployment, surface area, households, businesses, and other criteria.

Even the same data-set converted between the same areas will change its distribution over time – for example, due to the demolition or construction of housing; thus conversion tables should be time-dated for appropriate use. The 1991 and 2001 Censuses will in England and Wales be released for very different building-block areas, whose boundaries overlap. It will be sensible to convert both the 1991 and 2001 data to a single geography for comparison, but a single weighting criterion will not reflect changes in the distribution of characteristics. If a conversion table related to the household distribution in 2001 is used for 1991 data, then an area where many houses had been built during the 1990s will be over-allocated 1991 data.

If taken very seriously, the search for appropriate weighting criteria would develop principles such as:

- Each conversion table between any pair of geographies should contain alternative weights to suit the likely distributions of a range of different data.
- Each new geography implies new conversion tables numbering twice the number of existing geographies, once with each existing geography as source and once with each as target.

A new geography in this context may be an update of an old geography, and these may occur on at least an annual basis.

Such a potentially vast expansion of conversion tables demands some rationalisation and prioritisation. This will in practice be achieved by monitoring the usage of current conversion tables and the demand for new ones. As most conversion tables will change little over time and between alternative weights, it will be possible to flag where changes have taken place or where different kinds of data will be geographically heterogeneous, on a more limited series of less frequently updated core

conversion tables. The development of appropriate measures of quality will be more useful than a proliferation of conversion tables. Core conversion tables will include those between geographies of census, administrative and electoral regimes.

In the UK, the Economic and Social Research Council has recently awarded resources for the development of geography conversion tables to the University of Manchester. The UK Government also aims to resource geo-referencing tools to facilitate data conversion between geographies (National Strategy for Neighbourhood Renewal, 2000: Annex G).

## DISCUSSION

This paper has presented a conceptual framework to implement and to assess conversion of data between different geographical schemes. Such conversion is now commonplace as large data-sets become available on a regular basis, geo-referenced to small areas, from successive censuses and from administrative databases. Social and demographic research often needs to combine these data to make inferences about social and geographical relationships, to monitor social and economic profiles for local areas, and to maintain time series in the face of changing boundaries.

For schemes of neighbourhood statistics, data conversion from a complete coverage of small 'building block' areas offers the possibility of estimating profiles for areas with boundaries that were not foreseen by the collectors or the distributors of the data. Demand for statistics on new target areas is a characteristic of government policy to fund neighbourhood renewal through competitive bids, which must be justified, and if successful monitored through local evidence of social conditions.

By attaching a variable weight between 0 and 1 to each combination of source unit and target unit, this paper has emphasised that data for a source unit may be apportioned to more than one target unit which its boundary overlaps. But is such sharing a good idea? If the anonymised data held for building-block areas are small enough – whether census output areas, grid squares, or based on postal codes – will the aggregation of whole units be a

sufficiently accurate approximation to new target areas for all reasonable purposes?

In favour of aggregating data from whole source units, one might argue that the resulting statistics are not subject to any estimation error, and that there is no arbitrary choice of weighting criterion. Both these claims are less substantial than they may seem. If the end user's interest is in a specific target area or areas, then while the statistics for the aggregate of building block areas are exact counts, when taken as a proxy for the user's area of interest they are estimates of the true counts. The approximation is one of boundaries rather than directly of estimation, but the impact on the user is the same. And as the example of Table 4 shows, the error from approximating by boundaries of whole source units is likely to be considerably more than the error when estimating from a weighted sum of source units.

When aggregating whole source units, some rule must be used to decide which source units' data will be aggregated to approximate the target unit. If all source units that touch on the target are included, there will always be over-estimation and often this will be considerable. If some calculation is made based on a centroid or surface area or some other rule to identify the source units which are mainly within the target and shall be included, then there is a choice of criterion which is as unrelated to the data being converted as when a weighted sum of data from source units are used.

Thus the arguments in favour of aggregating whole source units are not strong. If a weighting criterion is available, then data conversion with weights is preferable. The main drawback of using weighted source units to convert data to a target area is the possibility that data are distributed within each source unit very differently from the weighting criterion.

If individual grid-referenced data exist and their confidentiality is at issue, then aggregation to small areas to make them safe prior to public data conversion to other areas is one avenue that some public data providers have followed. A different approach would be to perturb the individual data-points – either by smoothing with nearby points, or by perturbing the grid references themselves, thus main-

taining as much geographical heterogeneity as possible.

Measures of the quality of data conversion, however that conversion is achieved, are important for systems of neighbourhood statistics. This paper has highlighted the degree of fit between source units and each target unit, and the assessment and documentation of each conversion table. Additionally, measurement of the heterogeneity of common data-sets with respect to common weighting criteria such as resident households and surface area will boost the confidence of administrators and users of neighbourhood statistics. Where data are held for individuals separate from the anonymised building bricks of a public system, it would be possible to forewarn users of source units containing great heterogeneity. The more information available to the user, the more the user can bring intelligent interpretation to the estimated data.

One can make the general observation that the smaller the source units relative to the target units, then the less acute the above problems will be. The proportion of source units wholly within the target is high, the degree of fit is high, and therefore the extent of estimation whether by whole units or by weighted parts of units is low.

Maintenance of geography conversion tables is also an important issue that needs to be dealt with strategically. As discussed in the previous section, the maintenance of one well-documented up-to-date conversion table between each pair of key geographies is more important than an attempt to include a wide variety of weighting criteria. The maintenance of an accessible interface to data conversion is more important than the production of multiple tables, if they are not available to the wide community of social scientists. Developments in the UK are very positively moving towards widely accessible tools of geo-referencing and data conversion.

#### ACKNOWLEDGEMENTS

The work reported was mainly funded by the UK Economic and Social Research Council award H507255164 'Updated UK Area Masterfiles', which is fully documented on the website <http://convert.mimas.ac.uk>. Some of

the analyses reported here were undertaken variously by An Yu, Dan Abbott, Lou Daley, David Avenell and Kevin Kuscyk. The Office for National Statistics and the University of Manchester MIMAS service responded to many requests for help and information.

## REFERENCES

- Dorling D, Atkins D. 1995. *Population Density, Change and Concentration in Great Britain 1971, 1981 and 1991*. Studies on Medical and Population Subjects, 58. HMSO: London.
- Dorling D, Simpson S. 2001. The geography of poverty, a political map of poverty under New Labour. *New Economy* 8: 87–91.
- Ghosh M, Rao JN. 1994. Small area estimation: an appraisal. *Statistical Science* 9: 55–93.
- London Research Centre 1999. *SASPAC Manual*. LRC: London.
- National Strategy for Neighbourhood Renewal 2000. *Report of Policy Action Team 18: Better Information*. Stationery Office: London.
- Noble M, Penhale B, Smith G, Wright G, Dibben C, Owen T, Lloyd M. 2000. *Measuring Multiple Deprivation at the Small Area Level: The Indices of Deprivation 2000*. DETR: London.
- ONS Geography 1995. *The ED-Postcode Link*. Office for National Statistics: Titchfield.
- ONS Geography 2000. *All Field Postcode Directory 2000/1 Version Notes*. Office for National Statistics: Titchfield.
- Simpson L. 2001. *Updated UK Area Masterfiles: Final Report to the UK Economic and Social Research Council*. <http://convert.mimas.ac.uk> [2001]. Centre for Census and Survey Research, University of Manchester: Manchester.
- Simpson L, Yu A. 2001. On-line data conversion between geographies, with multiple look up tables derived from a postal directory. Working paper, Centre for Census and Survey Research, University of Manchester: Manchester.
- Skinner CJ. 1993. The Use of Synthetic Estimation Techniques to Produce Small Area Estimates. New Methodology series 18. Social Survey Division, Office of Population Censuses and Surveys: London.
- Southall H. 2001. *Conversion Factors for 1931 Census and 1939 National Registration Geographies*. University of Portsmouth Department of Geography: Portsmouth.
- Thomasson E. 2000. Small area statistics online. *BURISA* 144: 2–9.
- Westlake A. 2000. The introduction of formal structure into the processing of statistical summary data. In *Proceedings in Computational Statistics 2000*, Bethlehem JG, van de Heijden PGM (eds). Physica-Verlag: Heidelberg; 493–498.
- Wilson T, Rees P. 1999. Linking 1991 population statistics to the 1998 local government geography of Great Britain. *Population Trends* 97 (Autumn): 37–45.