



PERGAMON

Computers, Environment and Urban Systems
27 (2003) 283–307

Computers,
Environment and
Urban Systems

www.elsevier.com/locate/compenvurbsys

Public access to conversion of data between geographies, with multiple look up tables derived from a postal directory

Ludi Simpson^{a,*}, An Yu^b

^a*Centre for Census and Survey Research, University of Manchester, Manchester, UK*

^b*Department of Computer Science, University of Manchester, Manchester, UK*

Received 1 February 2002

Abstract

Comprehensive statistical analysis of local areas and variation between them demands that data held for a variety of geographical source units are made compatible. Geography conversion tables enable transfer of data from one set of geographical units to another, but have in the past been generated for specific projects, while their use has required database programming skills. The derivation of multiple geography conversion tables for public use is specified from the UK-wide 'All-Fields Postcode Directory' (AFPD). The results are validated using alternative constructions including geographical information systems. The accuracy of conversion tables is measured. Access is achieved through list-matching algorithms and a World Wide Web interface <http://convert.mimas.ac.uk/>, which allows the public user to convert their own data between postal, electoral, census, administrative and statistical geographical boundaries. The data conversion is improved by proportional allocation for those source units that do not fit within a single target unit. The architecture of the site allows further tables from any source to be incorporated. The current demand for computational geography to meet small area statistical requirements leads to a discussion of priorities for the improvement and maintenance of geography conversion tables.

© 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Consistent geography units; Geography conversion tables; Postcodes; Public access

* Corresponding author. Tel.: +44-161-275-4721; fax: +44-161-275-4722.

E-mail address: ludi.simpson@man.ac.uk (L. Simpson).

1. Introduction

A rapid increase in geographically coded and electronically held statistical and administrative data has renewed the demand for datasets to be geographically consistent. Only then can data from different sources be used in the same area profiles and be monitored over time. However, in many countries statistical census output areas are not coterminous with administrative boundaries used for other government datasets, and each are different from electoral boundaries on which further datasets are maintained. Over time, boundaries change to meet policy and electoral developments. Major projects to monitor social programmes and governmental impacts in sub-national areas are frustrated by the lack of consistent geographical measurement and must ignore many data or attempt to develop methods converting data from one geography to another.

Such interpolation of spatial data has taken many forms but generally has been motivated by short-term purposes to achieve consistency of specific datasets (e.g. Champion, 1995; Dorling & Atkins, 1995; Wilson & Rees, 1999). For these projects, considerable skills in database or geographical information system (GIS) software have been invested to create and manage geography conversion tables, but these resources are not usually designed for use in subsequent projects by other researchers.

An earlier paper reviewed geography conversion tables and the types of error to be expected from data conversion; the errors arise both from the inaccuracy of the conversion table and from the ways in which the source units do not exactly nest within the target units (Simpson, 2002). The purpose of this paper is to stimulate greater *public* access to resources for geographical data conversion. The problem it addresses is that of a user wishing to submit their own data values for *source* geographical units and be returned with the equivalent data values for each *target* geographical unit, without the need to intervene with database and geographical information system skills.

This introductory section reviews the nature of the problem of public access to data conversion, the elements to any solution, and existing approaches to the problem. Subsequent sections define and evaluate one particular solution, which has been successfully implemented for general use with UK geographical units at <http://convert.mimas.ac.uk/>.

1.1. *Public access to data conversion: a review*

If users are to submit their own data values measured for specified source geographical units, for conversion to a new or target geographical scheme, without the need for technical database or GIS skills, then three elements must be supplied within any solution:

1. A *geography conversion table*, which provides an allocation of each source unit to one or more target units. The table may be explicit as in the example later in this paper, or implied within algorithms that are executed ‘on the fly’.

It is the mechanism for interpolation of spatial data from one set of units to another. Successful public access will rely on many geography conversion tables to allow the user to specify many different combinations of source unit and target unit—electoral, administrative, census, and so on.

2. A *matching algorithm*. The algorithm matches the source units within the user's data with the geography conversion tables, in order to calculate data values for the target units. Successful public access will include intelligent handling of incomplete and alternative formats of geographical codes.
3. A *public interface*. The means of receiving input and providing intelligent output on the results of the matching process. Successful public access will include not only useful documentation of the site itself, but appropriate information on the results of matching, the quality of the converted data, and supplementary data about the target units. It will allow the submission and receipt of data to be embedded within a user's external program.

Some of these elements are implemented in the allocation of specific single codes to a map or to other information. Thus, the user-entry of a postal code is a common means to locate an area for US Census information (<http://www.census.gov/geo/www/garm.html>, 2002), for UK government Neighbourhood Statistics (<http://statistics.gov.uk/neighbourhood/>, 2002), and for the location of community services (<http://www.multimap.com/>, 2002). This is a straightforward use of a geography conversion table from source postal codes to target areas or a map grid reference. The matching algorithm is a simple database search function. Many service industries make use of similar postal directories that return an address for a postal code input by the user.

More generally the user's need is to change the spatial scale of aggregated areal data, where its distribution within each aggregate area is unknown. Atkinson and Tate (2000) review issues of rescaling, a 'problem of fundamental importance to geography', emphasising that the observed variation of data values is always a function of both the true underlying variation of the data and the scale of measurement. This is the cause of the modifiable areal unit problem (Openshaw, 1984). From a statistical perspective, geographical rescaling and data conversion can be seen as a special case of synthetic estimation (Simpson, 2002).

Sadahiro (2000) and Francis, Lowe, Rushton, and Rayco (1999) investigate the case where for confidentiality purposes raw data have been aggregated to, and released for, representative points within small source units, and these must be used to estimate data values in larger target units for which boundaries are provided. Sadahiro finds that for greatest accuracy the representative point should be the two-dimensional median, rather than the centroid of the original data points as is often the case in practice (Dorling & Atkins, 1995). However, where the source unit boundaries are also available, Sadahiro finds that the point in polygon methods are usually less accurate than areal weighting by the overlap between source and target areas. In areal weighting the matching algorithm is the process of intersecting two sets of boundaries; the conversion table is comprised of the areal weight of each intersection within its source unit; an interface for areal weighting is provided by

most GIS packages: the user requires only the source data and boundaries of source and target units.

Data conversion by areal weighting can in turn be improved by ancillary data on target areas (Goodchild, Anselin, & Deichmann, 1993; Moxey, McClean, & Allanson, 1995). Bloom, Pedler, and Wragg (1996) implement in MapInfo a statistical model of the unknown data in the areal overlaps, using the EM algorithm and ancillary target area data as proposed by Flowerdew and Green (1992). While the use of a proprietary package makes this method generally available, it is still a GIS technician's tool.

The errors and problems of data conversion are much reduced if the source units are much smaller than target units, and to this end various attempts have been made to disaggregate data to small areas. In the GIS framework, the disaggregated data can then be re-aggregated through point in polygon methods to any target units for which boundaries are provided. Bracken and Martin (1989) spread data from a centroid within a polygon to a smooth data surface of 50-m grid squares. Xie (1995) distributes aggregate population data along networks of streets, on the basis that population and housing is usually confined close to these communication routes. Similarly, Fisher and Langford (1995) favour 'dasyymmetric' maps, where boundaries are snapped around residential areas before using areal weighting for data conversion. Outside GIS, database functions can achieve the same spatial disaggregation of data, and re-aggregation to larger areas, if addresses or small areal units are coded to more than one geographical scheme. In this vein, the Bradford Community Statistics Project (BCSP) uses the number of addresses represented by each grid-referenced postcode point to spread aggregate data within source units (www.bdsp-web.org, 2002; Thomasson, 2000). The number of addresses or postal delivery points used by BCSP (and in this paper) as auxiliary information for data conversion, is geographically correlated to many social variables, though will be less appropriate for some, such as vandalism of public property or workplace characteristics.

While many Internet sites provide pre-specified data that has been converted in the ways described to a common geographical scheme, the BCSP begins to give the user control over the tools of data conversion. The user is encouraged to draw any polygon on-screen on a Web-served street map, and is rewarded with a profile derived from the data held within the polygon. Because the data has been pre-disaggregated to very small areas using the ancillary information of numbers of addresses, the profile is sensitive to the precise area drawn by the user. The result is a powerful dynamic planning tool directed towards community activists and local government planners.

The example used in the following sections achieves far more than the mainly theoretical or parochial applications reviewed earlier. It allows submission and data conversion of any data, without need for GIS or database software or boundaries on the part of the user. It derives 200 conversion tables between geographical schemes often used in government and commercial datasets, from the 'All Fields Postcode Directory'. As this database is not publicly documented it is described in some detail later, and the conversion tables are subjected to evaluation. The matching

algorithms deal with many formats of postal codes and intelligently complete partially missing postal codes. The interface demands that the user supplies his or her own data for one of eight common geographical schemes—postal, electoral or census—and chooses a target geography from a choice of 25 geographical schemes. The site does not hold any data itself; it is purely a tool for data conversion. While not completely flexible, the development has allowed demographers, social scientists and political scientists to estimate consistent data for analyses that would otherwise not be possible. Importantly, the three elements of conversion tables, matching algorithms and interface are independent, allowing planned addition of conversion tables from any source including GIS, and further development of the matching algorithms and interface.

2. The creation of geography conversion tables from the UK All-Fields Postcode Directory

This part of the paper describes the derivation of a United Kingdom (UK) postcode directory and its use to produce multiple geography conversion tables. The characteristics and quality of the directory and the conversion tables are investigated.

2.1. A UK postcode directory

For the most part the directory is a copy of the Office for National Statistics ‘All-Fields Postcode Directory’ (AFPD). The authors’ cleaning and the addition of 1991 Census codes for Scotland and Northern Ireland are described, to achieve a database suitable for the derivation of geography conversion tables.

The AFPD is unique as a list of all postcodes in the UK whether or not currently in use by the postal service. The version used in this paper is the final AFPD of 1999 (version 1999b), containing over 2 million postcodes. The description is based on the authors’ discussions with ONS, on public documentation first issued in 2000 (ONS, 2000), and includes subsequent developments, primarily those improving grid-referencing and allocation of postcodes to other areas (May, Standen, & Taylor, 2001).

Each record of the AFPD reports a postcode, the number of residential addresses in that postcode, the numbers of other types of address, and a code for each of several geographies from which those listed in Table 1 can be derived. Some indicators of the quality of the information are included, and many non-residential postcodes can be identified—those representing “Post Office boxes” and non-geographic locations such as those for mail held at the sorting office for large organisations. Also provided are a grid reference for each postcode, the postcode’s date of introduction and its date of termination if not currently in use.

Postcodes in the UK (Raper, Rhind, & Shepherd, 1992) do not themselves represent geographical areas but collections of addresses for the purposes of mail delivery. Postcodes in the UK are highly structured and always contain four elements. For example, the first half of M13 9PL refers to postal district 13 within the

Table 1

Geographies derived from the UK All-Fields Postcode Directory (with example codes)

1991 Census	Postal
Census ED/OA 91 (EGFA01)	Postal District 99 (BD1)
Census wardlevel 91 (EGFA)	Postal Sector 99 (BD1 5)
Census District 91 (EG)	Postcode 99 (BD1 5DT)
Electoral	Health
County 98 (47)	Primary Care Group 99 (4AA01)
Ward 98 (JAMA)	Health Authority 98 (QAA)
Local Authority District 98 (JA)	NHS Regional Office 99
Parliamentary Constituency 97 (001)	
European Electoral Region 98 (01)	Statistical
	Standard Statistical Region 98 (1)
Administrative	Travel To Work Area 98 (001)
Country 98 (044)	NUTS-1 98 (UK4)
Gov. Office for the Region 98 (A)	NUTS-2 98 (UK41)
Local Education Authority 98 (201)	NUTS-3 98 (UK411)
Training Enterprise Council/LEC 99 (001)	NUTS-4 98 (UK41100)
	NUTS-5 98 (UK41100001)

NUTS refers to the European Community's five levels of territorial units. In the UK, NUTS-1 equates to government office regions, and NUTS-5 equates to electoral wards.

Manchester postal area. The single number following the postal district, in this case the number 9, represents a postal sector within M13. Postal sectors have on average a population of 6000 within 2500 addresses. The final two letters complete the postcode. Residential postcodes—those with at least one residential address—are small; the median number of residential addresses is 14, the maximum 149, and 99% have less than 63 addresses.

Each postcode has only one record even if its addresses straddle the boundary between more than one unit of other geographies. The updating processes allocate each postcode to a single unit of each geography in Table 1.

2.2. The updating processes of the AFPD

The AFPD is issued four times a year and maintained by the Office for National Statistics from three regular sources of new information, and two one-off independent exercises:

- (i) *Electoral geography*: existing postcodes are re-allocated following boundary changes for ward or parliamentary constituencies reported by the ONS Boundary Commission. The Local Authority, Local Education Authority, European Electoral Region, Government Office, and Standard Statistical Region classifications are derived from the ward classification.
- (ii) *Health Authority geography*: existing postcodes are re-allocated when changes to Health Authority and Health Region boundaries are reported by the Department of Health.

- (iii) *New postcodes and terminated postcodes*: between 4000 and 5000 new postcodes are added each month, and between 2000 and 3000 existing postcodes are terminated, using information from the Royal Mail's latest Postal Address File (PAF). The majority of the movement is due to business postcodes. Records for postcodes terminated by the Royal Mail are kept on the AFPD, but replaced if the same postcode is subsequently re-used—the historical record is thus limited. With each new postcode comes four address counts: residential addresses, small business addresses, delivery points (which usually is the sum of the two previous counts), and multiple occupancy. The grid reference is of the first-listed address, but since 2000 it is the centroid of the postcode's addresses' grid references, resolved to the nearest 1 m.

When a new postcode is added from the PAF, its electoral and health geographies have been since 2000 derived by point in polygon by the Ordnance Survey (May et al., 2001). The address counts and grid reference fields, for existing current postcodes, are also updated from the PAF in every new version of the AFPD, but allocation to other geography codes is maintained as in the previous version of the AFPD. Terminated postcodes retain the address counts and grid references frozen from their last mention on the PAF.

Previous to 2000, new and re-used postcodes have been allocated to geographies by imputation from the existing postcodes alphabetically immediately before and after the new code, within the same postal sector. Where it is not possible to impute in this way, for example because it is the first or last postcode in a postcode sector, or for a new postal sector, then the postcode is marked for clerical coding, which is carried out as resources permit. Since alphabetically contiguous postcodes need not be geographically contiguous, imputation introduces significant approximation.

In addition to the regular updating, two independent exercises have added extra geographical codes:

- (iv) *Census geography*: for each postcode in England and Wales existing at the time of the Census, the 1991 Census Enumeration District containing most of the postcode's households was derived directly from the Census database. Postcodes added since the 1991 Census have a Census code imputed from alphabetically similar postcodes as above. Codes for Travel to Work Area and Training and Enterprise Councils (now superseded by Skills Learning Councils) are then generally derived from the Census code. It is expected that the 2001 UK Census Output Area code will be added to the AFPD in 2003.
- (v) *Primary Health Group/Local Health Group*: as they were constituted at the beginning of 1999, based on electoral, postal and 1991 Census look-up tables supplied by each Health Authority.

The Channel Islands and the Isle of Man are not included in the processes to update geographies. For these two areas, each with approximately 5000 postcodes, the AFPD provides only a means to check that a given postcode exists. ONS restricts their updating processes to postcodes which are clearly geographic and at

least partly residential in nature. Table 2 classifies the AFPD postcodes accordingly, and by whether the postcode is recorded as in use or terminated. Finally, occasional clerical changes are made to the AFPD in response to users noting postcodes not on the file, or improvements necessary to specific records.

2.3. Validation, removal of duplicates, addition of Census codes

When it is viewed as a database for statistical use, integrity checks during the maintenance of the AFPD are incomplete. Evident lack of internal consistency includes 673 incompatible dates of introduction and termination; 345 duplicate

Table 2
PAF classification of postcodes as held on the All-Fields Postcode Directory (AFPD)

	England	Scotland	Wales	Northern Ireland	Channel Islands	Isle of Man
Current postcodes	81%	78%	73%	89%	100%	95%
<i>Large user (business) postcodes</i>						
Neither PO Box nor non-geographic	77,349	7,159	4,279	2,091	37	156
PO Box	89,160	4,951	3,399	1,329	297	4
Non-geographic	2,879	10	81	6	87	1
PO Box and non-geographic	4,341	41	13	–	1,383	437
Unknown	14,532	149	810	220	44	28
<i>Small user postcodes</i>						
Neither PO Box nor non-geographic	1,223,952	138,037	80,893	40,917	4,317	4,000
PO Box	5	1	1	1	–	–
Non-geographic	2	–	–	–	–	–
PO Box and non-geographic	–	–	–	–	–	–
Unknown	6,332	657	330	200	14	19
Terminated postcodes	19%	22%	27%	11%	0%	5%
<i>Large user (business) postcodes</i>						
Neither PO Box nor non-geographic	8,082	732	1,294	192	–	17
PO Box	37,264	2,360	2,350	596	–	1
Non-geographic	3,245	3	4	–	3	108
PO Box and non-geographic	1,149	4	–	–	–	100
Unknown	157,476	12,728	6,859	1,925	1	7
<i>Small user postcodes</i>						
Neither PO Box nor non-geographic	27,302	3,328	13,706	1,124	2	13
PO Box	–	–	–	–	–	–
Non-geographic	–	–	–	–	–	–
PO Box and non-geographic	–	–	–	–	–	–
Unknown	99,720	23,685	9,111	1,832	7	5
Total, all unique postcodes (= 100%)	1,752,790	193,845	123,130	50,433	6,192	4,896

All unique postcodes on AFPD version 1999/1 (for exclusion of non-unique postcodes, see text). The categories of postcodes highlighted in bold are used for the production of geography conversion tables described in the text. PAF, postal address file.

postcode records; 0.3% of postcodes with delivery counts not equal to the sum of residential and small business address counts; missing codes not distinguished between ‘not applicable’, ‘not processed’, ‘unknown’ and other reasons; a small percentage of invalid geography codes when compared with the file specification and the code documentation files; and geographical references remaining for a large number of postcodes after they had been marked as non-geographical.

These inconsistencies, further documented in Yu and Simpson (2000), do not prevent the AFPD’s use in geographical and statistical manipulation, but cause users to invent a variety of strategies to cope, which would be better adopted in a consistent manner within ONS before release. For the current work, duplicate records were removed and census codes from Northern Ireland and Scotland added as follows.

Duplicate records in the AFPD arise mainly when ONS merge their data with records from Scotland files from which they take some area codes. In the AFPD, we identified 325 postcodes which appeared more than once in the file, involving a total of 670 records. Table 3 exemplifies the rules of excluding 345 duplicate records, the key being to keep the most recent record for a postcode.

The AFPD does not include 1991 Census codes for Northern Ireland and Scotland, which significantly restricts its use by social data analysts. We obtained a copy of the Census codes for each postcode from the respective national statistical agencies. To bring the records into line with the AFPD, Scotland postcodes were reformatted and multiple records for the same postcode were eliminated by retaining the overlap with a census area involving most housing units. While all but one of the

Table 3
Duplicate postcode records on the AFPD - examples

Different termination dates, latest termination date retained (second record):

IV1 2QR,12/1997,12/1998,,QT,57,0,0,,28270,85310,0,SH9,S00,179,0,,X,,650,11,S15,259,49998,
UKM4201057,,,,42,42,42,0,,QT

IV1 2QR,01/1999,12/1999,,QT,53,0,0,A,28164,85284,0,SH9,S00,179,,1,X,,650,11,S15,259,49998,
UKM4201053,,,,0,0,0,,QT

Identical termination dates, latest introduction date retained (first record):

AB1 0LT,01/1980,06/1996,,QB,40,0,0,A,38262,80048,0,SN9,S00,179,,X,,670,11,S08,248,49998,
UKM1102040,,,,7,7,5,0,SN9,QB

AB1 0LT,08/1973,06/1996,,QA,35,0,0,B,38244,79924,0,SN9,S00,179,,X,,603,11,S08,248,49998,
UKM1101035,,,,7,7,5,0,SN9,QA

Only introduction dates, latest date retained (third record):

EH526QJ,10/1980,,RH,03,0,0,A,30904,67644,0,SS9,S00,179,0,,X,,653,11,,49998,
UKM2800003,,,,10,11,11,1,SS9,RH

EH526QJ,10/1980,,QP,03,0,0,B,31034,67644,0,SS9,S00,179,0,,X,,632,11,,49998,
UKM2500003,,,,10,11,11,1,SS9,QP

EH526QJ,12/1995,,QP,03,0,0,C,30908,67759,0,SS9,S00,179,0,,X,,632,11,,49998,
UKM2500003,,,,10,11,11,1,SS9,QP

The postcode, date of introduction and date of termination are the first, second and third fields respectively, the latter expressed as month/year.

agency postcodes were already represented on the AFPD, over 15% of the AFPD postcodes for Northern Ireland and Scotland do not exist on the postcode lists provided by those countries' own statistical agencies. These are mainly terminated postcodes.

The AFPD version 1999b cleaned by the authors has 2,131,286 records, with 1991 Census codes extended to all of Great Britain and Northern Ireland. It is this file that is referred to in the remainder of this paper.

2.4. *Production of geography conversion tables from the AFPD*

Two hundred UK-wide geography conversion tables have been constructed from the cleaned AFPD postcode directory. A single stored procedure creates a geography conversion table for any pair of source and target selected from Table 1, as follows:

1. 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.
2. Group by source code, summing the total residential addresses in the source unit = S.
3. Group by source and target codes, summing the number of residential addresses in each overlap = O.
4. Output the conversion table: source, target, residential addresses (O), weight (O/S).
5. As an exception, when postcode is the source geography, all postcodes with non-blank target are used, with weight 1.

Postcodes are thus used as an intermediary to construct geography conversion tables. The process is illustrated in Fig. 1 with simulated records. Where the AFPD indicates that all the postcodes in a source geography unit lie within one target geography unit (as in Census areas 1 and 2 in Fig. 1), 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 (Census areas 3 and 4), the AFPD provides a weight based on the number of residential addresses in the overlap of the source unit with target geography units.

The same logic could be used to construct a conversion table from a database of properties or any other items, whose records contain codes for the source and target geography and a value for the weighting criterion.

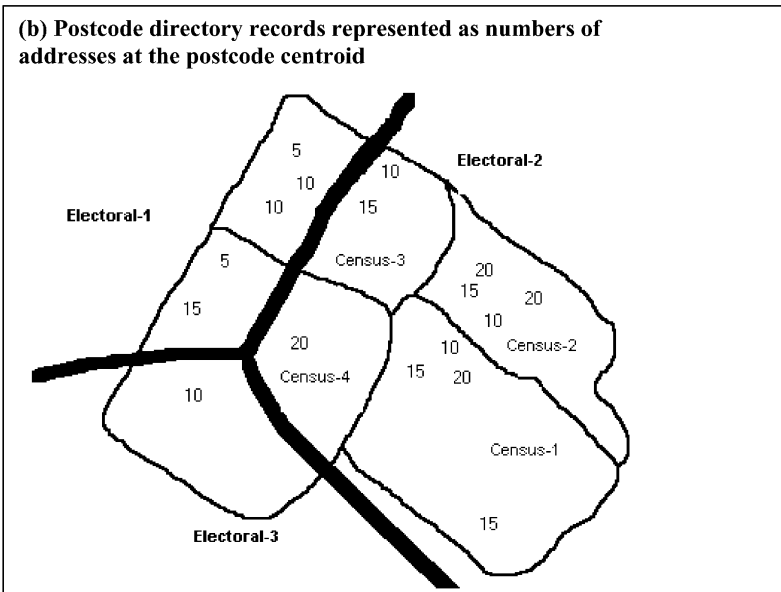
2.5. *The quality of the AFPD and geography conversion tables based on the AFPD*

Fig. 1(b) represented the production of geography conversion tables as if from a map of postcode points lying within the boundaries of source and target geographies. However the production did not use maps but database procedures as described in the previous section, with room for error when a postcode is not allocated on the database to correct units of other geographies. This section exam-

(a) Postcode directory records

Postcode	Census unit	Electoral unit	Count of residential addresses
Postcode 1	Census-1	Electoral-2	15
Postcode 2	Census-1	Electoral-2	20
Postcode 3	Census-1	Electoral-2	10
Postcode 4	Census-1	Electoral-2	20
Postcode 5	Census-2	Electoral-2	15
Postcode 6	Census-2	Electoral-2	10
Postcode 7	Census-2	Electoral-2	20
Postcode 8	Census-2	Electoral-2	15
Postcode 9	Census-3	Electoral-1	5
Postcode 10	Census-3	Electoral-1	10
Postcode 11	Census-3	Electoral-1	10
Postcode 12	Census-3	Electoral-2	10
Postcode 13	Census-3	Electoral-2	15
Postcode 14	Census-4	Electoral-1	5
Postcode 15	Census-4	Electoral-1	15
Postcode 16	Census-4	Electoral-2	20
Postcode 17	Census-4	Electoral-3	10

(b) Postcode directory records represented as numbers of addresses at the postcode centroid



(c) Geography conversion table derived from the postcode directory

Census unit	Electoral unit	Count of residential addresses	Total in source unit	Weight (Count/Total)
Census-1	Electoral-2	65	65	1.00
Census-2	Electoral-2	60	65	1.00
Census-3	Electoral-1	25	50	0.50
Census-3	Electoral-2	25	50	0.50
Census-4	Electoral-1	20	50	0.40
Census-4	Electoral-2	20	50	0.40
Census-4	Electoral-3	10	50	0.20

Fig. 1. Construction of a geography conversion table from a postcode directory: example.

ines the data conversion tables extracted from the AFPD, using both internal evidence and comparison with conversion tables derived from other postcode directories and using Geographical Information System (GIS) software. In most cases the completeness and accuracy of the conversion tables from the AFPD are found to be high.

Table 4 displays two measures (Simpson, 2002) of the goodness of fit between pairs of geographies extracted from the AFPD. The *degree of hierarchy* is the proportion of source units that are contained within a single target unit. For each of these source units there is only one record on the geography conversion table, with weight equal to 1. If the degree of hierarchy is 100%, the source units fully nest within the target units in a hierarchical arrangement, as do postal sectors within postal districts, and Census Enumeration Districts within Census Wards and Districts.

The *degree of fit* sums the largest weight from each source unit and expresses this as a proportion of the number of source units. It is always more than or equal to the degree of hierarchy. The degree of fit shows how closely on average the number of addresses in a source geography unit matches the number of addresses in its largest overlap with a target unit. It recognises that source geography unit boundaries may be close to target unit boundaries even when they do not exactly fit within them as demanded by hierarchy.

Based on the AFPD, the degree of hierarchy is 100% for all conversion from postcodes to other geographies, since the AFPD allocates each postcode wholly to other geographies. The error arising from the use of whole postcodes to create conversion tables reduces rapidly as the size of target unit increases, for example reducing to 0.5% when allocating postcoded household data to electoral wards (Simpson, 2002).

The degree of hierarchy is generally lower than 100%. In the UK, the boundaries of administrative and electoral areas do not follow postal boundaries; regular reviews ensure that they also do not follow boundaries from the previous census. The degree of hierarchy in Table 4 shows that all but 10.2% of local government District boundaries had changed by 1998 since the time of the 1991 Census. However, the value for the degree of fit for Districts from 1991 to 1998 is 97.1%, illustrating the minor nature of the boundary changes. Postal sectors on the other hand do not fit well within electoral wards, with a degree of hierarchy of 10.6% and a degree of fit remaining under 70%.

The degree of fit indicates the accuracy of the data conversion that the tables can provide. When a pair of geographies has a high degree of fit, data for the source geography may be allocated to target units with confidence. The amount of approximation also depends on how well the data within the source units are correlated to the weights, which are based on the number of residential addresses.

Missing values in the postcode directory affect the accuracy of geography conversion tables. For any pair of source and target geography, only AFPD postcode records that have *both* present can be used to compute their geography conversion table. However, the calculation of a conversion table and its weights will still be of high quality if the omissions are not concentrated in specific areas. On the AFPD, all records have postcode and country recorded, but every other geography is missing

Table 4
Degrees of hierarchy and fit for conversion tables extracted from the All-Fields Postcode Directory (AFPD)

		Source unit							
		Postal			Census		Electoral		
		Postcode 99	Sector 99	District 99	ED 91	Ward 91	District 91	Ward 98	District 98
Number of source units (= 100%)		2,131,286	9,252	2,780	150,909	11,103	459	11,134	434
target unit									
<i>(a) Degree of hierarchy</i>									
Postal	Postcode 99		0.4%	0.5%	4.6%	0.2%	0.0%	0.0%	0.0%
	sector 99	100.0%		19.4%	76.8%	17.2%	0.0%	13.7%	0.0%
	district 99	100.0%	100.0%		90.5%	46.5%	0.2%	43.4%	0.2%
Electoral	ward 98	100.0%	10.6%	6.3%	95.8%	13.5%	0.0%		0.0%
	district 98	100.0%	74.0%	49.8%	98.0%	81.7%	10.2%	100.0%	
	county 98	100.0%	92.4%	84.0%	99.3%	95.6%	63.5%	100.0%	100.0%
	parlcon 97	100.0%	59.3%	33.7%	96.3%	70.1%	5.7%	99.7%	28.6%
	european 97	100.0%	97.5%	94.0%	99.8%	98.1%	73.9%	100.0%	100.0%
Administrative	teclec 99	100.0%	88.6%	75.0%	99.3%	92.8%	45.3%	93.3%	40.7%
	lea 98	100.0%	82.2%	63.5%	98.7%	88.4%	25.1%	100.0%	100.0%
	gor 98	100.0%	97.5%	94.0%	99.8%	98.1%	73.9%	100.0%	100.0%
	country 99	100.0%	99.7%	99.2%	100.0%	99.7%	91.1%	100.0%	99.3%
Census	ed 91	100.0%	1.4%	1.3%		1.0%	0.0%	0.1%	0.0%
	ward 91	100.0%	13.1%	7.4%	100.0%		0.0%	12.7%	0.0%
	district 91	100.0%	71.1%	45.4%	100.0%	100.0%		80.1%	10.8%
Health	pcg 99	100.0%	70.3%	49.7%	97.4%	78.2%	18.7%	99.3%	58.6%
	hahb 98	100.0%	87.2%	72.7%	99.0%	90.6%	33.1%	98.2%	76.4%
	hahb 99	100.0%	87.4%	73.2%	99.0%	91.0%	34.9%	100.0%	96.1%
	nhsro 99	100.0%	97.8%	94.8%	99.8%	98.2%	74.7%	100.0%	99.5%
Statistical	nuts5 98	100.0%	10.6%	6.4%	81.5%	13.6%	0.0%	100.0%	0.0%
	nuts4 98	100.0%	74.0%	49.4%	98.0%	81.8%	10.0%	100.0%	99.1%
	nuts3 98	100.0%	86.0%	69.6%	99.0%	90.8%	27.5%	100.0%	99.1%
	nuts2 98	100.0%	93.1%	84.2%	99.5%	95.0%	50.3%	100.0%	99.3%
	nuts1 98	100.0%	97.5%	94.0%	99.8%	98.1%	73.9%	100.0%	100.0%
	ttwa 98	100.0%	81.1%	61.8%	98.4%	85.2%	28.1%	92.5%	31.3%
	ssr 98	100.0%	97.9%	94.5%	99.8%	98.4%	82.4%	100.0%	100.0%
<i>(b) Degree of fit</i>									
Postal	postcode 99		5.0%	3.1%	33.4%	4.7%	0.2%	4.2%	0.2%
	sector 99	100.0%		52.7%	95.4%	78.8%	12.2%	76.7%	12.0%
	district 99	100.0%	100.0%		98.4%	91.4%	34.1%	90.6%	33.6%

(Table continued on next page)

Table 4 (continued)

		Source unit							
		Postal			Census			Electoral	
		Postcode 99	Sector 99	District 99	ED 91	Ward 91	District 91	Ward 98	District 98
Electoral	ward 98	100.0%	68.5%	44.5%	99.1%	93.0%	8.3%		7.9%
	district 98	100.0%	97.0%	94.2%	99.8%	99.5%	97.1%	100.0%	
	county 98	100.0%	99.1%	98.4%	99.9%	99.9%	99.0%	100.0%	100.0%
	parlcon 97	100.0%	94.5%	89.5%	99.6%	99.1%	73.4%	99.9%	72.1%
	european 97	100.0%	99.7%	99.5%	100.0%	100.0%	98.1%	100.0%	100.0%
Administrative	teclec 99	100.0%	98.8%	97.7%	99.9%	99.9%	99.4%	99.8%	99.0%
	lea 98	100.0%	98.0%	96.2%	99.9%	99.7%	97.2%	100.0%	100.0%
	gor 98	100.0%	99.7%	99.5%	100.0%	100.0%	98.1%	100.0%	100.0%
	country 99	100.0%	100.0%	99.9%	100.0%	100.0%	98.1%	100.0%	100.0%
Census	ed 91	100.0%	16.8%	9.7%		20.1%	1.5%	17.7%	1.5%
	ward 91	100.0%	71.3%	46.4%	100.0%		9.2%	93.1%	8.6%
	district 91	100.0%	96.7%	93.7%	100.0%	100.0%		99.6%	95.8%
Health	pcg 99	100.0%	96.3%	93.2%	99.7%	99.4%	84.8%	99.9%	84.4%
	hahb 98	100.0%	98.6%	97.3%	99.9%	99.8%	97.0%	100.0%	99.0%
	hahb 99	100.0%	98.6%	97.3%	99.9%	99.8%	97.1%	100.0%	99.1%
	nhsro 99	100.0%	99.8%	99.6%	100.0%	100.0%	98.0%	100.0%	99.9%
Statistical	nuts5 98	100.0%	68.5%	44.5%	97.4%	93.0%	8.3%	100.0%	7.9%
	nuts4 98	100.0%	96.9%	94.2%	99.8%	99.5%	97.0%	100.0%	99.7%
	nuts3 98	100.0%	98.5%	97.1%	99.9%	99.8%	97.4%	100.0%	99.7%
	nuts2 98	100.0%	99.2%	98.6%	99.9%	99.9%	97.9%	100.0%	99.9%
	nuts1 98	100.0%	99.7%	99.5%	100.0%	100.0%	98.1%	100.0%	100.0%
	ttwa 98	100.0%	97.9%	96.5%	99.8%	99.7%	84.6%	99.7%	85.4%
	ssr 98	100.0%	99.7%	99.4%	100.0%	100.0%	100.0%	100.0%	100.0%

parlcon, Parliamentary Constituency; teclec, Training and Enterprise Councils/Learning and Enterprise Councils; lea, Local Education Authority; gor, Government Office of the Region; ed, Enumeration District; pcg, Primary Care Group/Local Health Group; hahb, Health Authority/Health Board; nuts, European territorial units; ttwa, Travel to Work Areas; ssr, Standard Statistical Regions.

on at least some postcode records. Outside the Channel Islands and the Isle of Man where most geographies are not recorded, the occurrence of missing geography codes is not generally clustered in particular areas or on particular records. Only 2,298 or 0.15% of the 1,491,318 ordinary postcodes within Great Britain and Northern Ireland are missing one half or more of their geography codes.

Table 5 shows the existence of each geography code for current residential postcodes. Only in the case of 1991 Census areas and Primary Care Groups does the proportion of current ordinary postcodes with missing code rise above 0.4%. For every 1991 Census Enumeration District (ED) represented on the AFPD there is at least one postcode record that is also coded with a current ward. However, the

Table 5
Completeness of coding for current ordinary postcodes in GB and N Ireland

Geography	Not missing	% Missing
Country	1,491,318	0.00
Census 1991	1,478,620	0.85
European Electoral Region 98	1,488,730	0.17
Electoral Ward 98	1,489,023	0.15
Government Region Office 98	1,488,836	0.17
Easting	1,486,402	0.33
Northing	1,486,402	0.33
Health Authority 99	1,490,233	0.07
Local Authority 98	1,489,440	0.13
Local Education Authority	1,488,836	0.17
Health Region 98	1,490,468	0.06
EU NUTS	1,488,835	0.17
Parliamentary Constituency 97	1,488,730	0.17
Primary Care Group 99	1,435,993	3.71
Travel to Work Area 98	1,487,355	0.27

reverse is not true. Ten current wards are lost from the conversion tables to 1991 Census geography, because none of the postcodes in those wards have a 1991 Census code. These are all in Northern Ireland. Missing codes for Primary Care Groups were more seriously concentrated in the AFPD, preventing the allocation of between 2 and 4% of 1991 Census areas and current wards to these health administrative areas. These two cases—Primary Care Group codes and Census codes—deserve some further comment.

The proportion of current residential postcodes without a Primary Care Group code (PCG, and Local Health Group in Wales) is higher than for any other geography code, at 3.71% (Table 5). This figure excludes Scotland and Northern Ireland where PCGs are not defined. The PCG codes within each Health Authority were added by ONS from various lists as described earlier, for example, based on current ward boundaries, or on postcode sectors. Where different types of allocation meet at the borders of a Health Authority, and where Health Authorities are yet to provide the constitution of all PCGs, some postcodes cannot be allocated. Missingness for the PCG field is therefore clustered more than for other geographies. The code is missing on the AFPD version 1999b for most postcodes in Shropshire and West Sussex Health Authorities, for a third of Cambridgeshire and Berkshire postcodes, and for over 10% of postcodes in West Pennine, Northamptonshire, Doncaster, South Cheshire, Salford and Trafford, and Manchester Health Authorities. ONS documentation suggests that this problem has persisted through the change from Primary Care Groups to Primary Care Trusts since the year 2000.

2.6. 1991 Census—an improved conversion table

Many analysts wish to allocate Census data to areas for which Census data have not been released. Approximation in the location of postcodes can have a relatively

large impact on conversion tables involving 1991 Census EDs, as they are the smallest unit of geography on the AFPD, apart from postcodes themselves. This section finds the AFPD wanting in its 1991 Census information. Improved data have been found and incorporated for the 1991 Census geography, as also described here.

The 1991 Census code is missing for almost 1% of current residential postcodes on the cleaned AFPD (Table 5). There are 155,448 1991 ED/OA Census areas in Great Britain and Northern Ireland, according to Census data files. Table 6 shows that of those that are not represented on the AFPD, 1050 were ghost EDs with zero population and are thus of no concern for those converting Census data to new geographies. A further 995 had so small a population that they would have been partially suppressed in the Census output itself. These will only be of concern for particular studies or areas (Cole, 1993). The remaining 129 EDs not found on any postcode record of the AFPD had substantial population at the time of the Census. Those that have been checked by local authorities in Nottinghamshire and Cheshire are not areas of housing change: they still have significant population. They indicate a serious problem with AFPD quality as regards the 1991 Census.

In 1999, ONS studied the quality of the allocation of postcodes to 1991 Census EDs, finding that 10% of postcodes assigned using census forms to 1991 Census EDs did not match the assignment using the Ordnance Survey product Codepoint. The match for the 25% of postcodes introduced since the 1991 Census was much poorer (ONS, 2000, p. 10). A postcode can be allocated to a different 1991 ward (by the Census code) and current ward (1998), even where the boundary had not changed. This is due to the inconsistent procedures for allocating postcodes to census and electoral geographies, referred to earlier. Fig. 2 shows an extreme example of how the postcodes allocated to a single ED have been allocated to a wide spread of neighbouring 1998 wards. Stafford 1991 Census ED 42QMFD06 is in reality entirely contained within 1998 ward 41UGFD, but is placed by the AFPD also in six adjacent wards (with weight shown as a percentage). Still more seriously, postcodes allocated to the same ED are in 4,149 cases allocated to non-contiguous 1998 wards. This indicates an error for one or more of the records. These inconsistencies and errors lead to significant approximations in the conversion tables between Census

Table 6
Enumeration Districts (EDs) not appearing in the cleaned All-Fields Postcode Directory (AFPD)

	All EDs	EDs not on cleaned AFPD 1999b		
		No residents	Small population	Ordinary EDs
England	106,843	986	799	26
N Ireland	3729	0	0	11
Scotland	38,255	0	151	89
Wales	6621	64	45	3
UK	155,448	1050	995	129

Small population, under 50 residents, or under 16 households; ordinary EDs, at least 50 residents and 16 households.

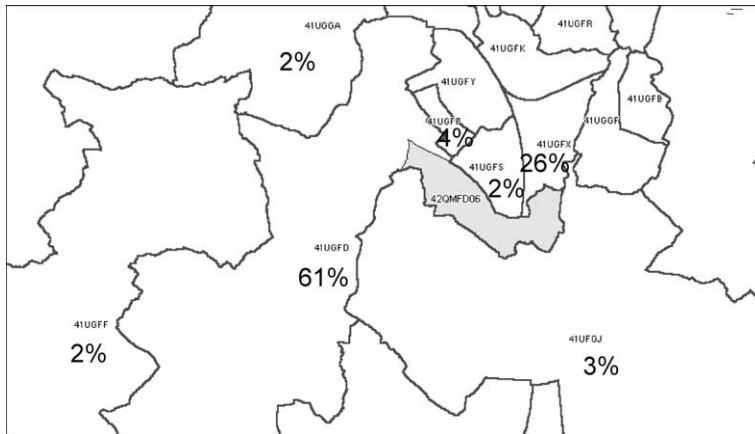


Fig. 2. Fuzziness in the conversion table from census to electoral ward derived from the All-Fields Post-code Directory (AFPD): an extreme case.

and other current ward geographies, implying changes in ward boundaries which have not in fact happened.

England and Wales have much poorer fits of EDs to the current ward geography relative to Scotland and Northern Ireland (Table 7). This is not itself proof of poorer quality, as it is possible that the re-organisation of local boundaries during the 1990s affected the relationship between EDs and current wards more in England and Wales. However, the use of an alternative file for England has allowed a closer examination and confirms a problem for that country at least, and provides a solution.

A team co-ordinated from Oxford University used GIS software to overlay two sets of digitised boundaries for England, the Ordnance Survey Boundaryline datasets for 1991 Census EDs and 1998 wards (Noble et al., 2000). Considerable manual effort removed ‘splinters’ of land created by slight imperfections in the alignment of the two sets of boundaries, where no population could be involved. The extent of surface area in the overlap between boundaries was used as the weight in a geographical conversion table to allocate Census data for 1991 EDs to the electoral wards existing in 1998. These area-based weights may correlate poorly and often negatively with the distribution of Census data within EDs, but an ED is unlikely to be allocated to the wrong ward under this procedure. This ‘GIS’ conversion table 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.

A new conversion table for England from 1991 Census ED to 1998 electoral wards has been created by combining the GIS-based and the AFPD-based tables. The following logic was used: (1) use unchanged the records for any ED that is recorded on only one of the two tables; (2) if the GIS table indicates an ED wholly within one 1998 ward, use this record and reject the AFPD records; (3) if the GIS table indicates an ED split between 1998 wards, use only the AFPD records for these same wards, and recalculate the weights using the AFPD address counts; and (4) if the

Table 7

Allocation of 1991 Census Enumeration District (ED) to 1998 electoral wards

	England	Wales	Scotland	Northern Ireland	All GB and NI
Number of records in the geography conversion table from 1991 Census ED to 1998 electoral ward	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%

Based on current ordinary postcodes in the All-Fields Postcode Directory (AFPD) version 1999b as cleaned by the project.

GIS table indicates an ED split between 1998 wards, for which there are no AFPD records for any of the wards, use the GIS records.

Table 8 compares the original and combined tables. 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 high at 97.4%. 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 as in the example of Fig. 2.

The combined file is an improvement in three ways. It records 635 EDs of England which were not on the AFPD file, including the 26 with substantial population in Table 6. It has greater degrees of hierarchy and fit than either file: the AFPD fuzziness has been greatly reduced while the overlaps in the GIS file that contain no population have been eliminated. It 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 GIS-derived file to be correct in its construction. Apart from the method of construction of the file preventing gross errors, its

Table 8

Comparison of AFPD and GIS allocation of 1991 census EDs to 1998 electoral areas, with the improved file combining their best features, England

	AFPD	GIS	Combined
Number of records	129,101	104,430	104,400
Number of 1991 EDs	103,090	103,168	103,725
Number of 1998 wards	8462	8442	8442
Degree of hierarchy	78.5%	98.8%	99.4%
Degree of fit	97.4%	99.8%	99.9%

AFPD, All-Fields Postcode Directory; GIS, geographical information system; ED, Enumeration District.

correctness has been further verified in six cases where the two files allocated EDs wholly to completely different wards, by checking with the local authorities concerned. In every case the GIS file was correct and the AFPD incorrect.

The improved records have been substituted for the records for England in the ED91to Ward98 conversion table derived from the AFPD. The improved table is used by the website for data conversion.

3. The algorithms for data conversion on-line

Two hundred UK-wide conversion tables created as above are available in <http://convert.mimas.ac.uk>. Data conversion is achieved online by the stored routines that this section describes, with no need for the user to have database and statistical programming skills. Special routines handle ill-formatted and incomplete postal codes to minimise the loss of data during conversion.

3.1. Data conversion

The user provides a file of data values D_s , each with source unit code s , and identifies the source geography S and the required target geography T . The algorithms match the codes s to the geography conversion table between S and T , and returns a file of data values D_t for target code t , calculated using the conversion table's weights as follows:

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

A single algorithm matches source codes to the conversion table and performs the data conversion: (1) filter records from the user's file keeping those whose source unit s is found on the conversion table defined by the user's choice of S and T ; (2) retrieve the target code(s) and weight(s) from the conversion table for each matched s ; and (3) multiply each source data value by the weight, group by target codes, and sum the weighted values. Fig. 3 provides an example, using the conversion table of Fig. 1. In practice the user's file may contain an unlimited number of data fields for each source unit. Each data field is converted to target units as earlier.

3.2. Formatting and imputation for a user's postcodes

When postcode is the source geography, two additional routines are executed before the data conversion. These cope with the non-standard and error-prone recording of postal codes on administrative and survey records. The routines minimise loss of data during data conversion by improving the match with the geography conversion table. They allow 'fuzzy matching' between the user's list and the conversion table (Chung & Jefferson, 1998).

Even when complete, users' postcodes may vary in length between 6 and 8 by the optional use of zero and blank characters. The first routine converts the user's

(a) User's source data		(b) Matched to conversion table				(c) Derived target data	
Source unit	Source data	Source unit	Source data	Target unit	Weight	Target unit	Target data
s	D _s	s	D _s	Target t	Weight w _{st}	t	D _t
Census-1	20	Census-1	20	Electoral-2	1.00	Electoral-1	9
Census-2	20	Census-2	20	Electoral-2	1.00	Electoral-2	49
Census-3	10	Census-3	10	Electoral-1	0.50	Electoral-3	2
Census-4	10			Electoral-2	0.50		
		Census-4	10	Electoral-1	0.40		
				Electoral-2	0.40		
				Electoral-3	0.20		

Fig. 3. Data conversion: example.

postcode to the same seven-character standard format used in the AFPD and many other systems. In the standard format the postal district, sector and code always occupy the first four characters, the fifth character and the final two characters, respectively, such that for example BD03_0JZ, WC1H_9RA and G1_2TE become BD3_0JZ, WC1H9RA and G1_2TE, where a _ represents a space. The key to the routine is to first strip out all blanks; complete postcodes have 5, 6, or 7 non-blank characters, of which the final three characters always represent the numeric sector and alphabetic postal code. The district code may be of variable length and is made to four characters by adding 0, 1, or 2 blank characters.

The second or imputation routine is provided to find a nearby postcode to substitute the user's postcode when after formatting it is not matched with a usable record on the conversion table. This will happen if the user's postcode is too new to appear on the AFPD, or is mis-typed or incomplete; the routine is also invoked where the user's postcode is found, but the target code is empty. The imputation of a nearby postcode prevents loss of data during conversion. However, the more complete results will be approximate to the extent that the imputed postcode is allocated to a target geography that is not the correct one. Imputed postcodes are listed on the log file so that the user may accept the results or decide to correct or delete those postcodes in their data file that had to be imputed. The hierarchical structure of UK postcodes is used to enable imputation of a nearby postcode. Three tables are pre-generated to assist postcode imputation, created from all postcodes in the AFPD having at least one residential address. Table A contains the sum of residential addresses for each unique postal sector code. It is created by summing the number of residential addresses in each postcode within a postal sector. Table B contains the address sum for each unique postal district, created in the same way. Table C contains postcode records with the cumulated number of addresses in each postal sector, and in each postal district.

The procedure for postcode imputation is used as follows for any postcode not matched with a postcode conversion file. If the postcode's sector is found in Table A, the address sum of the postal sector is obtained from Table A. A random number is generated within the range of zero and the address sum; the number is used to locate in Table C the postcode within the same postal sector which has sector cumulated sum closest to that random number, but not bigger than it. If the postcode's sector is not found in Table A, the postcode's district is searched in Table B, and a random

postcode within that district found in the same way. Thus, the postcode is chosen with probability proportional to the number of addresses in the postcode, to give less weight to the many postcodes with very few addresses.

4. Architecture and user interface

The architecture of the public-access website has three tiers: the user interface, the common gateway interface (CGI) and a database server (Abbott & Daly, 2001; Yu, 2000). The three tiers require different programming and design skills and are separated to ease development and maintenance.

The user interface uses Cold Fusion, ASP and HTML, taking advantage of the Web browser software in the user's own computer. It handles user input and displays data. The CGI uses Perl and C programs, responsible for file validation, postcode reformatting, file uploading and downloading as well as communicating with the Web browser and the database server. The programs can be run independently of a Web browser for ease of testing and development. The third tier is implemented within a SQL Server database using stored procedures. It contains the AFD and conversion tables and is responsible for data conversion, postcode imputation, log report generation and creation of lookup tables between any pair of area types. Typically these procedures parse strings and create a number of intermediate tables before creating a final product that will be copied to a file. They are also built with the capability of being tested in isolation by calling from the command line with appropriate arguments.

The site allows the user to create a geography conversion table between any pair of the area types listed in Table 1, directly from the AFD and using the algorithm described earlier, for a single local government area. Additionally, 200 UK-wide conversion tables are stored for downloading by the user. However, the most-used function of the site allows the user to convert their own data (Fig. 4). All functions are publicly available, but the use of postcodes for source or target geography is limited to academic users in order to protect the commercial copyright of the statistical office's original postcode records.

A Web interface allows the user to upload her or his own data file for geography conversion. The data file must contain at least two fields for each record, separated by a comma. The first field is the source geography, and the subsequent fields are value fields containing data. A number of validation checks are performed after the user's file is uploaded to ensure that the data conversion routine can correctly process the user's data. These checks identify disallowed formats, such as quotes around fields, irregular numbers of fields and characters in value fields. The comma-separated text demanded of input files is a format output by statistical and spreadsheet software commonly used by data analysts.

A log file is generated which the user can download for reference after the completion of data conversion. The first section records the location of the user's file that was uploaded, the name of the source and the target geographies, and the time of completion. The second section is a report of unmatched source units. The final

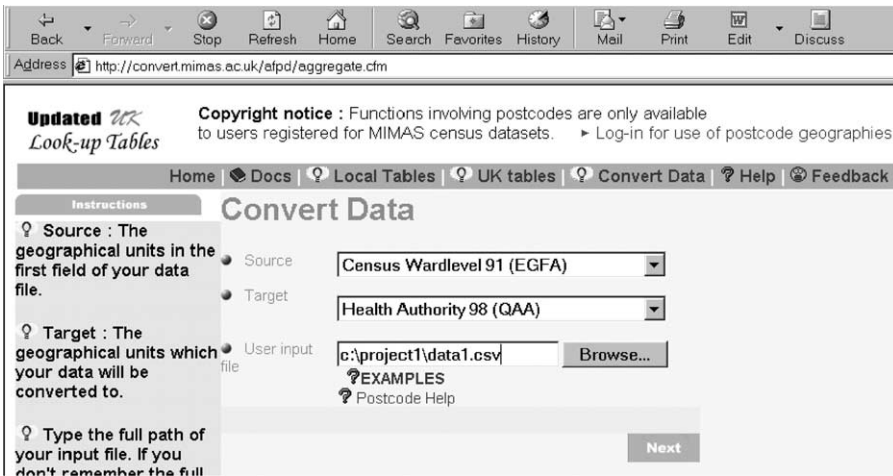


Fig. 4. Data conversion site: <http://convert.mimas.ac.uk>.

section lists for each degree of match, the number of records in this category and sub-totals of value fields.

Fig. 5 shows an example of the log file for a user's file of postcodes, where imputation has been necessary. In this case substitutes for two mis-spelled postcodes were found within their valid postal sector and postal district, respectively (BD21 9 is not an existing postal sector); one postcode had invalid postal district and was not substituted at all (BD52 does not exist), and a substitute was found for one incomplete postcode within its valid district. If, as is often the case, there is no more complete information from which to correct the data, the user may accept the results as the imputed values will lie in or close to the correct target areas, assuming of course that the first part of the postcode is correct. The data are thus allocated approximately correctly rather than ignored.

5. Discussion

The boundaries of interest to data analysts change over time and will continue to do so in the UK and elsewhere. Non-specialists need methods of transferring data from one set of geographical units to another, to create local statistical series, and to make disparate datasets compatible. This paper has described such a method and a website that gives public access to data conversion between the most common geographical units, and does not rely on specialist software or skills, or on geographical information systems. Its success results from treating separately the different elements that are needed for data conversion.

First, the algorithms and interface that achieve the data conversion are dependent on the existence of conversion tables but are independent of the origin of those conversion tables. Thus, while the site uses the 200 conversion tables produced as

```

File: 'C:\uamf\outputs\pc4odd.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

```

Fig. 5. Example log after data conversion.

earlier from the AFPD, it can and does also use other conversion tables, and can be developed as more tables are produced. Second, the site demands only that the data refer to source units contained on one of its geography conversion tables. The same conversion tables can be used with any data provided by the user at the time of data conversion.

The interface can be pragmatically developed independently of the conversion tables and the source data. For example, many users perceive as advantageous the fuzzy matching of postcoded records, which minimises data loss at the cost of approximate allocation of un-matched source units, and this could be developed for other source geographies. The UK academic datasets service now maintains the site, and will also consider the addition of target area labels on output files, conversion of data that is non-extensive such as proportions or rates, the facility to embed data conversion within an external program, and other developments. It expects to incorporate new versions of the AFPD from 2002, building a time series of geography conversion tables to allow the user to choose the table most relevant to the data that are to be converted.

Geography conversion tables allow weighted data conversion and are an increasingly important tool, meriting and receiving greater attention than in the past. Easy public access to data conversion will increase the demand for further improvements to the postcode directories which permit regular creation of multiple geography conversion tables as described in this paper. The AFPD itself is under development. The 2001 Census boundaries will be incorporated in 2003, while a separate project aims to bring 1991 Census boundaries within the same 'Gridlink' initiative to fix the allocation of postcodes by GIS methods, resulting in improved accuracy and geographical integrity in the AFPD geographical codes.

Considerable new developments would be required to tackle the main weaknesses of the approach to public access to data conversion described here. First, the source and target units could be entirely user-determined through a Web-served GIS allowing online submission of polygon boundaries attached to data. This would develop the approach to online point-in-polygon data conversion via local address points pioneered by the Bradford Community Statistics web site described earlier. Second, the provision of data for grid-referenced address points, perturbed or smoothed sufficiently to protect confidentiality, could replace aggregate data for areas. This would allow on-line accurate aggregation using point-in-polygon to any user-drawn or user-defined larger areas, and may be developed as part of the UK Government's Geo-Referencing Strategy (National Strategy for Neighbourhood Renewal, 2000). A national land and property gazetteer that distinguished and grid-referenced individual residential addresses could give rise to more precise national geography conversion tables than are possible at present, with a variety of weighting criteria (<http://www.nlpg.org.uk/>, 2001).

In the meantime, the AFDPD postcode directories provide a validated means of data conversion of great value to social science, within a publicly accessible on-line resource.

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>. Dan Abbott, Lou Daly and David Avenell undertook some of the work reported here. The Office for National Statistics and the University of Manchester MIMAS service responded to many requests for help and information. Five local authorities—Bridgend, Bradford, East Dorset, Peterborough and York—provided data to help the project, and others were helpful in validating findings. Justin Hayes and referees gave many improving comments.

References

- Abbott, D., & Daly, L. (2001). *UK Area Master Files system documentation appendix*. Manchester: Cathie Marsh Centre for Census and Survey Research, University of Manchester.
- Atkinson, P. M., & Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: a review. *Professional Geographer*, 52(4), 607–623.
- Bloom, L. M., Pedler, P. J., & Wragg, G. E. (1996). Implementation of enhanced areal interpolation using MapInfo. *Computers & Geosciences*, 22(5), 459–466.
- Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A*, 21, 537–543.
- Champion, A. (1995). Analysis of change through time. In S. Openshaw (Ed.), *The census user's handbook*. London: Longman.
- Chung, P., & Jefferson, M. (1998). A fuzzy approach to accessing accident databases. *Applied Intelligence*, 9(2), 129–137.

- Cole, K. (1993). The 1991 local base and small area statistics. In A. Dale, & C. Marsh (Eds.), *The 1991 Census user's guide* (pp. 201–247). London: HMSO.
- Dorling, D., & Atkins, D. (1995). *Population density, change and concentration in Great Britain 1971, 1981 and 1991. Studies on Medical and Population Subjects No. 58*. London: HMSO.
- Fisher, P. F., & Langford, M. (1995). Modeling the errors in areal interpolation between zonal systems by Monte-Carlo simulation. *Environment and Planning A*, 27(2), 211–224.
- Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *Annals of Regional Science*, 26(1), 67–78.
- Francis, R. L., Lowe, T. J., Rushton, G., & Rayco, M. B. (1999). A synthesis of aggregation methods for multifacility location problems: strategies for containing error. *Geographical Analysis*, 31(1), 67–87.
- Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25(3), 383–397.
- May, K., Standen, P., & Taylor, A. (2001). Gridlink—the new standard for postcode location data. *BURISA*, 147(April), 5–7.
- Moxey, A., McClean, C., & Allanson, P. (1995). Transforming the spatial basis of agricultural census cover data. *Soil Use and Management*, 11(1), 21–25.
- National Strategy for Neighbourhood Renewal. (2000). *Report of Policy Action Team 18: better information*. London: The Stationery Office.
- 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*. London: DETR.
- ONS. (2000). *All-Fields Postcode Directory 2000/1 version notes*. Titchfield: Office for National Statistics.
- Openshaw, S. (1984). *The modifiable areal unit problem (CATMOG38)*. Norwich: Geo Books.
- Raper, J., Rhind, D., & Shepherd, J. (1992). *Postcodes: the new geography*. Harlow: Longman.
- Sadahiro, Y. (2000). Accuracy of count data estimated by the point-in-polygon method. *Geographical Analysis*, 32(1), 64–89.
- Simpson, L. (2002). Geography conversion tables, a framework for conversion of data between geographical units. *International Journal of Population Geography*, 8(1), 69–82.
- Thomasson, E. (2000). Small area statistics on-line. *BURISA*, 144, 2–9.
- Wilson, T., & Rees, P. (1999). Linking 1991 population statistics to the 1998 local government geography of Great Britain. *Population Trends*, 97(Autumn), 37–45.
- Xie, Y. C. (1995). The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, 19(4), 287–306.
- Yu, A. (2000). *UK Area Master Files system documentation*. Manchester: Cathie Marsh Centre for Census and Survey Research, University of Manchester.
- Yu, A., & Simpson, L. (2000) *The All-Fields Postcode Directory (AFPD): validation for use as a look-up table* (Progress report prepared for Census Programme workshop, Leeds, 3–4 May 2000). Manchester: Centre for Census and Survey Research, University of Manchester.