

Updated UK Area Masterfiles

Full report of research activities and results

University of Manchester October 1999 – March 2001 ESRC award H507255164
Director: Ludi Simpson. Research assistant: An Yu. Other research and support: David Avenell, Dan Abbott, Mike Noble. Website: <http://convert.mimas.ac.uk>

1. Background

2. Objectives

3. Methods

3.1 All Fields Postcode Directory: validation and cleaning

3.2 Geography conversion tables: terminology

3.3 Construction of geography conversion tables without GIS

3.4 Construction of geography conversion tables with GIS

3.5 Assessment of the quality of geography conversion tables

3.6 Data conversion

3.7 Database and website construction

4. Results

4.1 Validation of the AFPD postcode directory; its use in creating geography conversion tables

4.1.1 Validation of the AFPD postcode directory

4.1.2 Production of geography conversion tables from the AFPD

4.2 Other geography conversion tables: comparison with those derived from the AFPD

4.2.1 Comparison with tables based on part postcodes

4.2.2 Comparison with table produced using GIS: Census 1991 ED to ward 1998

4.3 Data conversion

4.3.1 Best fit whole allocation rather than weighted allocation

4.3.2 Incorrect allocation of the weights

4.3.3 The source units are not wholly contained within single target units

4.3.4 The weighting criterion is not correlated to the data (synthetic estimation within source units).

5. Activities

6. Outputs

7. Impacts

8. Future research priorities

References

1. Background

Estimation of social statistics for small areas is a high priority for central government (National Strategy for Neighbourhood Renewal, 2000). This priority is stimulating many 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 project was devoted to the creation and study of geography conversion tables – or ‘look-up tables’. Their main purposes are to:

- Aggregate data to units sufficiently large to provide reliable results (for example from postcoded events to local authority Districts)
- Present results for areas that are familiar to the audience for the research (for example from small census units to current parliamentary constituencies).
- Estimate a time series on a consistent basis (for example electoral data from wards before and after boundary changes).
- Merge data sets drawn from different sources (for example for neighbourhood profiles containing both census data and data based on postal geography).

2. Objectives

As reproduced here from the proposal with annotations, the aim and objectives were met in full:

“Main aim: to make available database look-up tables allowing conversion between UK census, postcode and administrative geography.”

UK geography conversion tables have been made available. As detailed below, the work has extended considerably beyond the main aim.

Objectives:

- to manipulate existing statistical service postcode directories into several look-up tables.

Postcode directories for Scotland and Northern Ireland from GRO(S) and NISRA have been merged with the ONS All Fields Postcode Directory (AFPD). Together they have provided 200 UK geography conversion tables between census, electoral, administrative, postal, health and statistical geographies (see Section 3). These manipulations are stored procedures that can be used with future postcode directories.

- to measure and improve the quality of the directories through GIS and other approaches.

Measures have been developed to describe and assess the quality of geographical conversion tables, separately addressing the quality of the design of a table, the quality of the data used to construct the table and the quality of the small area estimates based on the table’s use (see section 3).

For England as a whole and for five Local Authority Districts experiencing different degrees of boundary changes to electoral wards since 1991, tables converting from census ED 1991 to ward 1998 boundaries have been produced using two different GIS techniques based on Ordnance Survey products Addresspoint and Boundaryline (section 3.4). The relatively poor quality of census codes on the England whole-postcode directories has been highlighted. The

geography conversion tables involving 1991 census EDs have been much improved using the GIS work (see Section 4).

- to mount those look-up tables on the Manchester academic data service MIDAS.

All 200 conversion tables mentioned above are available to academic researchers on the Manchester academic data service, renamed MIMAS since the proposal.

- to provide a web-based front-end to make access to the look-up tables easy for a range of academic researchers using census and other data.

<http://convert.mimas.ac.uk> gives access to all the geography conversion tables to all academic researchers registered for use of census data. The website also allows users to submit their own source datasets for on-line immediate conversion to a user-specified target geography. The site implements database routines of matching, weighting and aggregation such that users have no need for the database or programming resources that have limited this work in the past. With permission from the statistical agencies and the DETR, all the website routines are available also to non-academic users, except those that involve postcode units as source or target geography. The website has been adopted by MIMAS, ensuring its future maintenance. The site contains full documentation of the project and its products.

3. Methods

3.1 All Fields Postcode Directory: validation and cleaning

The All Fields Postcode Directory (AFPD) is a list of all postcodes in the UK. It is compiled twice yearly by the Office for National Statistics (ONS) from information supplied by the Ordnance Survey, the Royal Mail, Boundary Commissions, the General Registrar's Office for Scotland and the Northern Ireland Statistics and Research Agency. Each record includes reference to the number of residential addresses in that postcode, a code for each of several geographies from which those listed below with an example code can be derived. The code indicates the unit of that geography that the postcode mainly lies within. Some indicators of the quality of the information and the type of postcode are included.

Census ED/OA 91 (EGFA01)	
Census wardlevel 91 (EGFA)	Postal district 99 (BD1)
Census District 91 (EG)	Postal sector 99 (BD1 5)
	Postcode99 (BD1 5DT)
County 98 (47)	
Ward 98 (JAMA).	Primary Care Group 99 (4AA01)
Local Authority district 98 (JA)	Health Authority 98 (QAA)
Parliamentary constituency 97 (001)	NHS regional office 99
European electoral region 98 (01)	
	Standard Statistical Region 98 (1)
Country 98 (044)	Travel To Work Area 98 (001)
Gov. office for the region 98 (A)	NUTS-1 98 (UK4)
Local Education Authority 98 (201)	NUTS-2 98 (UK41)
Training Enterprise Council/LEC 99 (001)	NUTS-3 98 (UK411)

NUTS-4 98 (UK41100)

NUTS-5 98 (UK41100001)

Procedures were written to:

- a) Read the AFPD into a SQLserver database.
- b) Identify, describe and remove duplicate postcode records
- c) Add census codes for Scotland and Northern Ireland, identifying mismatches between the AFPD and the files supplied by GRO(S) and NISRA.
- d) Validate the contents of the files against the format and code lists supplied with the AFPD by ONS.
- e) Summarise the extent and nature of missing values on the AFPD.
- f) Export the validated UK postcode directory without duplicate records and with census information for Scotland and Northern Ireland, to a fully documented SPSS system file.

The procedures are stored and documented for future use. At the time of this work, ONS had no information about the file beyond its layout and code lists. The project reported its description of the file derivation and its contents to the first census programme workshop, along with the validation results.

3.2 Geography conversion tables: terminology

The project defines a *geography* as a set of non-overlapping units of geography, and a *geography conversion table* as a list of records, each including three fields to identify:

- (s) A source geography unit, for example a code for a 1998 electoral ward.
- (t) A target geography unit, for example a code for a postal district.
- (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*.

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

The weighting criterion may be the 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 sum 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 1991 Census EDs are all contained within 1991 Census wards). The percentage of source units with weight equal to one is the *degree of hierarchy*. It may be calculated as at the right.

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

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 90% within a single target unit, the *degree of fit* sums the maximum weight among records for each source unit, expressed as a

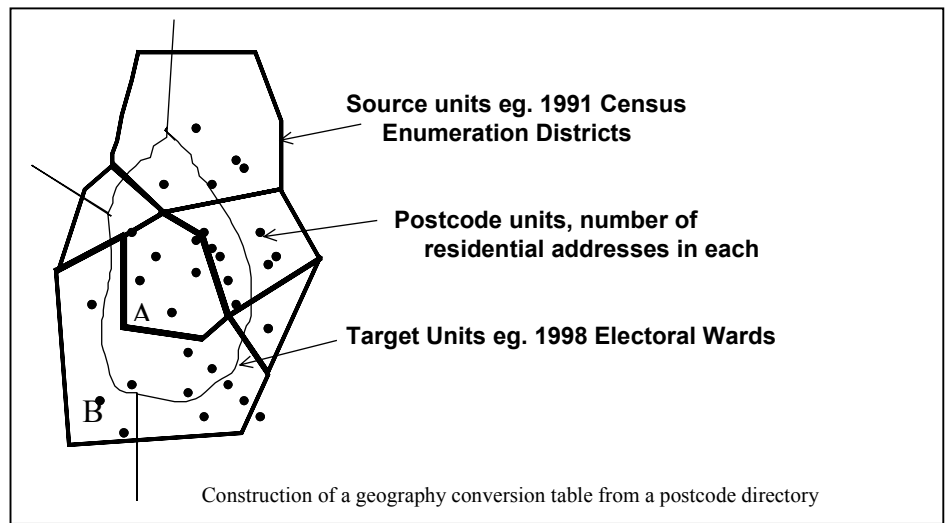
percentage of all source units. The degree of fit can be calculated for the whole table as at the right, or for a single target unit when the calculation is simply the sum of the weights with target t divided by the number of source units with target t. The degree of fit is an important measure of uncertainty when tables are used to convert data from source to target geography (see sections 3.6 and 4).

$$\frac{\sum_s (\max_t w_{st})}{\sum_{s,t} (w_{st})}$$

3.3 Construction of geography conversion tables without GIS

The project's main geography conversion tables were constructed from the AFPD. Within Great Britain and Northern Ireland, each geography is present for more than 99% of ordinary residential postcodes, with the exception of Primary Care Group codes for which the coverage is 96.3%.

Postcodes were used as an intermediary to construct geography conversion tables as indicated in the diagram. Where the AFPD indicates that all the postcodes in a source geography unit lie within one target geography unit (A), 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 (B), the AFPD provides a weight based on the number of residential addresses in the overlap of the source unit with target geography units.

A single procedure (stored and documented) creates a geography conversion table for given source and target as follows:

- a) Select all current postcodes with non-blank source and target, which are geographic (not post boxes) and not large users (as defined by the Royal Mail), and have at least one residential address.
- b) Group by source code summing the total residential addresses in the source unit.
- c) Group by source and target codes, summing the number of residential addresses in each overlap.
- d) Output the conversion table: source, target, residential addresses, weight (= residential addresses as a proportion of the total in the source unit).

As an exception, when postcode is the source geography, all postcodes with non-blank target are used, with weight 1.

The same logic could be used to construct a conversion tables from a database not of postcodes but of properties (or any other items) which contains on each record codes for the source and target geography and a value for the weighting criterion.

The project used a further ONS postcode directory (the pc2ed file) to examine the allocation of parts of postcodes to 1991 Census units, and the quality gained compared to geography conversion tables derived from the AFPD as above using allocation of whole postcodes.

3.4 Construction of geography conversion tables with GIS

Digital boundaries (ie in electronic form) were obtained for the Census ED 1991 and electoral ward 1998 geographies. GIS has been used to produce geographical conversion tables by examining the overlap of the boundaries.

AREA file. Mike Noble of Oxford University supplied a file in which the hectarage of each overlap was the weighting criterion. The file (referred to as the AREA file) covers all of England and was derived after cleaning the OS Boundaryline products for a project commissioned earlier by DETR, who gave permission for it to be used and made available by this project.

5 Districts file. For five districts with varying degrees of change to ward boundaries between 1991 and 2001, the OS Addresspoint product was obtained from the local authorities concerned. Each whole postcode was allocated by the centroid of its addresses to ED91 and ward98 overlaps using the point in polygon functionality of GIS.

The spatial analysis within GIS ensures consistency in some aspects of conversion tables. In particular it ensures that target units associated with a source unit are adjacent to each other. This is not guaranteed when deriving conversion tables from a database. Beyond this, whether or not GIS is involved, the accuracy of the conversion tables depends on the quality of the data used including the digitised boundaries.

3.5 Assessment of the quality of geography conversion tables

Using the concepts above (section 3.2) the project described and compared conversion tables constructed from postcode directories and from GIS, with a variety of weighting criteria. The geography conversion tables may themselves be in error. One principal source or error arises when the source geography units are allocated to target units on a *best fit* basis, assuming hierarchy when this is not the case. Additionally or alternatively, error can arise from *misallocation* of geographical units due to clerical error, poor quality digitising of boundaries, or other errors in the construction of the table. One criterion for lack of serious errors, implemented in this project with GIS techniques, is that the target units to which a source unit is allocated should always be adjacent to one another, achieving *geographical integrity*.

The project used GIS to create a matrix showing adjacent wards for the whole of England.

These investigations indicated weaknesses in the conversion tables involving 1991 Census EDs. They led to an improvement of one frequently used table (section 4.2.2), and recommendations for further improvements.

3.6 Data conversion

In the project and its website, data for source geography units (D_s) are converted to target geography units (D_t) using the weights recorded in the conversion table:

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

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.

Regardless of the accuracy of the construction of the geography conversion table, data conversion will induce error in estimates to the extent that the weighting criterion is not highly correlated to the data being converted. While the estimation of data for target units is usually unbiased, the variation between target units is likely to be under-estimated.

The weighting criterion and data conversion are exactly equivalent to the ancillary variables in a synthetic estimator for small areas described in statistical texts (Ghosh and Rao, 1994).

When an independent data set represents the truth, the approximation involved in data conversion has been measured by the *mean and median absolute error* and the *mean and median absolute percentage error*.

3.7 Database and website construction

The database and website construction was undertaken and documented by An Yu. Latterly, the website was completed and ported to MIMAS with the support of Dan Abbot and Lou Daley. The AFPD and conversion tables are held within an MS SQL server database. The website functions as far as possible are composed of discrete modular parts to ease testing, maintenance and development.

The user interface is based on html and Cold Fusion. It accepts user choices and uploads files, resulting in a simple standard string of parameters.

The cgi (common gateway interface) component is a C program that takes the string, sets the appropriate queries for the database, and accepts results from the database.

The database server itself using several internal stored procedures does the main work.

Detailed documentation and copies of the routines are held by the project website manager at MIMAS, Justin Hayes.

4. Results

4.1 Validation of the AFPD postcode directory; its use in creating geography conversion tables

Summary: 200 UK-wide conversion tables from the cleaned AFPD are sufficiently comprehensive to use, though with warnings about the lack of coverage of Primary Care Groups in some areas. Quality in later versions of the AFPD is likely to be improved due to the use of Ordnance Survey's 'Gridlink' (see below).

4.1.1 Validation of the AFPD postcode directory

The AFPD version 1999b contains 2,131,631 records. The project:

- Removed 345 duplicate postcode records.
- Filled gaps in Census geography with data from GRO(S) and NISRA.
- Documented the resulting 'cleaned' AFPD is documented as an SPSS system file available via MIMAS.

The AFPD is unique as a directory of all UK postcodes including those historically but not currently used, with a wide range of geography indicators. The allocation of geography codes is achieved by a variety of means, resulting in variable completeness and quality.

The geography conversion tables are based on ordinary residential postcodes, that is excluding postcodes characterised by the Royal Mail as 'large users' (usually large organisations), PO Box postcodes, or otherwise non-geographic.

The table overleaf shows the completeness of coding for the remaining 1,491,318 postcodes in Great Britain and Northern Ireland, taken from the file as cleaned by the project. Geography codes are generally not provided for the postcodes listed from the Isle of Man and the Channel Islands.

No field on the AFPD is complete except the allocation to a country. The incompleteness however is not systematic. It is spread geographically and is concentrated neither in particular postcode records (except see Primary Care Group below), nor in specific geographical areas. Two geography fields merit further warnings however.

Primary Care Group 99

3.71% of postcodes current in 1999 did not have PCG recorded. Shropshire, West Sussex, Cambridgeshire and Berkshire Health Authorities are most affected. The PCG codes were allocated to postcodes by ONS using look-up tables provided by Health Authorities, variously based on current wards, 1991 Census wards and EDs, or postcode sectors.

Completeness of coding for ordinary postcodes in Great Britain and N Ireland

Currently used and terminated postcodes on the AFPD as cleaned by the project.

Geography as recorded on the AFPD	Current postcodes		Terminated		All postcodes	
	Not empty	% empty	Not empty	% empty	Not empty	% empty
Easting	1,486,402	0.33%	178,918	0.49%	1,665,320	0.35%
Northing	1,486,402	0.33%	178,918	0.49%	1,665,320	0.35%
Census 1991 ED	1,478,620	0.85%	137,609	23.47%	1,478,763	11.51%
County ¹ 98	680,480	54.37%	52,620	70.74%	733,100	56.13%
Local Authority district 98	1,489,440	0.13%	179,670	0.08%	1,669,110	0.12%
Ward 98	1,489,023	0.15%	179,498	0.17%	1,668,521	0.16%
Parliamentary Constituency 97	1,488,730	0.17%	179,451	0.20%	1,668,181	0.18%
European electoral region 98	1,488,730	0.17%	179,451	0.20%	1,668,181	0.18%
Country	1,491,318	0.00%	179,808	0.00%	1,671,126	0.00%
Govt Office for the Region 98	1,488,836	0.17%	179,496	0.17%	1,668,332	0.17%
Local Education Authority 98	1,488,836	0.17%	179,496	0.17%	1,668,332	0.17%
Training & Enterprise Council ² (TEC or LEC) 99	1,446,905	2.98%	176,053	2.09%	1,622,958	2.88%
NHS Region 99	1,490,468	0.06%	179,783	0.01%	1,670,251	0.05%
Health Authority, Health Board 99	1,490,233	0.07%	179,704	0.06%	1,669,937	0.07%
Primary Care Group 99	1,435,993	3.71%	167,132	7.05%	1,603,125	4.07%
Standard Statistical Region 98	1,310,474	12.13%	149,813	16.68%	1,460,287	12.62%
Travel to Work Area ³ 98	1,487,355	0.27%	178,964	0.47%	1,666,319	0.29%
NUTS 98	1,488,835	0.17%	179,495	0.17%	1,668,330	0.17%

Notes:

¹ County is only applicable in English two-tier areas, where it is missing for less than 0.5% of current postcodes.² TEC/LEC is not applicable in Northern Ireland. Elsewhere it is missing for less than 0.5% of current postcodes.³ TTWA is not applicable in Scotland or Northern Ireland. Elsewhere it is missing for less than 0.5% of current postcodes.*1991 Census ED*

Postcodes were allocated to 1991 Census ED by census offices using the Census database, but to other geographies manually, or by methods using the centroid of the postcode's addresses, or the first address in the postcode. The difference in allocation methods causes some inconsistency; for example a postcode may be allocated to a different census ward from the current ward even when ward boundaries have not changed. Postcodes introduced after the census were often allocated in England and Wales to a census ED by imputation, with significant local error. A significant number of 1991 EDs are not represented by any of the AFPD's current postcodes. Finally, because EDs are the smallest geography coded on the AFPD, there is most approximation in the allocation of each postcode wholly to a single ED. The project investigated the extent of these approximations and improved the ED allocation to current wards (section 4.2.2).

The project's first report described the cleaning and validation of the AFPD in detail (Yu and Simpson, 2000).

Gridlink is a new means of ensuring that the AFPD allocates postcodes consistently to electoral and administrative boundaries, using GIS point-in-polygon methods to place the

centroid of the digitised addresses in each postcode into areas defined by digitised boundaries (May et al., 2001). This positive development used from the 2000b version of the AFPD will not however improve the allocation of 1991 Census codes.

4.1.2 Production of geography conversion tables from the AFPD

200 Geography conversion tables were produced from the cleaned AFPD using the method described in section 3.3. Eight source geographies (1991 Census ED/OA or ward, 1998 ward or district, 1999 Postcode unit – current or all postcodes, 1999 Postal sector, 1999 Postal district) are tabulated against each of the 25 target geographies listed in section 3.1. The table below summarises a selection of them.

Selected UK geography conversion tables derived from the AFPD

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 section 3.2.

The degree of hierarchy is only 100% for source geographies that lie entirely within the units of the target geography, when there is a one-to-one lookup between geography units. Postcodes fit hierarchically in this way within all other geographies, due to the construction of the AFPD from whole postcodes. The lack of exact hierarchical fit of Census areas to electoral areas is partly a result of the inconsistent methods of construction of the AFPD, which are further explored in section 4.2.2.

The degree of fit reflects units that do not lie wholly within one target unit but may nonetheless fit well within one target unit. It is high in most cases, showing the high degree of certainty involved in data conversion using these tables, even without strict hierarchy.

4.2 Other geography conversion tables: comparison with those derived from the AFPD

Summary: the two main weaknesses of the AFPD as a basis for geography conversion tables are:

- (a) the use of whole postcodes and
- (b) the lack of geographical integrity within the AFPD's allocation of postcodes to other geographies.

The project's investigation of these issues is reported here. The drawbacks are measurable and significant in specific situations. The project has shown how they can be overcome to a large extent, and that the drawbacks are not significant in most situations.

4.2.1 Comparison with tables based on part postcodes

A separate ONS 1991 census postcode directory (called here the pc2ed) provides the number of resident households in each intersection of postcode and census Enumeration District (ED) in England and Wales. It is based on a count from the Census database where every household record contains both postcode and ED code. From this directory nine conversion tables have been created, using the number of households as the weighting criterion, which are summarised in the table below.

Nine conversion tables between postal and 1991 census geography, using part postcodes

Degree of hierarchy and degree of fit

		ED e.g. EGFE01	Ward EGFE	District EG
Postcode unit e.g. PE1 7RU	Hierarchy:	78%	96%	100%
	Fit:	94%	99%	100%
Postal Sector PE1 7	Hierarchy:	2%	8%	66%
	Fit:	15%	68%	96%
Postal District PE1	Hierarchy:	0%	3%	38%
	Fit:	7%	41%	93%

Notes:

Degree of hierarchy: percentage of source units that lie wholly within a single target unit.

Degree of fit: Sum of maximum weight for each source unit, as a percentage of the number of source units.

Example codes are from Peterborough

While the AFPD allocates postcodes wholly to each census (and other) geography, in fact only 78% of postcodes lie wholly within a single 1991 ED, 96% wholly within a ward, and all (to the nearest 1%) lie within a District. However, most postcodes that overlap a census boundary lie mostly in one area, such that the degree of fit is much closer to 100%. The degree of fit when allocating whole postcodes to EDs is thus 94% – that is the proportion of households that is correctly allocated.

When using the tables to convert postcoded data to census geography some of the error introduced by using whole postcodes will balance out – as investigated in section 4.3. Generally, the addition of error due to the use of whole postcodes rather than part postcodes, in constructing conversion tables exists but is not great.

Similarly the table shows that conversion from postal sectors to electoral wards inevitably involves approximation, because only 8% of postal sectors are wholly contained in electoral wards. While most postal sectors lie mostly within a single ward, there is considerable uncertainty remaining, as shown by the degree of fit of 68%.

A recommendation of the project is that the most up-to-date part-postcode census directories are used to create the nine directories referred to in the table above, for use in the conversion website. Postcodes not in the part-postcode directory should be retained from the AFPD. GRO(S) and NISRA may supply part-postcode directories for Scotland and N. Ireland.

4.2.2 Comparison with table produced using GIS: Census 1991 ED to ward 1998

The imperfect quality of census-to-electoral conversion tables is quantified and is improved by combining with information from the 'AREA' table based on overlapping digitised boundaries.

The conversion tables between Census geography and current electoral boundaries are of great importance for social analysts, as they allow the use of 1991 census data together with political demographic and social data collected after electoral boundaries have changed. There have been and are expected to be major revisions to electoral boundaries between each census year in Britain.

The '5 Districts' file described in section 3.4, when compared to the AFPD for the same districts showed clearly some weaknesses of the AFPD. The AFPD lacks geographical integrity in that

- *A postcode can be allocated to a different 1991 ward (by the census code) and current (1998) ward, even where the boundary had not changed since 1991.* This is due to the inconsistent procedures for allocating postcodes to census and electoral geographies, where there is more than one option. For example current ward allocation was often made by placing the first address in the postcode visually using maps; the allocation of census code had been made from the census database, by identifying the census area in which most of the postcode's households were located. After *Gridlink*, the allocation methods will remain inconsistent.
- *The goodness of fit of 1991 EDs to current wards is under-estimated.* This is a consequence of the inconsistency of coding methods just described, but also of incorrect allocation of postcodes introduced since 1991.
- *Postcodes allocated to the same ED can be allocated to different and non-contiguous 1998 wards.* While not a frequent occurrence it clearly involves an error in allocation of the code for either ED or ward or both.

ED91toWard98 conversion table

Based on current ordinary postcodes in the AFPD version 1999b as cleaned by the project.
Degree of hierarchy and degree of fit, each country

	England	Wales	Scotland	Northern Ireland	All GB and NI
Number of records	129,101	8,312	41,718	4,271	183,402
Number of 1991 EDs	103,090	6,378	37,811	3,630	150,909
Number of 1998 wards	8,462	877	1,246	572	11,157
Degree of hierarchy	78.5%	75.0%	90.6%	83.6%	81.5%
Degree of fit	97.4%	97.3%	97.6%	96.2%	97.4%
Ordinary EDs¹ not included	222	13	293	99	627

¹EDs with at least 16 households and 50 residents.

The vast majority of the 155,448 1991 ED/OA census areas in Great Britain and Northern Ireland are represented by postcodes on the AFPD. Most of those that are not had zero population or were had so small a population that they were partially suppressed in the census itself. Among the 627 remaining ordinary ED/OAs not included as shown above, many represent housing demolished since 1991 – the project’s tables are based on current postcode locations. However 129 missing ordinary ED/OAs are neither represented by terminated postcodes – in existence at the time of the 1991 census. A list of these has been reported to ONS.

That England and Wales have relatively poor fits of EDs to the current ward geography is not itself proof of poorer quality, as it is possible that the re-organisation of local boundaries during the 1990s affected this relationship more in England and Wales. However, the equivalent AREA file for England has allowed a closer examination:

The AREA file is a conversion from 1991 EDs to 1998 electoral wards, based on digitised boundaries. The OS Boundaryline datasets of ED and ward boundaries was cleaned (for another project, see section 3.4) by removing ‘slices’ or ‘splinters’ between wards in their intersection with EDs, where no population could be involved, and then a conversion between the two sets of boundaries was based on the weighting criterion of hectarage in each remaining real intersection.

Hectarage is not the appropriate weighting criterion for most conversion of social data, because large tracts of land may contain few or no residents. However, the AREA file accurately identifies EDs which are wholly contained in one ward, and identifies which wards an ED overlaps if it is not contained wholly within one ward.

New ED91toWard98 records for England have therefore been created using the following logic:

- (i) Use records for all EDs that are recorded on only one of the AREA or AFPD files.
- (ii) If the AREA file indicates an ED wholly within one 1998 ward, use this record and reject the AFPD records.
- (iii) If the AREA file indicates an ED split between 1998 wards, use the AFPD records for these same wards, and recalculate the weights.

- (iv) If there are no AFPD records for any of the wards indicated by AREA as overlapping and ED, use the AREA records.
- (v) In each case recalculate the weight using the number of addresses indicated by AFPD. Only where this is not possible (as in case iv) use the AREA weight based on hectarge.

Comparison of AFPD and AREA files, and the improved file combining their best features, England

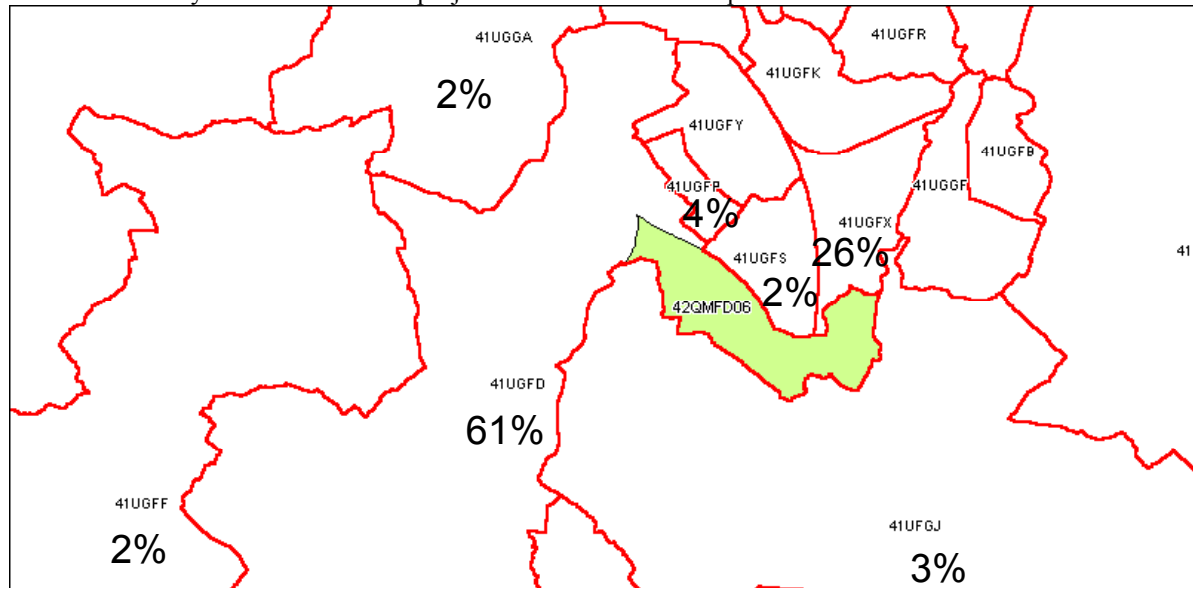
	AFPD	AREA	Combined
Number of records	129,101	104,430	104,400
Number of 1991 EDs	103,090	103,168	103,725
Number of 1998 wards	8,462	8,442	8,442
Degree of hierarchy	78.5%	98.8%	99.4%
Degree of fit	97.4%	99.8%	99.9%

The AFPD tendency to allocate 1991 EDs to too many 1998 wards is shown by its greater number of records and its lower degree of hierarchy: less than 80% of England EDs fit wholly into a 1998 ward according to the AFPD, while over 98% do so in fact. However, this ‘spreading too far’ of EDs by the AFPD is also shown to be fuzziness at the edges rather than divisions of EDs down the middle, in that the degree of fit is 97.4%.

An extreme case of the AFPD incorrect spreading of EDs to adjacent wards is shown in the figure below. The AREA file’s allocation to a single unchanged ward has been verified with the local authority. Much of the incorrect placing of EDs is at the boundary of wards and districts, especially where the shape of the electoral units is unusually irregular.

Fuzziness in the census to electoral conversion table derived from the AFPD: an extreme case.

Stafford 1991 Census ED 42QMFD03 is entirely contained within 1998 ward 41UGFD, but is placed by the AFPD also in six adjacent wards (with weight shown as percentages). The correct allocation as confirmed by the local authority is contained on the project’s conversion table improved with information from AREA file.



The combined file is an improvement in that it:

- includes 635 EDs which were not on the AFPD file;
- has greater degrees of hierarchy and fit than either file: the AFPD fuzziness has been eliminated as well as the overlaps of the AREA file that contain no population;
- excludes 20 Welsh wards wrongly coded on the AFPD file as overlapping English EDs on the boundary between the two countries.

This discussion has taken the AREA file to be correct in its construction. Apart from the construction of the file preventing gross errors as described in section 3.4, its correctness has been further verified in six cases where the two files allocated EDs to completely different wards, by checking with the local authorities concerned.

The improved records have been substituted for the records for England in the ED91 to Ward98 conversion table derived from the AFPD. The improved table is used by the website for data conversion.

It is recommended that from this improved conversion table, further improved tables are derived between other census and electoral geographies that are aggregations of 1991 EDs and 1998 wards. The errors in the AFPD are not so great for these larger areas, and therefore this work was not of such high priority for the project.

4.3 Data conversion

The project and its website (see section 6) have made quick and relatively accurate data conversion for non-technicians a reality. The website's function to convert data has been widely used already by social scientists who no longer need specialist database and programming skills.

The conversion of data from geographical areas rather than from points requires conversion tables with non-unity weights such as those produced by this project and entails some inaccuracy. Conversion with no estimation is only possible when data are accurately coded to dwelling units or other points, which can then be aggregated hierarchically to any target geography. This is one goal of the government's current National Geo-referencing Strategy. This strategy will not eliminate the need to convert data that is only held for geographical areas as will continue for practical reasons and to prevent data disclosure.

This section takes a critical look at data conversion, categorising and quantifying with examples the approximation that is an inevitable consequence of converting data from one set of geographical units to another.

A variety of sampling and non-sampling errors may afflict the source data before data conversion. The additional errors from data conversion may or may not be relatively insignificant and should be assessed in that context.

There may be error arising independently from inaccurate construction of the conversion table and from data conversion itself. In each case the source of error can be usefully sub-categorised to identify four sources of error:

- Errors in the construction of the conversion table
 1. Best fit whole allocation rather than weighted allocation.
 2. Incorrect allocation 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 (synthetic estimation within divided source units).

There is no general level of error that should be expected, since it will depend on the geographies involved, the weighting criterion and its correlation with the data, and the construction of the conversion table. For this reason, the examples in this section must be illustrative rather than comprehensive. Section 6 returns 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 sometimes cancel out.

4.3.1 Best fit whole allocation rather than weighted allocation

To avoid estimated statistics, or because no other information 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 best fits by some stated criterion. The loss of accuracy in using a best-fit approach to estimating neighbourhood statistics has been tested for the development of an online community statistics system for the region of Bradford.

Estimating benefit claimants in 70 Neighbourhoods from whole source units and weighted source units

Source units: Census EDs in Bradford. Average 726 claimants in 2,477 residential addresses

Allocation	Absolute error	
	Maximum	Root Mean Square Error
Whole source units	245.0	73.5
Weighted source units	66.4	21.5

The true distribution of welfare claimants is known within the Bradford region to the Council which administrates 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 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. The average error, 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%. When

allocating EDs 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 amount of error added from using best fit allocation will be less when the source units are small compared 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 postcode to Census directory referred to in section 4.2.1. The use of postcodes to allocate survey and administrative data to other geographies is common. The directory records the actual number of enumerated households in each census area, as well as the number falling into each overlap of postcode and census geography. The recorded true number of households is compared in the table below with the number that would result by allocating census households in a postcodes wholly to the census area in which most of the postcode's households lie.

The median 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, because the errors balance each other for larger areas, the median error is hardly different for wards or for Districts in spite of their much longer boundaries over which wrong allocations can be made. The percentage errors are thus reduced far below 1% (calculated as the absolute error divided by the true recorded number of households).

Error induced by use of whole postcodes

Resident households allocated from whole postcodes to census areas

Census unit (target)	Median absolute error	Median absolute % error
ED	10	5.70%
Ward	8	0.50%
District	12	0.03%

Notes: Source: pc2ED part-postcode directory, England and Wales 1991

4.3.2 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, using the improved table constructed as in section 4.2.2, the table below measures the error in data conversion caused by wrong allocation of UK 1991 census geography (EDs) to 1998 electoral geography (wards) as derived from the AFPD.

The errors in the table derived from a postcode database occurred for a variety of reasons. While the use of whole postcodes contributed some error (as above), but the main problem lay in an inconsistency in the construction of the database. The errors in the database were not

clustered in a few areas but were a general phenomenon presenting itself as an allocation of source census units to too many electoral wards. Because both geographies are relatively small, errors in allocation are likely to have a relatively large impact. The error shown in the table is that which has been eliminated by the use of the improved equivalent table.

The error is considerable especially for the estimation of 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.

Error induced by incorrect allocation of weights

Resident households allocated from 1991 EDs to 1998 wards and Districts (England)

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

4.3.3 The source units are not wholly contained within single target units

Imagine there were 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 error in the data conversion to the extent that the source units do not lie wholly within single target units, and in these cases the error would depend on the correlation of the data to the weighting criterion.

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 in section 3.2. Examples were given in section 4.1.2.

The degree of hierarchy shows simply the proportion of source units that lie wholly within one target unit according to the conversion table (weight = 1). 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 edges, or amalgamations of more than one District together with minor other changes.

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; the remainder is subject to weighting and error according to the weights in the conversion table. Thus if a source unit is shared between 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 geography does not fit so well within electoral geography, with a degree of fit under 70%.

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. The table

below shows for England and Wales the degree of hierarchy and fit between postal geography and Census geography in 1991, based on the same part-postcode directory used in section 4.1 above. 1991 Census EDs were not based on postal geography (unlike in 2001). Only three quarters of postcode units fit wholly within a census ED.

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 section 2.

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 aggregate whole units but also estimated parts of source units. The degree of fit to a single target unit measures the reliability of each value output from data conversion.

4.3.4 The weighting criterion is not correlated to the data (synthetic estimation within source units).

Where the fit between a pair of geographies is not exact – and the previous section showed that this is usually the case – data conversion uses the conversion table weight to share data from source units to target units. The estimate for the 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 available for the intersections of source and target units.

However, one can closely simulate the heterogeneity of census variables within larger census units by the heterogeneity between Enumeration Districts. 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, and are used in the table below to identify the error involved in conversion of census characteristics from census wards to postal districts.

The table compares the approximate conversion from wards to Postal Districts with the more accurate conversion from the much smaller EDs. The differences are due to the imperfect correlation between the weighting criterion (the number of residential addresses) and the data as represented by variation between EDs inside each ward.

Error induced by data conversion due to synthetic estimation

Households and household characteristics.

Difference between estimates by allocation of 1991 EDs (taken as correct) and by allocation of 1991 wards.

(a) 2,142 postal districts in England and Wales

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)
		Resident households	9,220	46	0.6%
With pensioners	2,092	27	1.1%	1,942	1,932
Only pensioners	2,293	23	1.3%	1,462	1,454
Lone pensioner	1,385	16	1.4%	915	910
Lone parent	378	5	2.6%	354	349
Crowded households	45	1	6.8%	84	82

(a) 92 postal districts of West Midlands County

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)
		Resident households	9,900	115	1.0%
With pensioners	3,391	98	3.0%	2,072	2,037
Only pensioners	2,427	90	3.5%	1,509	1,483
Lone pensioner	1,508	52	2.9%	918	904
Lone parent	477	44	9.9%	350	334
Crowded households	55	6	18.4%	65	58

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

Source: 1991 Census SAS; ED91to Postal District 99 and ward91 to Postal District 99 conversions using tables based on the All Fields Postcode Directory.

The ward data was created as the sum of ED data, so that none of the discrepancy is due to census data modification.

The census number of households in a postal district is well estimated from data conversion from wards, because it is highly correlated to current the number of addresses on which the conversion weights are based. The median absolute percent error is only 0.6%. The number of pensioner households of different types is geographically heterogeneous within census wards – not correlated to the number of addresses in each area. They are thus less well estimated by data conversion, with percentage errors of over 1%. Relatively rare events that are 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 magnitudes of error 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 has considerable 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. 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 EDs then this 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 is available (as is the case for many 1991 Census data, and for non-census demographic and other data).

Exploration of the accuracy of data conversion is beyond the project brief, but will be expanded in later publications arising from the project. These will also look at overall error, eg by comparing postal estimates as released from census database with those created by data conversion from ED and ward datasets.

5. Activities

ESRC Census development programme presentations:

- Leeds (3-4 May 2000)
- London (1 November 2000)
- Leeds (15-16 May 2001).

Presentations to users:

- Office for National Statistics Central Postcode Directory Users' Group. London, September 20th 2000.
- Office for National Statistics Neighbourhood Statistics unit. October 10th 2000.
- Leeds University School of Geography seminar series. March 1st 2001.
- Manchester University Centre for Census and Survey Research seminar series, March 19th 2001.

Informal demonstrations at various venues.

6. Outputs

Substantial progress reports were written for the ESRC Census development programme workshops in May 2000, November 2000 and May 2001.

A paper on related topics referring to the project's work has been submitted to the Journal of the Royal Statistical Society series A (Simpson et al., Small area statistics on-line, submitted 2000). A further three papers will report the work of the project, tackling respectively the general theory of geography conversion tables and data conversion, the specific developments using the All Fields Postcode Directory, and a user's perspective on data conversion:

- 'Small area statistics online' Journal of the Royal Statistical Society Series A.
- 'Geography conversion tables'. International Journal of Population Geography.
- 'The UK All Fields Postcode Directory and an on-line facility for geography data conversion'. Computers, Environment and Urban Systems.
- 'Boundary free data and their use in social investigation'. Journal of Social Policy.

The stored routines for the validation, cleaning and use of the All Fields Postcode Directory are documented outputs that will assure future development of the site.

ESRC Data Archive have been offered the 200 geography conversion tables and are considering the format in which they wish to receive them.

The major output from the project is the web site for:

- creation and downloading of geography conversion tables
- data conversion using geography conversion tables.

Figure 2 below shows the main menu reached from <http://convert.mimas.ac.uk>. Figures 3 and 4 show two of the option's main screens, including the drop down list of geography units by

which the user chooses the source and target geographies for data conversion or downloading conversion tables. Figure 5 shows the log of a data conversion which is produced by the user to save.

When the user's source units are postcodes, the website routines convert the user's postcodes to a standard format. Any not matched with the site's conversion tables are imputed within a valid postal sector or postal District, to minimise data loss. The imputed postcodes are listed so that the user can correct or omit them should they wish to.

Figure 2: The project's geography conversion site – main menu

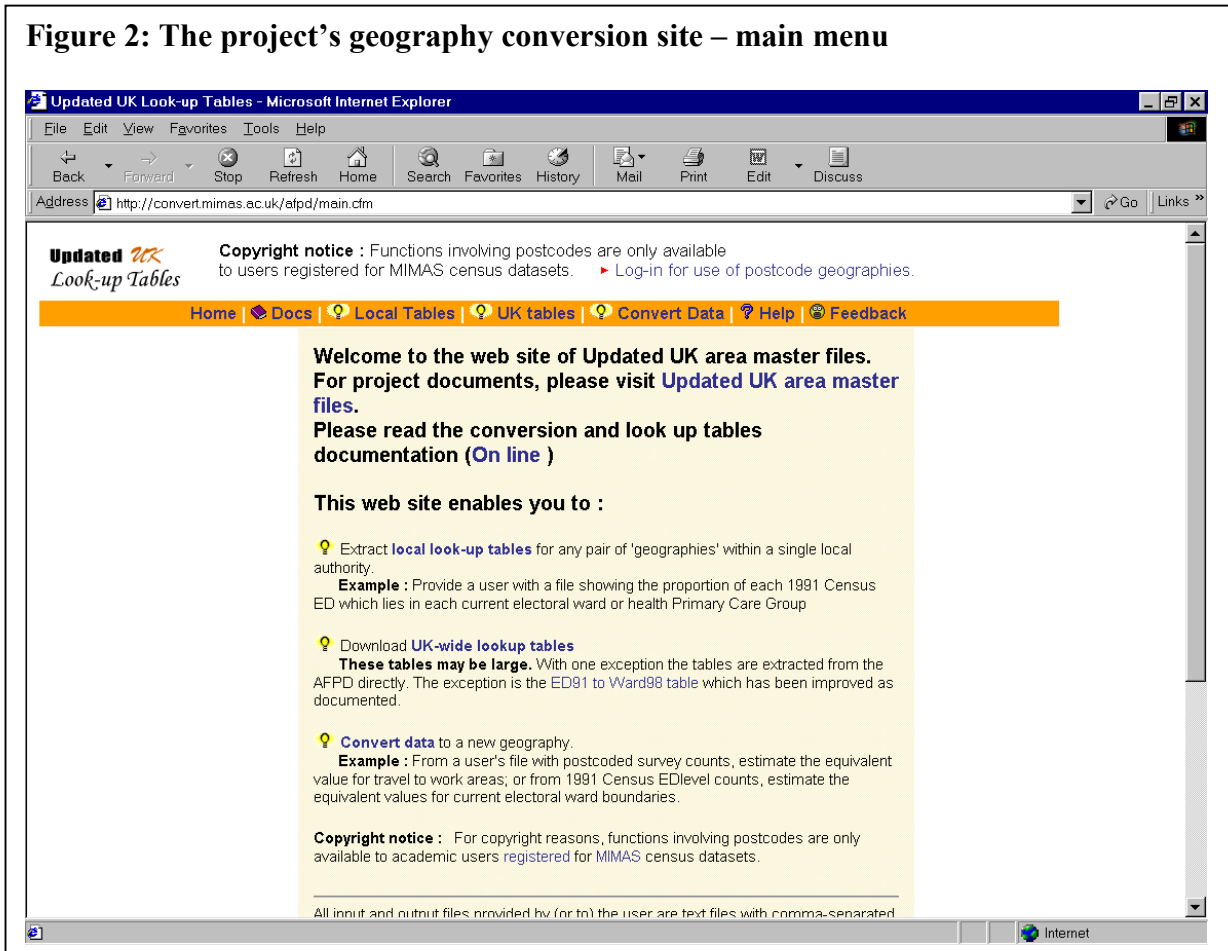


Figure 3: The project’s geography conversion site – choosing a lookup table to download

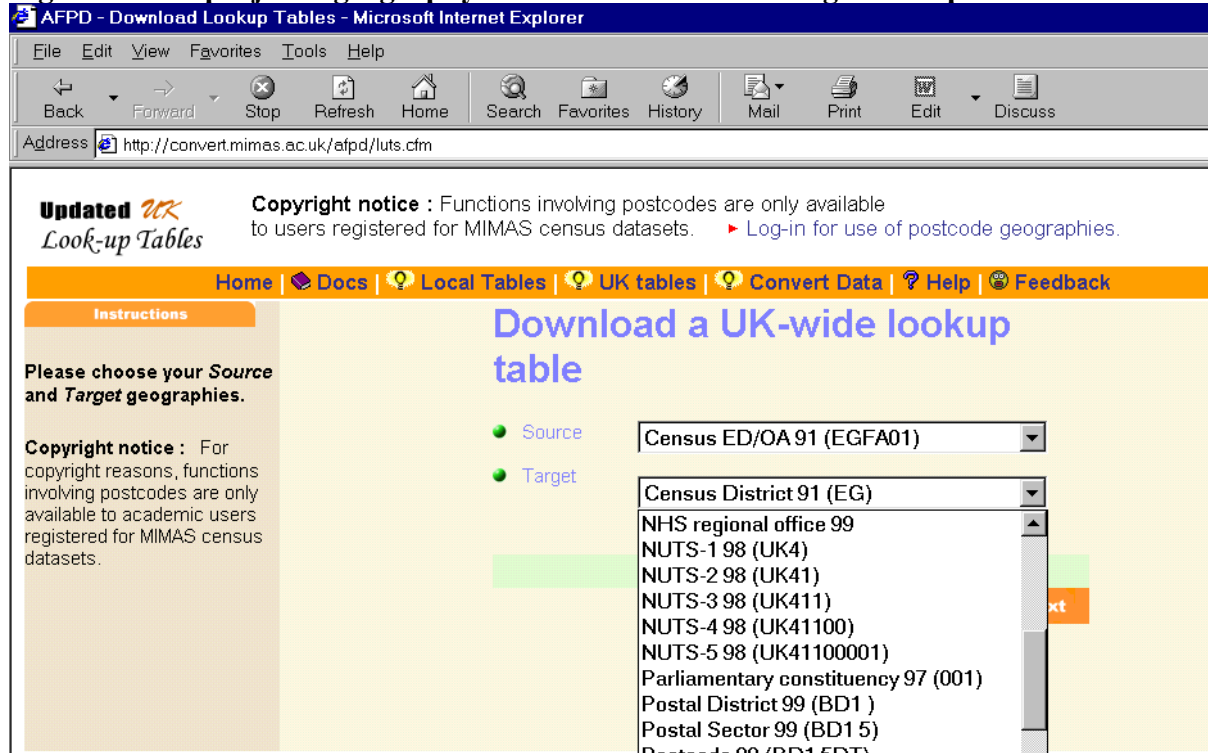


Figure 4: The project’s geography conversion site – converting data

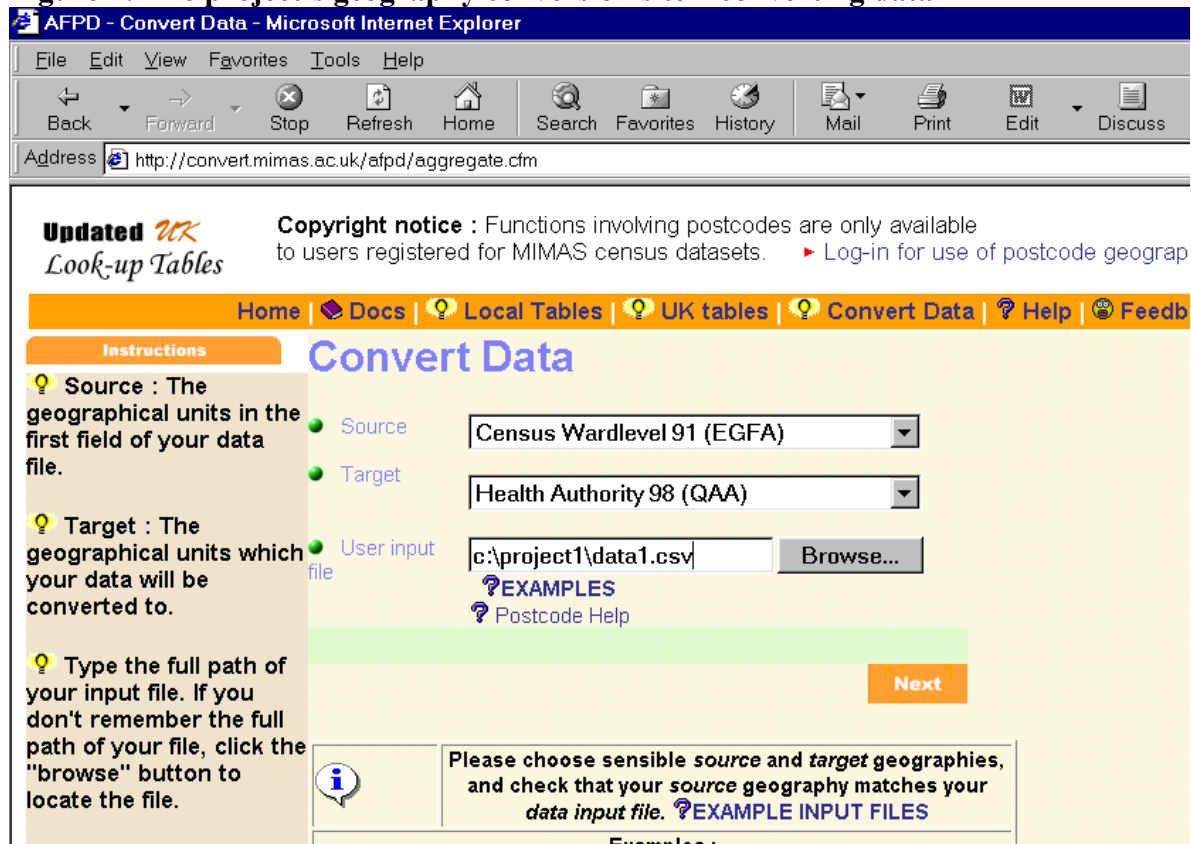


Figure 5: example log after data conversion

File: 'C:\uamf\outputs\pc5odd.csv'
 converted from 'Postcode 99 ALL' to 'Local Authority District 98 (09UC)'
 completed at Thu May 03 11:59:16 2001

Input records with source geography not fully matched:

Postcode	Degree of match	Imputed Postcode
BD21 4A2	2	BD214ER
BD21 9EU	3	BD213HQ
BD52 9SS	0	N/A
BD22	3	BD226EG

Summary: number of records, degree of match, data field sub-totals

1, Match 0 unable to use, 0,1,0,0,0
 14, Match 1 matched, 27,38,33,39,41
 1, Match 2 post unit imputed within valid sector, 0,1,0,0,4
 2, Match 3 post unit imputed within valid district, 1,2,1,1,1
 0, Match 4 post unit matched but imputed because no target geography on AFPD,
 0,0,0,0,0
 18, All records processed, 28,42,34,40,46

7. Impacts

The project website and its routines have been used already during the course of the project by many researchers, the data conversion routines being used many times each day. Uses have included:

Sociological research: Social capital and health outcomes (Kelvyn Jones, Steve Barnard, University of Portsmouth). Convert 1998 government indicators of deprivation to 1991 wards for use with census data.

Political science: Poverty and political representation (Danny Dorling, University of Leeds). Convert data for electoral wards with 1991 and 1998 boundaries to 1997 parliamentary constituencies.

Resource allocation: Deprivation measures (Margrethe Anderson, University of Manchester). Convert postcoded data for all London, for comparison with government Indices of Deprivation 2000.

Census dissemination and research: Estimation of census data for non-standard areas (Lou Daly, MIMAS, University of Manchester). Used project routines to enhance the functionality of CASWEB software. Allocation of 1999 census rehearsal data from 1998 wards to 1991 wards for comparison with 1991 census data (David Voas, Universities of Liverpool then Sheffield).

Demography: Convert 1991 Census area population estimates to 1998 wards (Paul Norman University of Leeds); allocate postcoded pupil records to 1991 census areas (Paul Williamson, University of Liverpool).

Higher education participation: aggregate postcoded student administrative records to larger census and other areas for which relevant denominators exist (Mark Corver, HEFCE).

8. Future research priorities

The demand for data conversion using geography conversion tables to aid research is increasing. The ESRC Census Programme centre responsible for developing look up tables' can build on the current project in a number of ways (section numbers are given where further information is given in the text of this report). Some of these are straightforward production work, others involve research into the quality of data and the priority needs of research users.

- Obtain the latest All-Fields Postcode Directory (2001a or later), with Gridlink allocation of postcodes, merging GRO(S) and NISRA census files, supplementing rather than replacing lookup tables on the current MIMAS site (4.1.1).
- Contribute to the specification of 2001 Census geography on the AFPD as soon as possible, and add this to the MIMAS data conversion facilities.
- Obtain subsequent All-Fields Postcode Directories on an annual basis.

- Continue to liaise with the UK statistical agencies to improve the allocation of postcodes to 1991 census EDs.
- Supplement the MIMAS site with other geography conversion tables, for example from Experian, and the Historical census projects.
- Derive from the improved ED91 toward 98 table other census-electoral tables (4.2.2).
- Derive improved postcode-census tables from the part-postcode directories already obtained by ESRC (4.2.1).
- Consider the production and appropriate use of varied weighting criteria: addresses, households, population, unemployment, hectareage (4.3).
- Consider the production of the same table for different years: for example 1991 census data with 1991 weights.
- Add functionality to the data conversion. In particular to add further geography codes to a user's file with one geography code, without converting data, would be straightforward.
- Add to the output from data conversion, statistics of the degree of fit for each target geography (4.3).
- Add a program interface to the functions of the datasite, so that data conversion may be carried out within other on-line research functions.

References

Ghosh, M and JN Rao (1994) Small area estimation: an appraisal, *Statistical Science* **9**: 55-93.

May, Keith, Peter Standen and Alan Taylor (2001) Gridlink – the new standard for postcode location data. *BURISA* **147** (April): pages 5-7.

National Strategy for Neighbourhood Renewal (2000) *Report of Policy Action Team 18: better information*. London: The Stationery Office.

<http://convert.mimas.ac.uk> (project website with links to further documentation).

Yu, An and Ludi Simpson (2000) The All-Fields Postcode Directory (AFPD): validation for use as a look-up table. Progress report to the ESRC Census Programme workshop Leeds 3-4 May.