THE DIFFICULTY OF LINKING TWO DIFFERENTLY AGGREGATED SPATIAL DATASETS: USING A LOOK-UP TABLE TO LINK POSTAL SECTORS AND 1991 CENSUS ENUMERATION DISTRICTS

Daniel Vickers

School of Geography
University of Leeds
Leeds LS2 9JT
United Kingdom

E-mail d.vickers@geog.leeds.ac.uk

January 2003

CONTENTS

ABSTRACT

The use of postal geography as areal units has developed significantly since its first introduction in the 1971 Census of Scotland. The 1987 Chorley report advocated the use of postal codes as the standard areal unit for publication of geographic data across the board. The change to a postal base in the census of England and Wales finally took place in 2001. Aggregation of population data is essential, both to protect the identity of individuals and make the data manageable. The question however remains, whether the independent aggregation of two similar datasets covering the same geographical area makes the data two separate information sources, or whether they can be successfully be linked together and used as one. Through the comparison of two aggregated areal datasets of British population based statistics, this paper examines the reliability of commercially produced, undocumented data, with the use of a look-up table linking postal sectors to enumeration districts of the 1991 Census of British population. The investigation finds that the ability to link the Experian dataset is to the census is questionable. The two datasets contain many obvious and significant differences when linked, it can be concluded that look-up tables are a poor and inaccurate way of linking differently aggregated spatial datasets.

ACKNOWLEDGEMENTS

Firstly, I would like to thank the European Social Fund (ESF) for the funding they provided during the completion of my MSc. Course at the University of Nottingham during which most of the work on this paper was completed. This project would not have been possible without the provision and availability of data, for this I must thank Experian Ltd, the Office for National Statistics and Manchester Information & Associated Services (MIMAS). I wish to thank the staff of the University of Nottingham Geography Department, especially Bob Abrahart, who was responsible for the initial idea behind the project and Michael McCullagh who oversaw the development of the idea into a working project. Phil Rees of the University of Leeds for answering many of my queries about census geography.

The opinions expressed are solely of the author and not of any organisation or individual mentioned within the text.

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In 1987 the Department of the Environment chaired by Lord Chorley published a report titled 'Handling Geographic Information'. The report discusses the way in which, data should be spatially referenced to increase uniformity. The report concluded that two different systems should be employed as standard as they were ideally suited to different situations and could be accurately linked together. The Ordnance Survey's British National Grid co-ordinate system and the Royal Mail's postcode system were recommended.

Since the publication of the Chorley report some changes have taken place, for example the 1991 Census in Scotland was largely built upon postcodes, as were many commercial datasets. However, the 1991 Census of England and Wales was still based on arbitrary output units, called enumeration districts (EDs). Experian postal sector data contains socio-economic data based on the geography of the Royal Mail's postcode system, it also contains a look-up table linking the postal sectors, to Census EDs. The linking of the 1991 Census data and the Experian postal sector datasets through the look-up table will enable an assessment of how well two datasets can be accurately linked. This would also enable the accuracy of the undocumented Experian data set to be assessed against the transparently produced census dataset.

The ability to link datasets in such a way relies upon them being based on the same geography or the availability of an accurate and reliable system to link the different geographies. The ability to reliably link the Census and Experian datasets via the look-up table would be a valuable tool, as it not only enables the two datasets to be linked

geographically, but also applies the 1991 Census data to the geography of Royal Mail's postal system.

2. POSTCODE GEOGRAPHY

The British Postcode system was created with the sole aim of enabling the automated sorting and delivery of mail. This purpose is clearly reflected in the geography of postcodes. Postcodes are made up of an '*outcode*', used to establish which sorting office to send the mail *'out to'* and an '*incode*', used to establish what part of the local area the mail is to be sent to when it comes *'in to'* the local sorting office. Both the incode and outcode are made up of two geographic parts creating a four tier hierarchy to postal geography.

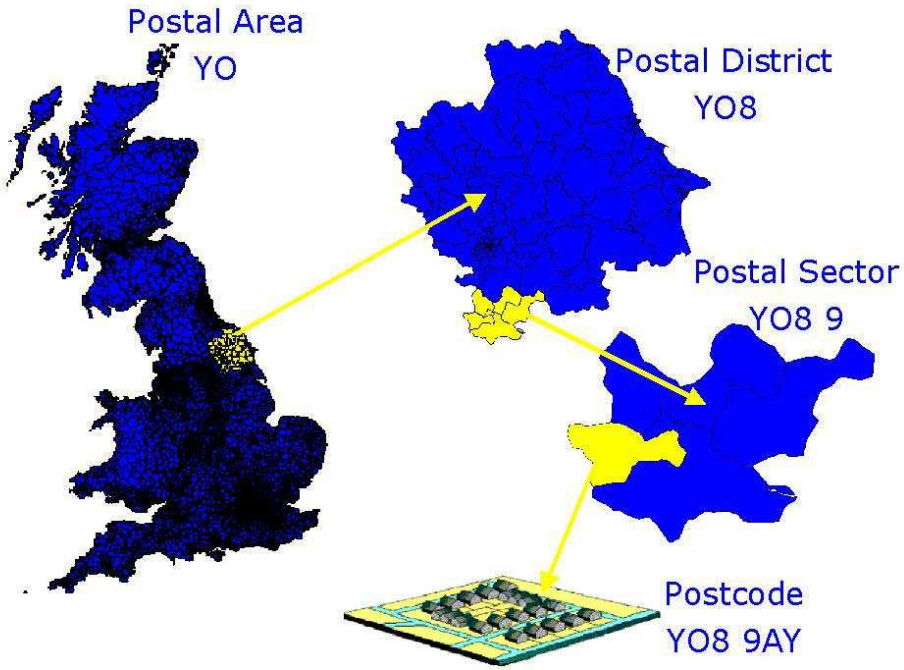


Figure 1: An Illustration of the hierarchy of Postal Geography in the UK

The geography of the British postal system is illustrated in figure 1, the outcode (e. g. YO8) made up of the postal area, of which there are 120 in Britain (e.g. YO representing York and the surrounding region), and the postal district (e.g. YO8) of which there are approximately

2700. The incode (e.g. 9AY) is made up of the postal sector (e.g. YO8 9) of which there are approximately 9200, and the unit postcode (e.g. YO8 9AY) made up of on average of 14 addresses, of which there are approximately 1.7 million in Britain (Martin 1992).

Postcodes are the most widely recognised spatial referencing system. If you were to ask an individual which Census ED it falls within they would be dumbfounded, yet ask them their postcode and they will know automatically (Martin 1992). Postcodes have been widely adopted as the primary reference codes by a wide variety of organisations, therefore the use of postcodes within spatial analysis could provide useful and valuable information, which is more difficult to gain from other areal units (Raper et al 1992).

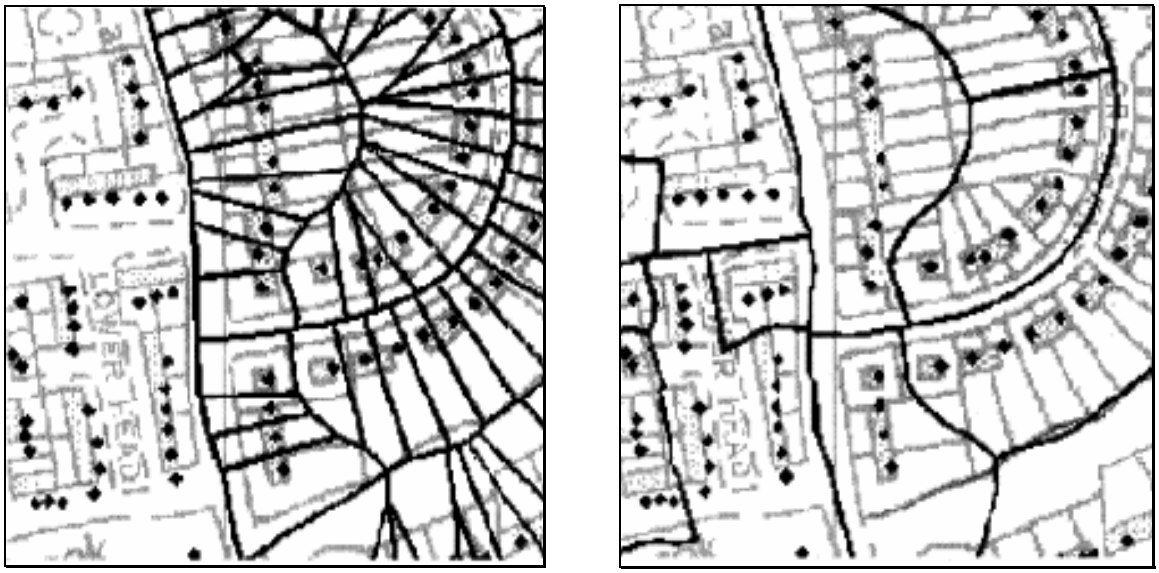
Postal geography clearly has many advantages as a spatial scale at which geographic data can be aggregated. It would be untrue to suggest that postcodes themselves don't have problems in being used as a geographic base, especially when linking them to a different form of geographic aggregation. Postcodes are based on the nearest major city rather than an administrative area, hence postal geography can cross county lines and administrative boarders (for example, Chesterfield in Derbyshire is in the S postal area (Sheffield), which is in South Yorkshire). A problem with using postcodes as a base for socio-economic data, is that they are based on locations of postal delivery rather than the locations where people live, therefore there are large areas of offices and factories etc. which have postcodes but few residents.

2.1 The development of postcodes as areal output units and their changing role in census geography

In the 1971 Census of Scotland EDs fitting into larger postal sectors were used, representing the first step in the British Census being based around postal geography (Raper et al 1992). In 1981 full integration of the Scottish Census with postal geography was decided upon, this contrasts with the more cautious way in which postcodes and EDs were phased together in England and Wales (Martin 2000). It would have seemed beneficial to use the postcode as the design basis for the 1991 Census, which was the case in Scotland. Following the 1987 Chorley report, the 1991 British Census was to be based around postal geography. However, due to spiralling development costs and opinion at the time that it was more important that the areal units used linked reliably to the units used in 1981, the idea was shelved until the planning of the 2001 Census (Martin 1992). The growth of Geographical Information Systems (GIS) through the 1990's led to a dramatic increase in the use of digital geographic data. The popularity of the postcode as a spatial unit was reinforced by the Chorley report of 1987 and 'Postcodes the new geography' by Raper et al. (1992). This led to the reengineering of census output geography in the planning of the 2001 Census count.

Postcodes, which on average contain 14 address, fall below the statistical disclosure threshold level of census output areas, which was set at 25 people in 1981 and 50 people in 1991. In Scotland postcodes were used as the building blocks to create larger output areas rather than the postcodes themselves. The 2001 Census of England and Wales was different to all previous undertakings of the population count, for the first time the collection and output of the Census were based on separate geographies. Data collected during the enumeration of the 2001 Census of England and Wales was for the first time stored at individual level rather than

being accumulated into EDs and then stored, as had previously been the case. When the data is stored at the individual level it can be aggregated in to many different spatial units, including for the first time postal geographies.



(a) (b)

Figure 2: The process of creating areal unit boundaries for the output of the 2001 Census, based on postal address points and postcode boundaries. (Source Martin 2002 pp 10)

Postcodes have no boundaries, so to enable the production of census data based on postal geography postcode boundaries was to be created. Figure 2 explains how this was done, (a) illustrates the creation of Thiessen polygons around the centre of each address using the 'Ordnance Survey Address Point™' (accurate to 1m). Figure 2 (b) demonstrates that by merging the polygons of each address within a postcode together, a postcode boundary is created (Martin 1997 & 2000). Postcodes fall well below the minimum disclosure level of population, which has risen once more for the 2001 Census to 100 people. Several postcodes must therefore be grouped together, in order to create output areas, which have a population above the disclosure level, these output areas do not cross any major postal, or administrative boundaries (Martin et al 2001).

2.2 The Applications of postcodes through data linkage

Linking geographic datasets together provides *'added value'*, it increases the number of applications of the dataset and enables comparisons between datasets to observe consistency and compare accuracy (Raper et al. 1992). The ease of linking of datasets depends upon the format of the data. When only one of the datasets are in the form of postcodes, linking data through postcodes becomes more complicated and other strategies of exploiting the postcode need to be employed, in order to provide a meaningful answer.

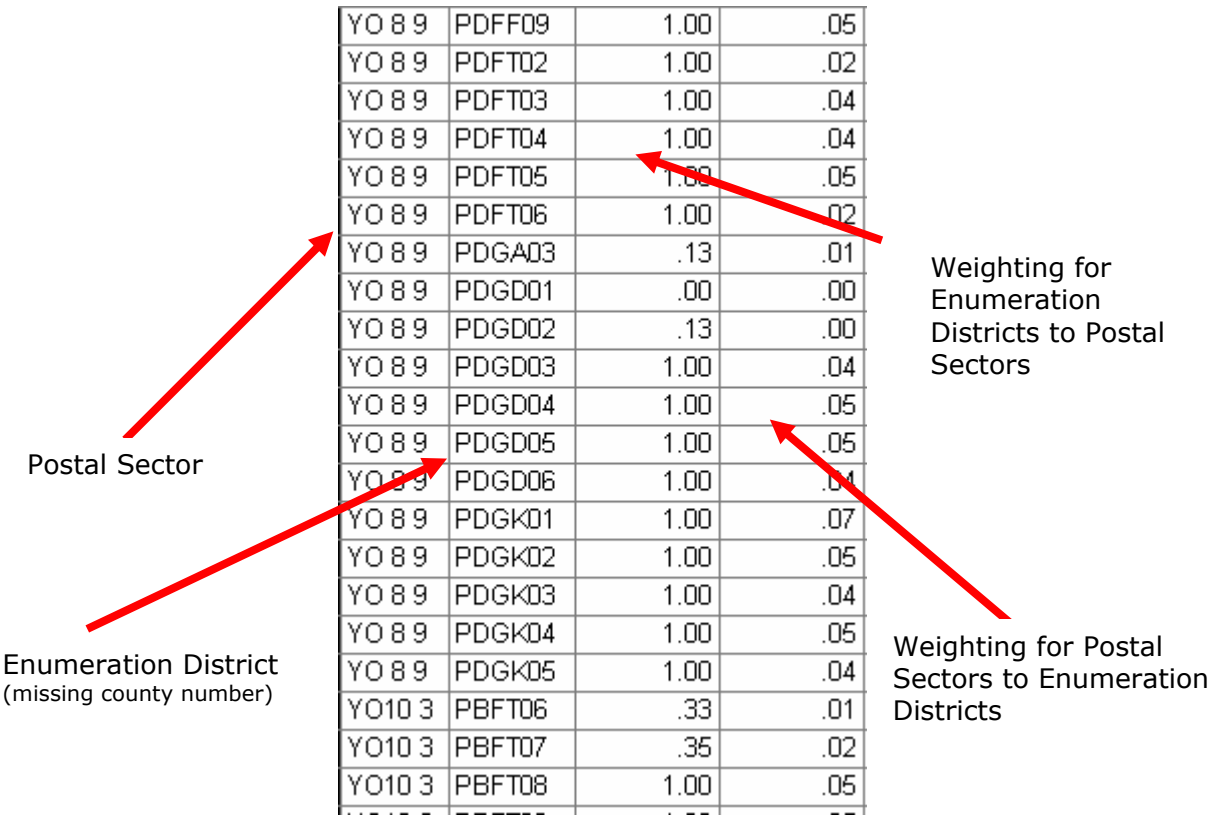
Postcode boundaries enabled the creation of the Office of Population Censuses (OPCS) ED to postcode directory. The OPCS ED to postcode directory was created with the use of population weighted EDs. This was done by finding the population centre of each ED, the ED centriods could then be linked to postcodes using a *'nearest neighbour analysis'* where the centre point of each postcode is assigned to the nearest population weighted ED centriod (Raper et al. 1992). Analysis of the OPCS ED to postcode directory has questioned the accuracy of the process. Long thin EDs or inner city areas where EDs are very small were found to be especially unreliable when linked in this way (Collins et al. 1998). 50% of the postcodes were found to link to the correct ED, just over 40% of postcodes linked to the adjoining ED and the remaining 10% linking to an ED more than one ED away (Raper et al. 1992, Reading & Openshaw 1993). The linking of data based on postal geography to data based around census geography can therefore be seen as an inexact science prone to random difference and inaccuracy.

2.3 The role of GIS in the development of postcoding

The rapid growth of GIS during the 1990's has changed the way, in which spatial data can be created, updated and used. The use of GIS has a two-fold benefit when used in conjunction with postal geography, namely the analysis of postal based geographic information and the maintenance of the geography of the postal system. This is especially relevant to the constantly changing postal geography of Britain, which requires constant updating. Existing boundaries can be changed and new boundaries created and immediately saved to the postcode file, which is stored digitally. Previously all maps of postal geography affected by a change would require reprinting and reissuing. GIS not only enables the rapid and easy update of data but provides a means of quick and simple data analysis (Raper et al. 1992).

2.4 Experian Postal Sector Data

Experian Ltd. is primarily a credit-checking agency, their Micromarketing division is able to produce several geographic data sets from the information that they gain through the main arm of their business. Their neighbourhood classification is supplied as an undocumented data set in terms of its method of production level of accuracy. Although the Experian data is free to academics for research purposes, the charge to commercial users is considerable. The data contains a look-up table (figure 3), which relates to the national Census; this can therefore be used to test the qualities of their product in terms of homogenous groupings and spatial patterns. The Experian postal sector data and the 1991 Census data represent similar data that is aggregated into different areas at different scales, it is therefore likely that this could result in variations in the values produced for certain variables in some areas between the two data sets.



YO 8 9	PDFF09	1.00	.05
YO 8 9	PDFT02	1.00	.02
YO 8 9	PDFT03	1.00	.04
YO 8 9	PDFT04	1.00	.04
YO 8 9	PDFT05	1.00	.05
YO 8 9	PDFT06	1.00	.02
YO 8 9	PDGA03	.13	.01
YO 8 9	PDGD01	.00	.00
YO 8 9	PDGD02	.13	.00
YO 8 9	PDGD03	1.00	.04
YO 8 9	PDGD04	1.00	.05
YO 8 9	PDGD05	1.00	.05
YO 8 9	PDGD06	1.00	.04
YO 8 9	PDGK01	1.00	.07
YO 8 9	PDGK02	1.00	.05
YO 8 9	PDGK03	1.00	.04
YO 8 9	PDGK04	1.00	.05
YO 8 9	PDGK05	1.00	.04
YO10 3	PBFT06	.33	.01
YO10 3	PBFT07	.35	.02
YO10 3	PBFT08	1.00	.05

Figure 3: A sample section from the Experian postal sector to ED look-up table

3. CENSUS GEOGRAPHY

A census is a device for counting people and recording characteristics about them, they are usually carried out once every ten years. Research from census data is becoming increasingly sophisticated. The census process requires three main components, firstly the people who complete census returns, secondly the census offices who are responsible for the collection, editing and production of data and the licensed census partners who disseminate census data and produce value added data products (Rees et al. 2002). Many lesser-developed countries have only recently begun census taking. Some European countries (e.g. The Netherlands) no longer hold a census, as they have very well developed population registers and household survey systems, which enable them to estimate the necessary information from other sources (Rees 1996).

Population censuses held by national statistical offices have the following *favourable* characteristics:

- they are comprehensive;
- they represent the gold standard of data collection;
- they provide data for all geographical scales;
- they provide objective attributes for the population;
- They have the confidence of the people.

They have the following *unfavourable* characteristics:

- they suffer from underenumeration;
- there are always arguments about how to estimate and locate the missing population;
- the data are only collected at periodic intervals;
- the range of characteristics gathered is very limited
- Respondents make a great many "errors" when they fill in the census questionnaire.

(Rees 1996 pp 2)

3.1 The Geography of the 1991 Census of England and Wales

The British Census of population, which is carried out every ten years (1981, 1991, 2001 etc) is the most complete source of information on the number, characteristics and location of the British population (Dale & Marsh 1993). The importance of the Census should not be underestimated, as results from the Census provide the basis of the majority of the reports, findings and activities of both national and local government. The opening and closing of facilities such as schools, hospitals, and clinics are largely based on information revealed by the Census. Census information is also a valuable tool for marketing companies, business

planners and academic researchers. The total cost of the 1991 Census was £135 million but its benefit to the country went far beyond its outlay (Raper et al. 1992). Census data is produced at several different levels as illustrated in figure 4, the smallest of which being enumeration districts.

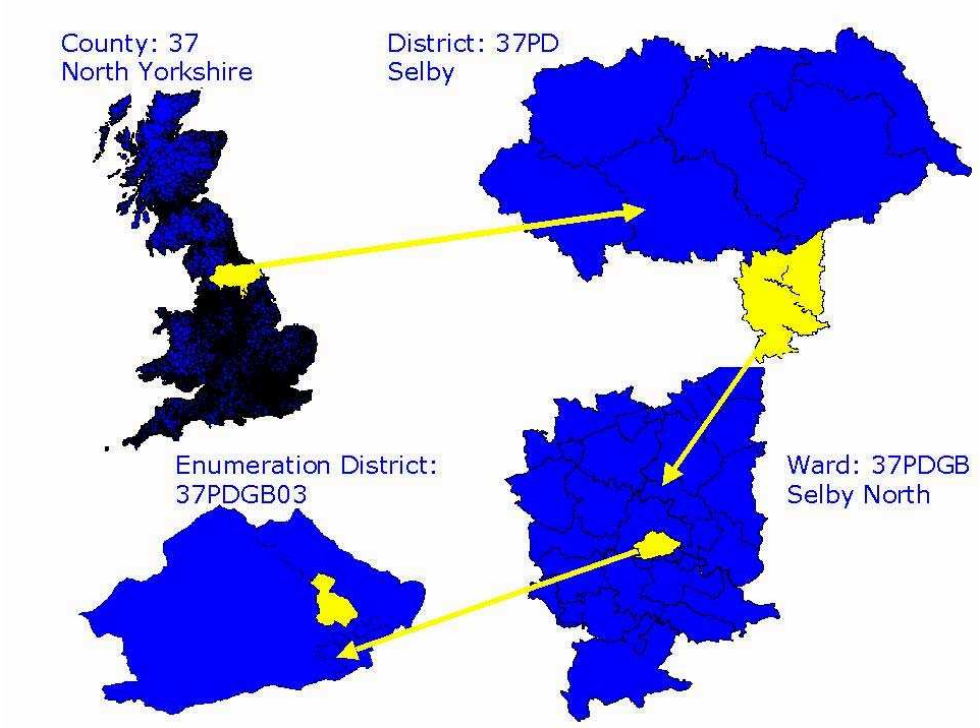


Figure 4: An Illustration of the hierarchy of Census Geography of England and Wales

None of the scales at which the Census is published are representative of any real world features, for example the size of an ED represents the amount of ground a person can cover in a day. The Census data is not published at a more detailed level to ensure anonymity of individual census respondents. No information on individuals or their homes must be released singly or in a form where it can be interpreted from information for an area (Duke-Williams & Rees 1998). EDs that fall below disclosure level have their data added to an adjoining ED and are given a value of 0 population. The 1991 Census was not without problems, famously it suffered from underenumeration where over one million people were not counted in the

Census. Haynes et al. (1995) partly blames the under-count in the 1991 Census for differences observed in ward population estimates between the 1991 Census and National Health Service patient registers. The number of people on National Health Service patient registers on the day of the 1991 Census count exceeded the number of people counted in the Census in the counties of Norfolk and Suffolk (Haynes et al. 1995).

3.2 The Role of GIS in the changing nature of Census geography

GIS has played an ever-increasing role in the development of the census (Openshaw and Rao 1995). Table 1 explains how census management has evolved from an entirely manual process before the 1960's, to computerisation with the advent of the first computers in the 1960's data encoding was computerised. With the birth of GIS geographic encoding could take place on computers and now with new concepts and techniques, the geographical design of census geography can now also take place in the digital environment (Martin 1998).

Table 1: The technological development of Census design through time

Stage	Approximate Date (in UK)	Data encoding	Geography encoding	Geography design
1	Pre 1960	Manual	Manual	Manual
2	1960s	Digital	Manual	Manual
3	1980s	Digital	Digital	Manual
4	2000s	Digital	Digital	Digital

(Martin 1998 pp2)

Figures 5 and 6 show how the design of the census developed in the years between the 1991 and 2001 counts. Figure 5 explains the way in which the 1991 Census was designed consisting principally of inputs, including paper mapping from the previous Census, working practises referring to enumerator workload and boundary placement. Issues relating to the

geography of the Census are resolved by re-digitising, either manually or on screen (Martin 1998).

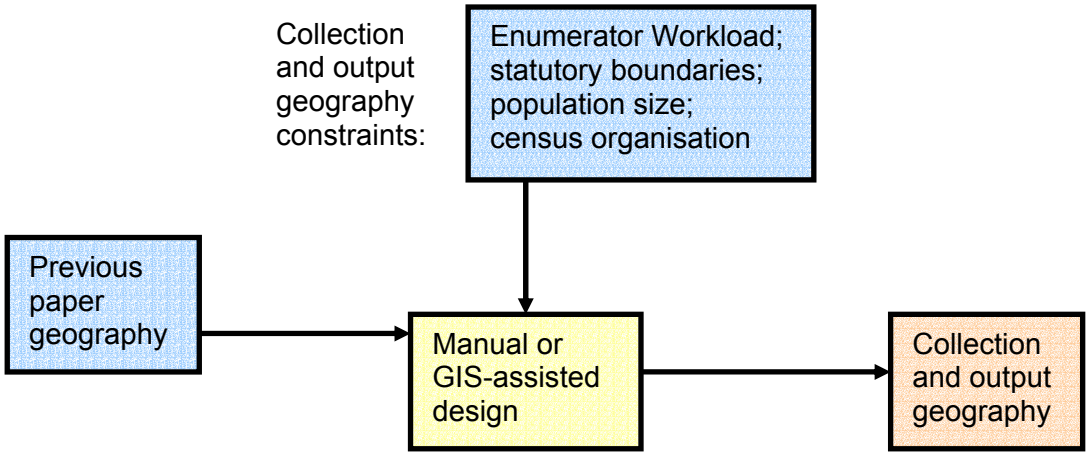


Figure 5: An illustration of the process of census design in England and Wales pre 2001 (Adapted from Martin 1998 pp 5)

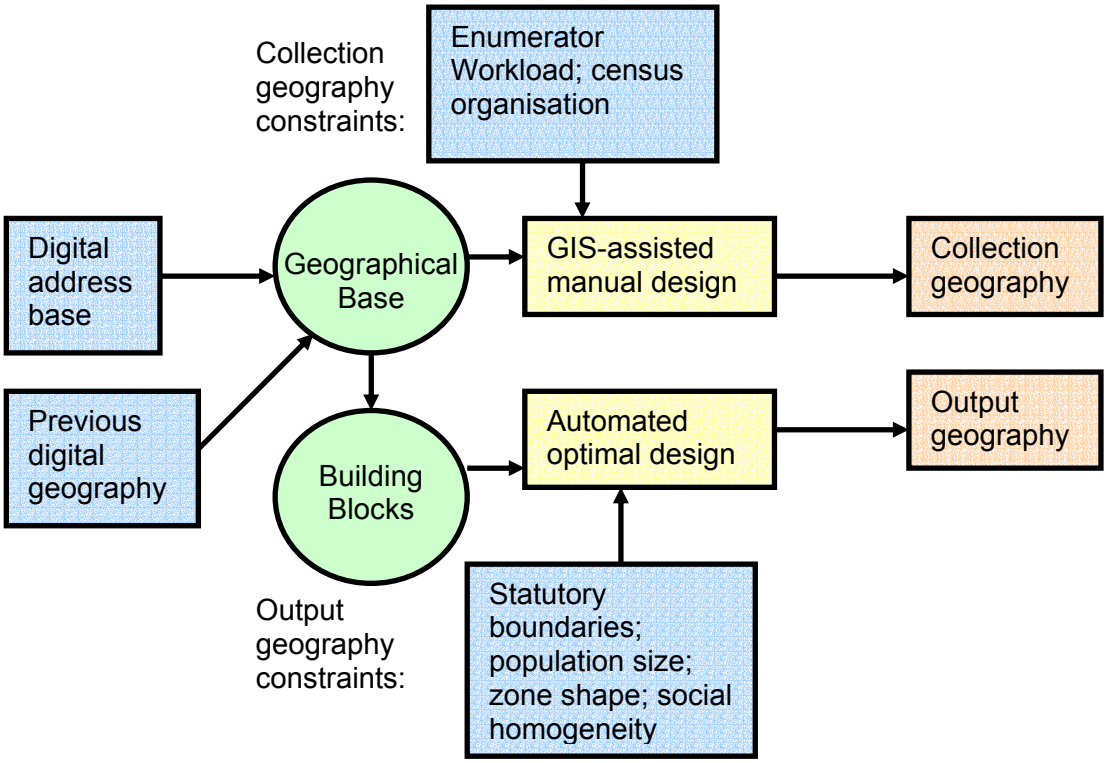


Figure 6: An illustration of the process of 2001 census design in England and Wales (Adapted from Martin 1998 pp 5)

Figure 6 illustrates the way in which the 2001 Census was designed, including the two main differences to its predecessor. The use of digital mapping rather than paper maps as the geographic base for the design of collection geography (Martin 2000), and more importantly the separate design of output geography based on postal geography, which is created through an automatic optimal design process (Martin 1998).

4. ECOLOGICAL FALLACIES AND THE AGGREGATION EFFECT OF THE MODIFIABLE AREAL UNIT PROBLEM

Since the demise of the region as the primary method of geographic study, very few people have expressed a view as to the nature and definition of spatial objects/units being studied (Openshaw & Taylor 1981). The areal units used in the majority of geographical studies are arbitrary, they do not reflect real world geography, and are subject to whims and fancies of whomever aggregated the data (Openshaw 1984a).

4.1 Ecological Fallacies

An ecological fallacy arises when statistics relating to an aggregated areal unit are incorrectly assumed to represent an individual or a smaller unit within the original area (Tranmer & Steel 1998). Figure 7 is a simple example of ecological fallacy based upon census geography, it displays the number of males who live in the area who are pensioners. At the most detailed level over 16% of males are over 65. However if a wider view of the area is taken this level drops to around 6%. If information at the ward level is used to represent data at ED level then ecological fallacy could occur seen (Tranmer & Steel 1998). By using ward level data to establish statistics for EDs uniformity is assumed within the ward therefore ignoring possibly significant local variations (Holt et al. 1996).

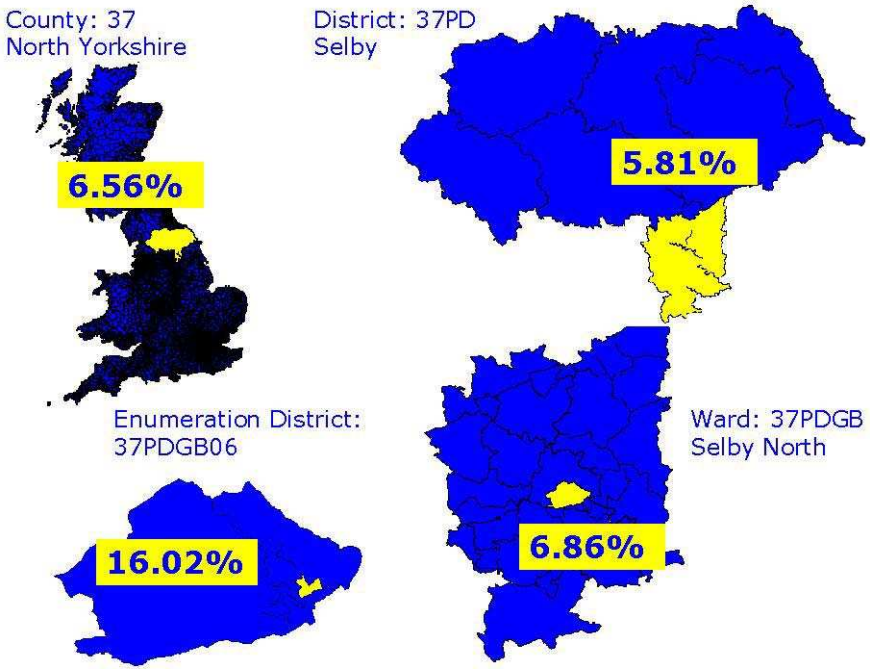


Figure 7: An example of ecological fallacy using census data (percentage of Males above 65 years of age in Selby, North Yorkshire)

in 1991 census the threshold level of which population had to exceed to be published was set at 50 people, below this level data was suppressed. Therefore the level/size of area at which the census information is released is critical. If the area is too small the data is suppressed or not available. If the area is too large the data is smoothed to a level where it becomes unrepresentative of its constituent parts. In both instances the likelihood of inaccuracy and ecological fallacy is great (Raper et al. 1992).

4.2 The Modifiable Areal Unit Problem and the consequences of data aggregation

A dataset can appear to have significantly different values depending upon how and where the data is aggregated, this is called the Modifiable Areal Unit Problem (MAUP) (Monmonier 1996, Openshaw & Taylor 1981). The MAUP is a fundamental geographic problem that is endemic to all studies of spatially aggregated data (Wrigley 1995). Sensitivity to MAUP is unpredictable and further conclusions cannot be made, as the severity of the problem appears

to be specific to each dataset (Openshaw 1984a). The MAUP is especially relevant to this study in that, it examines data for the same area aggregated in two different ways. It is likely that some of the differences observed between the two datasets can be attributed to the fact that they are aggregated into two contrasting geographic systems.

It is possible to produce significantly different correlation rates by choosing an appropriate size or shape of unit area on which to base a study. Therefore the results of studies based on modifiable units will depend on the units used (Openshaw 1984a). With this in mind, the use of the postal sectors could produce significantly different results in comparison to the use traditional census boundaries such as EDs. Figures 8 and 9 show an example of MAUP. Figure 8 represents a grid of 24 fictitious census EDs each one square mile in area. The numbers inside the boxes represent the number of people who live in each ED.

624	587	543	237	321	501
475	509	652	720	526	175
442	605	731	551	599	116
549	574	504	585	235	376

Figure 8: Sample dataset to illustrate the effect of aggregation on areal units

Figure 9 (a – l) illustrate the values that are created if the values in figure 8 are aggregated in different ways, the numbers can be made to look significantly different by splitting the grid in different places. Split into two equal areas horizontally (figure 9 a) both areas have the same value, when split into two equal areas vertically (figure 9 b) there is a 27% difference in value between the two areas. If the region is split diagonally in two different ways thus producing four areas of the same size and shape as in (figure 9 c & d), all values produced are different to each other and the previous examples.

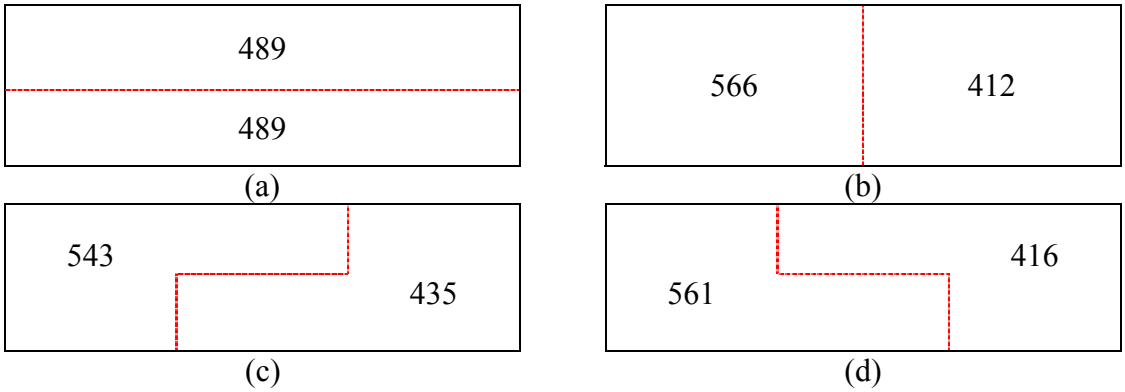


Figure 9 a – d: An illustration of the effect of data aggregation of socio-economic data

It is clear that by splitting the grid in different places it is easy to make the population of the two areas look both uniform and irregular. By splitting the grid into different numbers of areas, of varying shapes and sizes further manipulations of the population of the area can be made. The average population of the whole area (figure 9e) is 489 people per square mile, by aggregating the data into six areas (figure 9f) a difference of up to 40% can be produced. Many more different values can be produced by aggregating the data into different sized and shaped areas illustrated in figure 9 g - l.

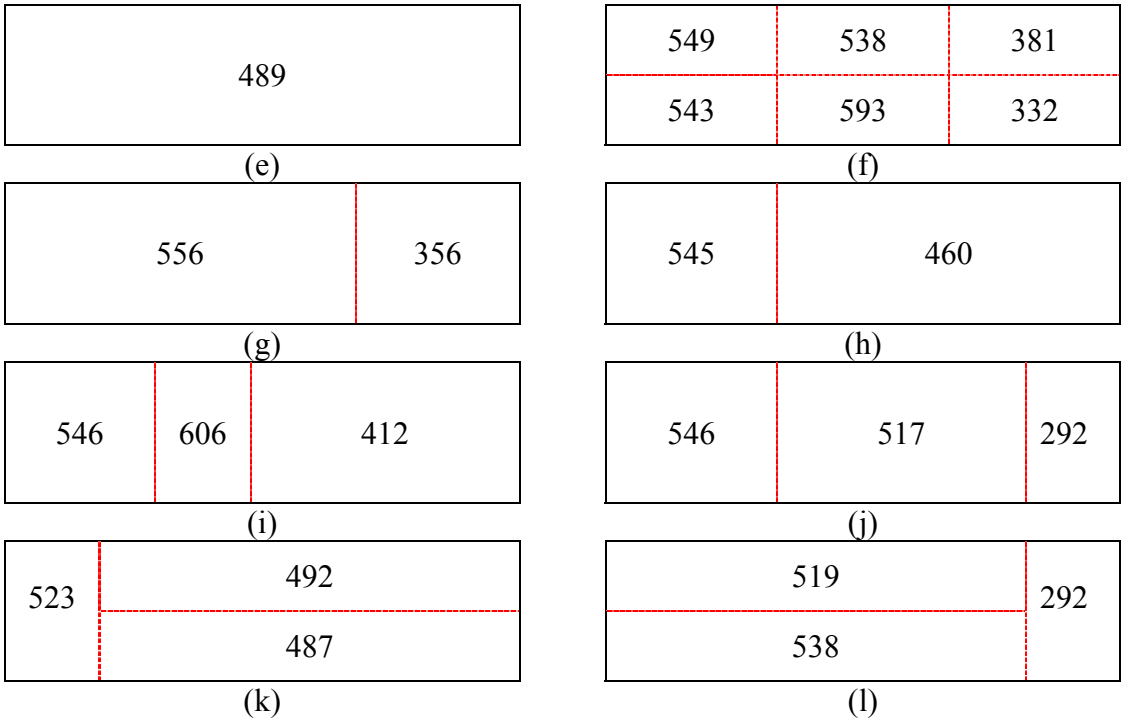


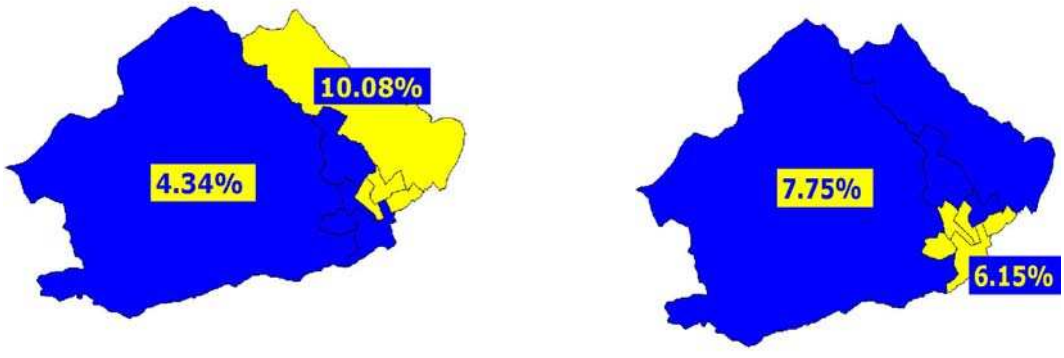
Figure 9 e – l: An illustration of the effect of data aggregation of socio-economic data

Table 2 demonstrates that by aggregating the data in many different ways it is possible to make the red square in figure 8 have many different values ranging from 412 (-139/25%) to 593 (+42/8%) a difference of 181 or 31%. None of the aggregations of the data kept the original value of the square.

Table 2: The effect of data aggregation on the sample dataset (shown in figure 8)

Example Number	Original Value	(a)	(b)	(c)	(d	(e)	(f)
Value	551	489	412	435	561	489	593
Difference		-62	-139	-116	+10	-62	+42
% Difference		-11	-25	-21	+2	-11	+8
Example Number	Original Value	(g)	(h)	(i)	(j)	(k)	(l)
Value	551	556	460	412	517	487	538
Difference		-62	+5	-91	-139	-34	-64
% Difference		-11	+1	-17	-25	-6	-12

Figure 10 displays a simple example of the MAUP within census geography, it is clear that the way in which the data is aggregated is responsible for two very different patterns of values produced for the same area. Aggregation (a) produces two extreme units with one value being significantly larger than the other. In contrast (b) produces two areas with comparatively similar values.



(a) (b)

Figure 10: An illustration of the effect of the modifiable areal unit problem (MAUP) using census geography (example: a census ward in Selby, North Yorkshire percentage of males above 65).

There is no real solution to ensuring that multi-level and scale aggregations of data that display the similar geographic patterns whatever the aggregation. The only real way of getting round the problem is to store all data in the least aggregated form possible (i.e. individual level where possible). When stored at this level the data can then be aggregated to the required level or scale, whether this is postal sectors, census EDs, or electoral wards.

5. HOW WELL DOES THE EXPERIAN DATA REFLECT THE CENSUS?

The Experian dataset was too large to examine in its entirety so a sample was selected as a representation of the dataset. Figure 11 shows the 11 postal areas, representing 1005 of the 9216 postal sectors (10.9%) randomly selected. The sample areas are spread throughout the country, in both urban and rural areas to give a representative cross-section of both the population and geography of Britain.

AB	Aberdeen
CA	Carlisle
CF	Cardiff
E	London (East)
EH	Edinburgh
GU	Guilford
NG	Nottingham
NR	Norwich
TQ	Torquay
WR	Warwick
YO	York

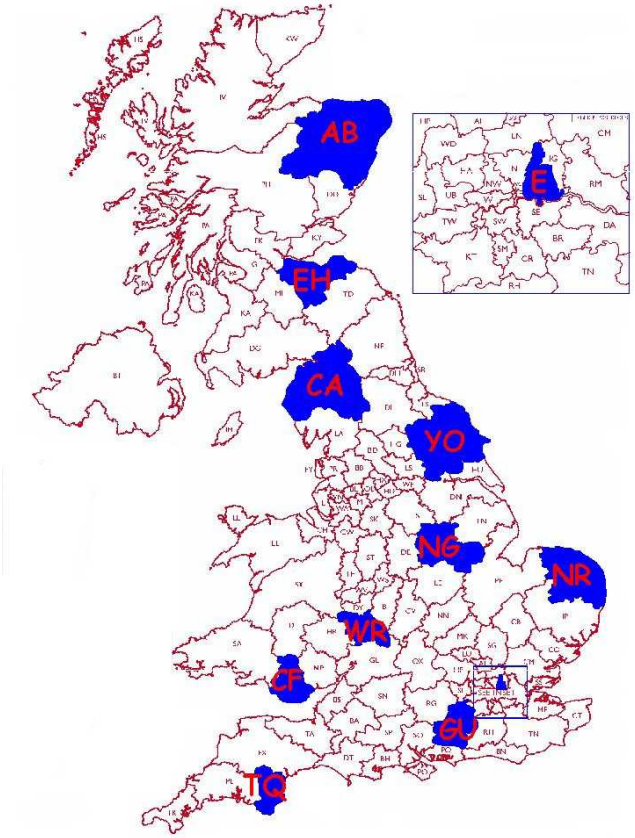
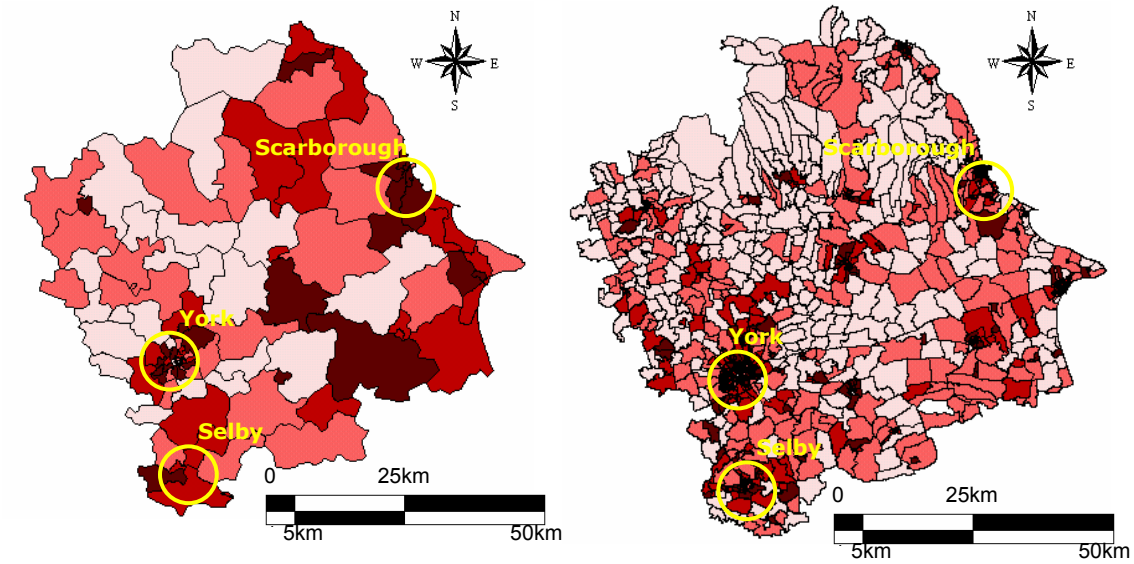


Figure 11: The location of the 11 sample Postal Areas used in this study

Although both datasets represent socio-economic data for Britain very few variables in each of the datasets have identical data, which is necessary to compare the accuracy of the Experian data to that of the census. Fields that are represented on both the Experian and Census datasets are the population and car ownership variables. From these total population and total numbers of cars were chosen as the link between the two datasets. Scottish sectors should be more accurate as the Scottish census is based on postal geography. However the postal geography of Scotland has changed significantly since the time of the 1991 census and many of the sectors have been renamed and changed in their geography. Therefore the Scottish postal sectors in the Experian datasets don't relate to those when the 1991 census was published. The Experian look-up table links Scottish postal geography to the 'output areas' in the Scottish census (equivalent to EDs in England and Wales).

In order to compare the two datasets they must be viewed at the same geographical level and aggregated using the same geographic system so that the data values are of a similar size. Figure 12 compares the population of the YO (York) postal area at two different aggregations. The map in figure 12 (a) illustrates the data aggregated by postal sectors, figure 12 (b) shows the data aggregated by census EDs. The two datasets seem to display the same general pattern, it is relatively easy to identify the major settlements in the area such as York, Scarborough, and Selby in both sets of data.



Quartile	Postal Sector Value	ED Value
Lower Quartile	203 - 2715	0 - 245
Second Quartile	2795 -4905	246 - 398
Third Quartile	4937 - 6248	399 - 497
Upper Quartile	6383 - 11197	498 - 1128

Figure 12: An initial comparison of two datasets aggregated using different spatial systems (population in North Yorkshire), (a) shows Experian data applied to postal sectors. (b) Shows Census data applied to enumeration districts.

There are two main reasons for difficulties in comparing the different aggregations at different scales. The populations of the postal sectors are several times those of the EDs and the smaller size of the EDs displays greater local variations that are not visible in the postal sector data, and the Experian data appears more smoothed. The data cannot be compared at ED level, as this is much more detailed than the postal sector data, it would assume uniformity within the postal sectors if the Experian data is applied to EDs this would result in an ecological fallacy. Therefore the census data at ED level must then be applied to the postal sectors using the weightings in the third column of the Experian look-up table (as seen in figure 3).

5.1 Descriptive statistics and correlations between and within the datasets

Table 3 compares descriptive statistics about the population variable in the Experian dataset and the Census dataset applied to postal geography with the use of the ED to Postal sector look-up table. The two datasets appear similar. However the Experian dataset has greater maximum and mean values and a larger standard deviation.

Table 3: Descriptive statistics of the differences between the Experian and Census datasets, population variable

	Min	Max	Mean	Standard deviation
Experian	0	18,571	6,665	3,534
Census	0	17,563	6,061	3,285

Table 4 compares descriptive statistics about the cars variable in the Experian and the Census datasets applied to postal geography with the use of the ED to Postal sector look-up table. It is clear from these simple statistics that there are obvious differences between the two datasets for the number of cars variable. The maximum value of car ownership in the Experian dataset is four times that of the Census dataset. The mean of the Experian dataset is over 20% greater than the Census and the standard deviation of the Experian dataset is over 30% greater than the Census dataset. This suggests that there are pertinent differences between the number of cars in the two datasets.

Table 4: Descriptive statistics of the differences between the Experian and Census datasets, cars variable

	Min	Max	Mean	Standard deviation
Experian	0	28,723	2,808	1,748
Census	0	7,407	2,367	1,304

As the two datasets represent exactly the same information near perfect correlations should be expected rather than significant ones. If the datasets are not significantly correlated this

would represent large differences within one or both of the datasets. To assess the strength of correlation between the datasets both Pearson Product Moment Correlation and Spearman's Rank Correlation were both used.

Figure 13 displays the relationship between the population variable in the two datasets. This exhibits not just a significant but near perfect correlation between the population of the postal sectors in the two datasets. Correlations of 0.977 Pearson's two tailed test significant at the 0.01 level, and 0.978 Spearman's Rank two tailed test significant at the 0.01 level were observed. This high level of correlation should be expected, the datasets are meant to be exactly the same so anything but a near perfect correlation would be surprising. This result suggests that the population numbers in the two datasets are not only nearly identical in terms of rank but also actual values. However there are some outliers, one is especially noticeable and the question should therefore be asked whether these point to differences in the Experian or census datasets?

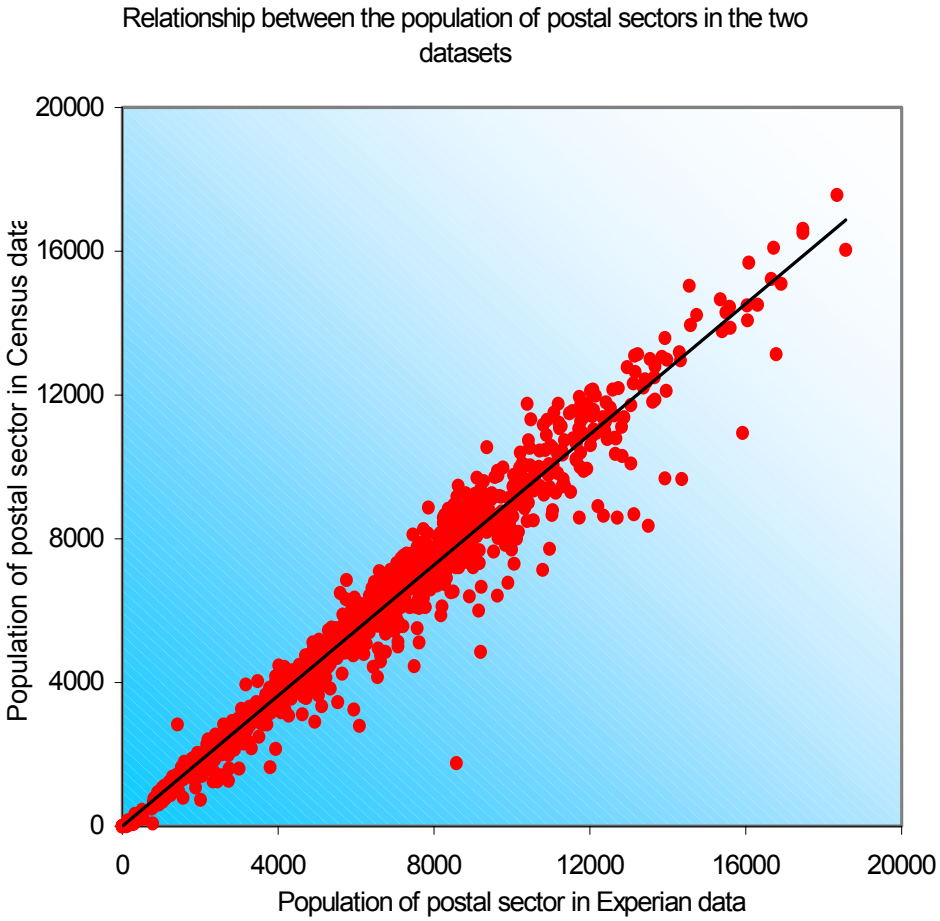


Figure 13: The relationship between the population of postal sectors in the Experian and Census datasets using the look-up table to apply Census data to postal sectors.

Figure 14 demonstrates the relationship between the population variable in the two datasets, showing a near perfect correlation for the Spearman's Rank correlation (0.902 two tailed test significant at the 0.01 level). However the simple Pearson Correlation although showing a significant correlation (0.661 two tailed test significant at the 0.01 level) is far from perfect. This suggests that the car ownership figures have a very significant level of correlation in terms of their rank. However the Pearson's correlation suggests that the values within the two datasets are not as correlated and could suggest a significant differences between the two datasets. Figure 14 reveals that the number of cars in the Experian dataset is greater than in the census dataset. This would account for the lower Pearson correlation between the two

datasets for the car ownership variable. However, looking at figure 14, a few outliers could be responsible for the lesser level of correlation, as the overall trendline (solid line) shifts away from the main trend of the majority of the dataset (dashed line).

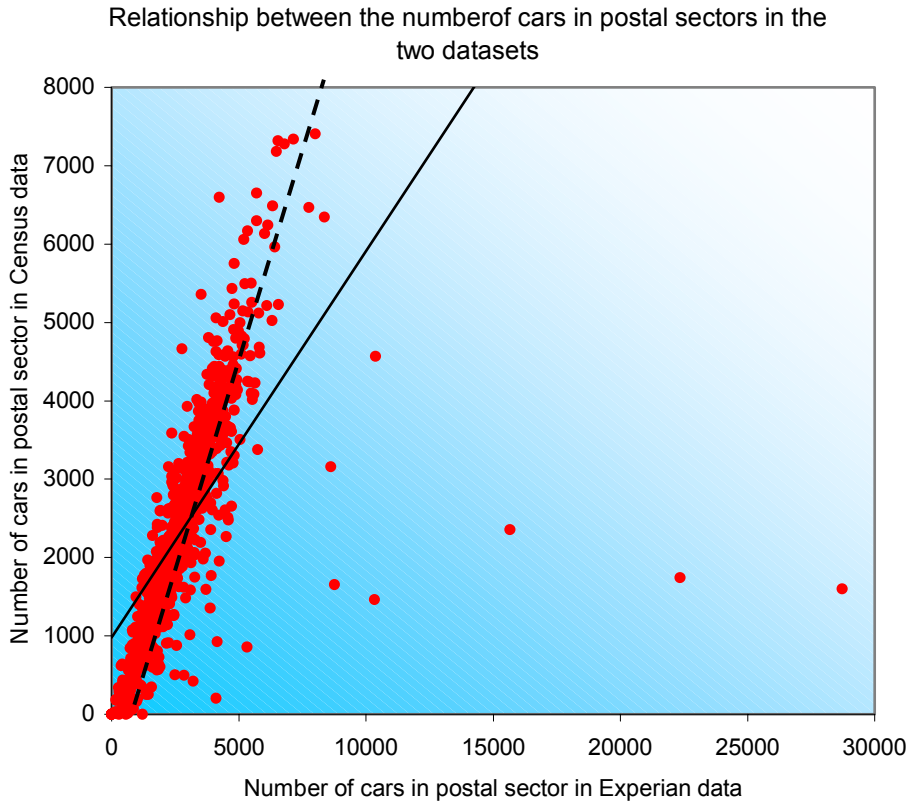


Figure 14: The relationship between the number of cars in postal sectors in the Experian and Census datasets using the look-up table to apply Census data to postal sectors.

Figure 15 reveals that the percentage difference between the value in the Experian and Census datasets. In the majority of cases the value in the Experian dataset is higher than in the census for both population and car ownership variables, very few values were higher in the census dataset for both variables. The car ownership variable displays more difference than the population variable.

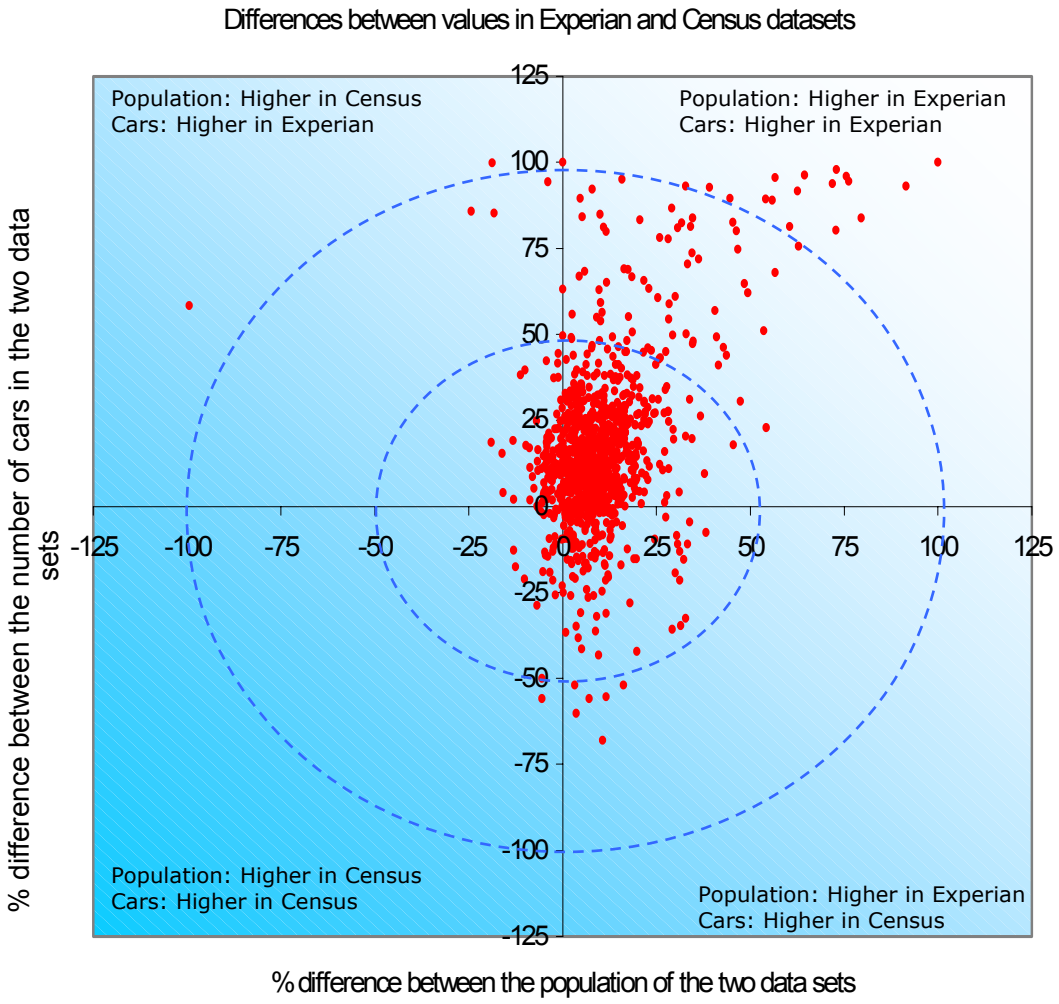


Figure 15: The distribution of the differences between the values in the Experian data, and the Census data, which has been applied to postal sectors using the look-up table.

There is a significant correlation between the variations observed between the two datasets for both variables. A correlation of 0.415 Pearson's (2-tailed test significant at the 0.01 level), and a 0.284 correlation Spearman's Rank (2-tailed test significant at the 0.01 level). The correlation of the differences between the two variables in the two datasets suggests that the same factor is responsible for a substantial amount of the difference between the two datasets. This therefore indicates that a significant quantity of the differences between the two datasets is in the way they are linked together through the look-up table.

Figure 16 shows car ownership per person for both datasets, this is a combination of both the population and car variables. Tests of correlation were run on the number of cars per person calculated for each dataset it was found that, Pearson's 2-tailed test (at the 0.01 level) produced a not significant correlation of 0.014. This is a surprising result suggesting that there is no correlation between the two datasets, the correlation is greatly affected by several extreme outliers. The Spearman's Rank correlation, which is less affected by outliers, produced a correlation of 0.697 (significant at the 0.01 level). The trendlines in figure 16 represent the effect that the outliers have on the relationship between the two datasets, the solid trendline represents the trend of the whole dataset, the dashed line represents the trend within the data ignoring outliers and represents a greater relationship between the two datasets.

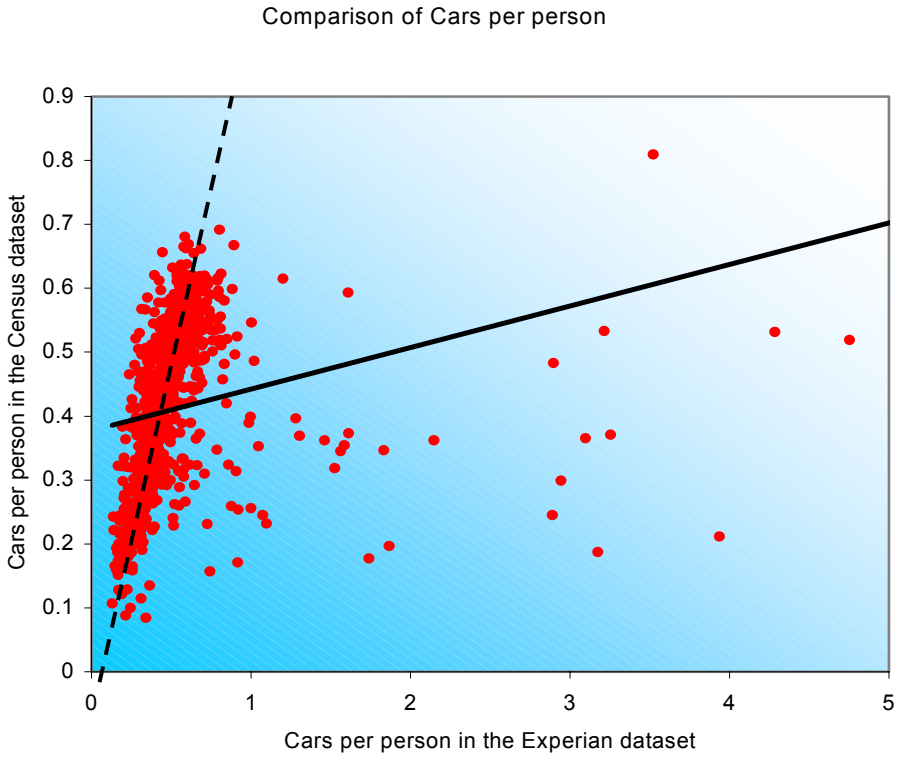


Figure 16: A comparison of the number of cars per person in postal sectors, for the Experian and Census data sets using the look-up table to apply Census data to postal sectors. (N.B. outlier (Experian 303.25: Census 0.416842) removed from graph to enable display)

5.2 Erroneousness and observed differences between the two data sets.

There is a 9-year time difference between the two datasets, the census from 1991 and the Experian data projected for the year 2000. The Experian data set shows 8.30% population growth shown by the 4,337,192 million extra people in the Experian dataset, in comparison to the census data. The Experian data also contains 4,274,428 (20.82%) more cars than the census dataset. These differences could be due to a number of factors, including, real time growth in the period between the two datasets, differences in the Experian dataset, and the well-documented underenumeration in the 1991 census. This difference between the two datasets has been taken into account when making comparisons between them and creates an range of fuzzy accuracy when the Experian value is greater than the census value by an amount less than the average difference between the two datasets for that variable.

Table 5: Selected examples of the 642 postal sectors in the Experian dataset, which have more cars than people.

Sector	Population	Cars	Difference	Cars per person
SA99 1	0	3,974	-3,974	Infinite
TW 6 1	0	1,288	-1,288	Infinite
EC3N 3	1	328	-327	328.00
NG17 6	4	1,213	-1,209	303.25
L 38 2	1	163	-162	163.00
MK 6 1	7	839	-832	119.86
SN 5 6	5,034	106,644	-101,610	21.18
M 33 7	4,342	76,817	-72,475	17.69
M 5 2	6,786	76,489	-69,703	11.27
SL 1 4	856	52,738	-51,882	61.61
B 16 0	7,590	56,894	-49,304	7.49
HU 7 4	23,899	4,707	19,192	0.20
Average	6,140.60	2,691.28	3,449.32	0.44

The large number of obvious inaccuracies in the Experian data set is illustrated perfectly by the 642 sectors, which have more cars than people. Some of the worst examples (shown in table 5), SN5 6 (Swindon) has 5,034 people and 106,644 cars, 21 cars for every person that

lives in the sector. This is not an area of incredible car ownership, it is the location of the Honda car factory where all the cars have been pre-registered before they are sold. There are 145 sectors that have 0 residents but have cars the worst case being SA99 1, which has 3,974 cars, the question here is how do people who do not exist own cars? These incongruities are almost certainly point to differences within the car variable in the Experian dataset. It is possible that some of these differences occurred during the input of data. However the number of this type of difference suggest that they are endemic to the original source of the Experian cars variable data

At the ED level there are 14 EDs that have more cars than people (table 6), however unlike the Experian dataset these high rates of car ownership do not necessarily show serious differences within the data for several reasons. Firstly, only a small proportion of the over 100,000 EDs have more cars than people. EDs are much smaller than the postal sector in terms of population, therefore local variations can have a much bigger effect on the result, for example someone with a large personal car collection could significantly affect the rate for the whole ED. The ED with the highest rate of car ownership is 01AAFT01 in Inner London, which has 31 people and 64 cars, which equates to 2.06 cars per person. The areas with more cars than people in the census data all have a less than average population for a ED. This is not the case in the Experian data where many of the instances have larger than average population.

Table 6: The 13 English Census EDs, which contain more cars than people

ED	Population	Cars	Difference	Cars per person
01AAFT01	31	64	-33	2.06
03BNFK30	31	46	-15	1.48
01ADFR03	71	105	-34	1.48
16FAFP09	54	74	-20	1.37
01APFZ30	76	93	-17	1.22
01APFZ27	43	50	-7	1.16
25JQFM01	45	52	-7	1.16
25JLFK09	107	120	-13	1.12
01ADFR02	137	150	-13	1.09
01AGFH36	62	66	-4	1.06
04BYFQ21	52	55	-3	1.06
01AMFJ30	98	103	-5	1.05
03BRFL23	72	75	-3	1.04
03BRFL22	100	104	-4	1.04
Average	419.63	168.02	251.61	0.39

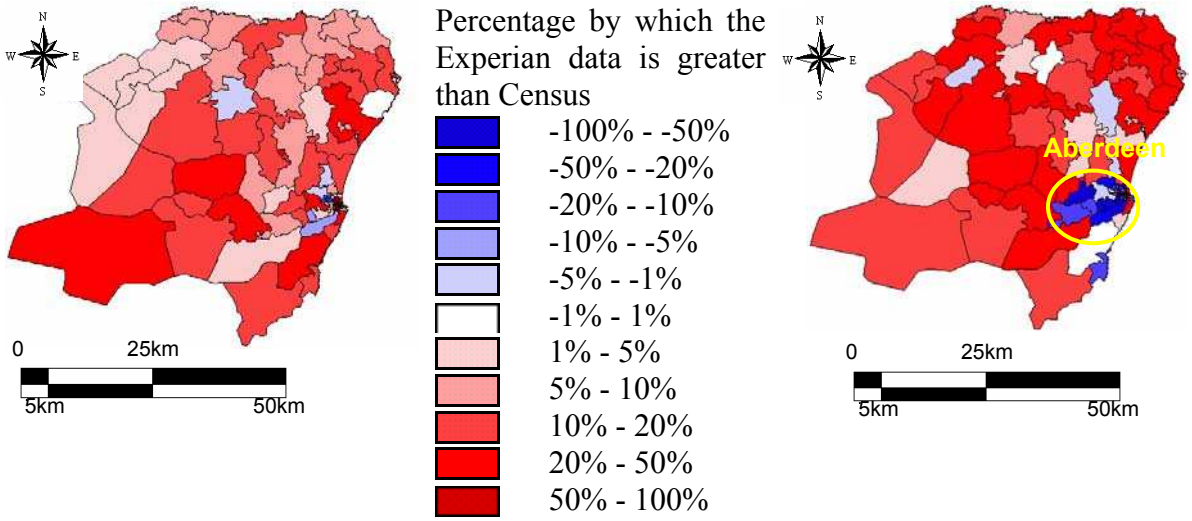
If the Experian data were correct a comparatively larger percentage of instances of more cars than people would be expected in the census data than the Experian data as the EDs are smaller than the postal sectors and are more likely to show local extremes. In fact only 0.013% of EDs (1 in every 7700) in the census data set had more cars than people, compared to 6.966% of postal sectors (one in every 14) in the Experian postal sector data.

The Experian car data is likely to have come from a separate data source to it's population data as it contradicts the population, even the aggregations of car data e.g. insurance group/type of car contain entries when there is 0 population. Once the data from the different sources was put together it would appear that no basic check of accuracy or consistency has taken place.

5.3 Neighbourhood variations between the two datasets

It has been established that the Experian dataset contains significant variations from the census data, but the question remains, whether the differences between the two datasets show any geographic patterns. Figures 17 – 27 show how the percentage differences between the two datasets vary geographically. The time difference between the two datasets means that many of the instances where the value for the Experian dataset is slightly higher than the census could be down to growth over that period. However, instances where the census data is greater than the Experian data are unlikely, as on average the Experian values are 8.3% greater than the census in terms of population and 20.8% greater than the census in terms of number of cars. However the large number of outliers observed in the Experian cars data suggests that the large difference between the two datasets for this variable could be as much due to differences in the dataset as an increase in car ownership. Therefore, when the Experian data value is significantly greater than census or the census value greater than the Experian value by any level this is likely to represent an difference in either the Experian dataset or the look-up table used to join the EDs to the postal sectors.

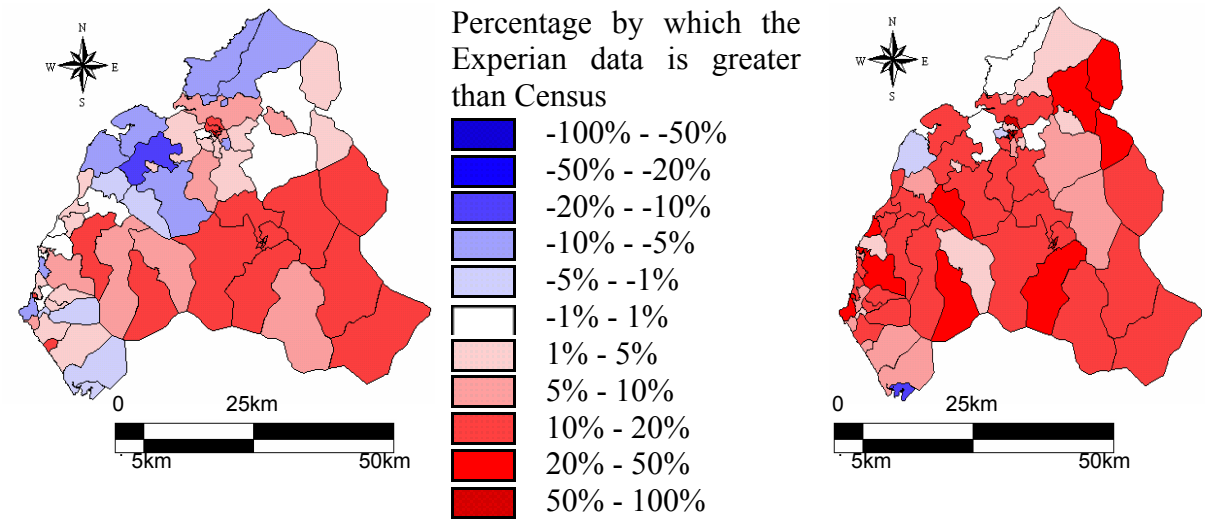
The most noticeable pattern that figure 17 displays is in the city of Aberdeen where the census data is greater than Experian data. This is in contrast to the rest of the area where the Experian dataset is significantly greater than the census data in most cases. The question therefore arises whether this suggests that the Experian dataset underestimates values in urban areas? Figure 17 also illustrates that the cars variable shows greater variation in the value of the two datasets than the population variable. This demonstrates that the cars variable contains more a more apparent geographic difference between the two datasets than the population variable.



(a) (b)

Figure 17: Postal Area AB (Aberdeen) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

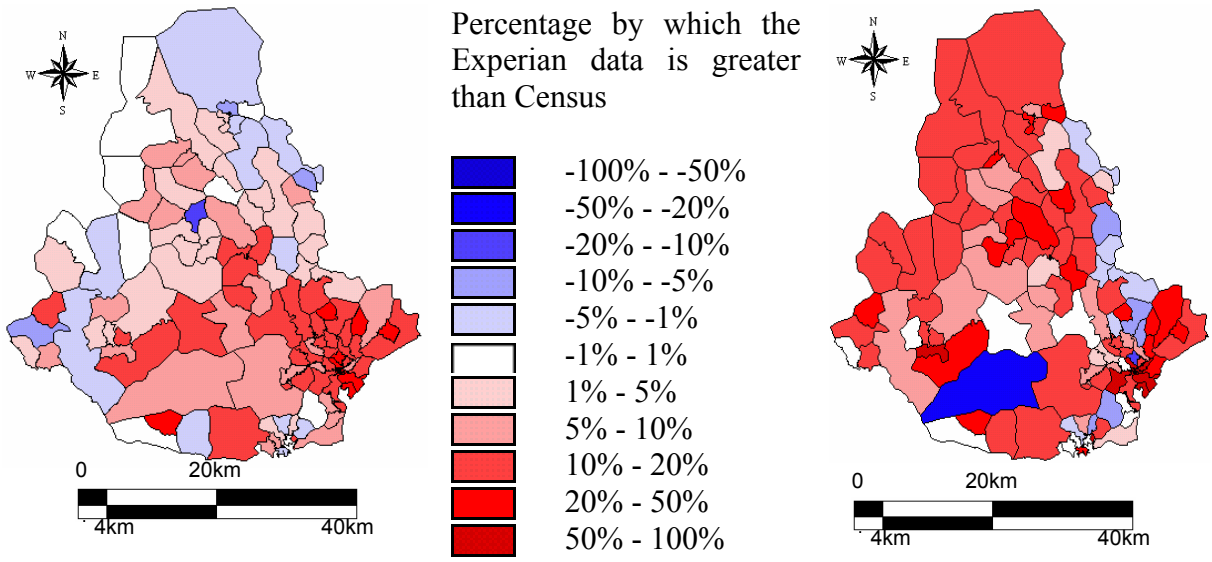
Figure 18 displays greatest variation in the population variable for the CA postal area which is in contrast to the previous example The large underestimation of the number of cars in Aberdeen is not repeated in Carlisle the main urban area in the CA postal area.



(a) (b)

Figure 18: Postal Area CA (Carlisle) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

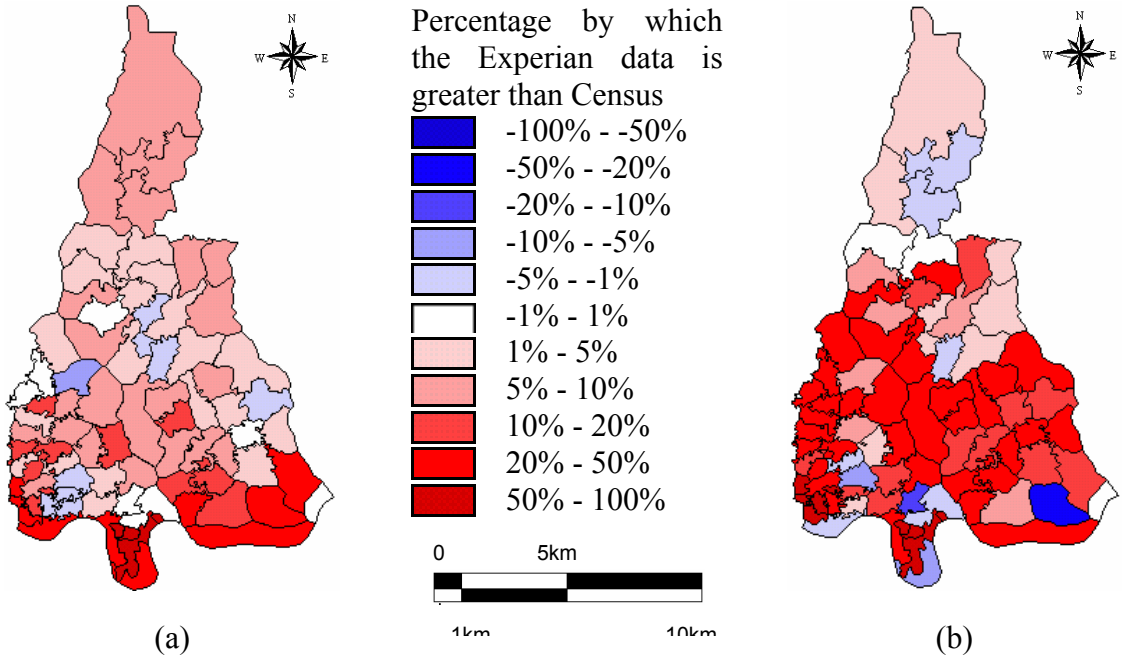
Figure 19 shows postal area CF illustrating significant variations between the two datasets, interestingly the location of sectors that show the census values greater than the Experian values differs between the population and cars variables. This suggests that there is much inconsistency between the variables in the Experian dataset. If the population and cars data were obtained from the same data source whether it is accurate or not the difference for each variable should be similarly large or small, variation in the location of high and low difference values for the two different variables. This suggests that they are from different data sources and have different levels of accuracy.



(a) (b)

Figure 19: Postal Area CF (Cardiff) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

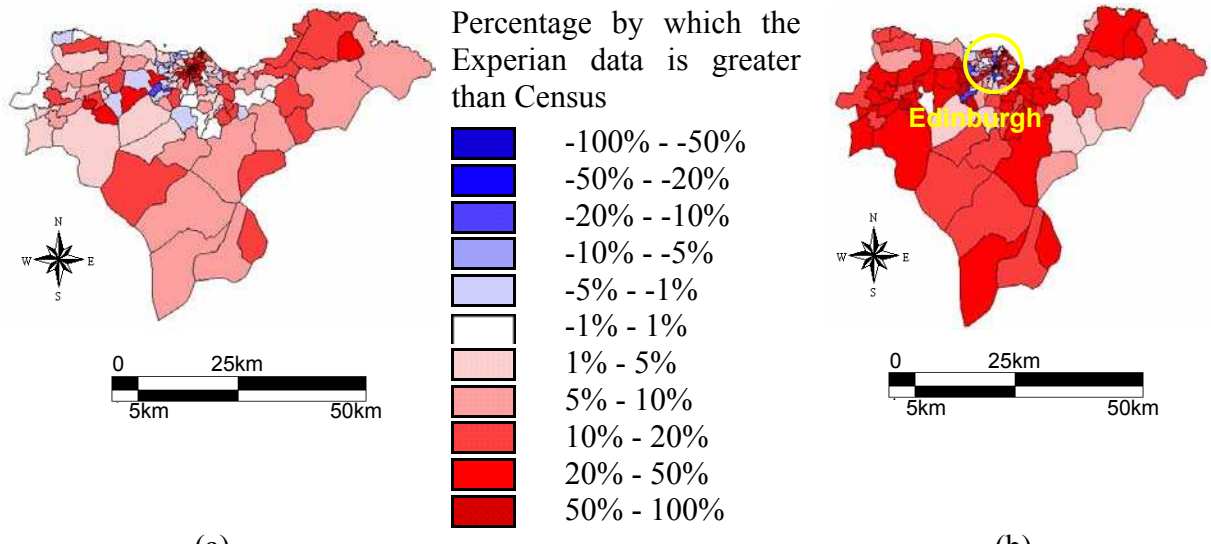
Figure 20 illustrates that postal area E has no distinct pattern, although the difference, both positive and negative, appears to increase towards the south for both variables.



(a) (b)

Figure 20: Postal Area E (London East) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

Figure 21 reveals how in postal area EH the Experian dataset underestimates both the population and the number of cars for several sectors around the city of Edinburgh, a very similar pattern to that seen in the AB postal area (figure17).



(a) (b)

Figure 21: Postal Area EH (Edinburgh) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars.

Figure 22 appears to show no distinct patterns, random both positive and negative difference can be seen within the Guilford area. There is not as much clustering evident as in other postal areas. However some of the sectors with higher census than Experian values do relate to some of the towns in the area.

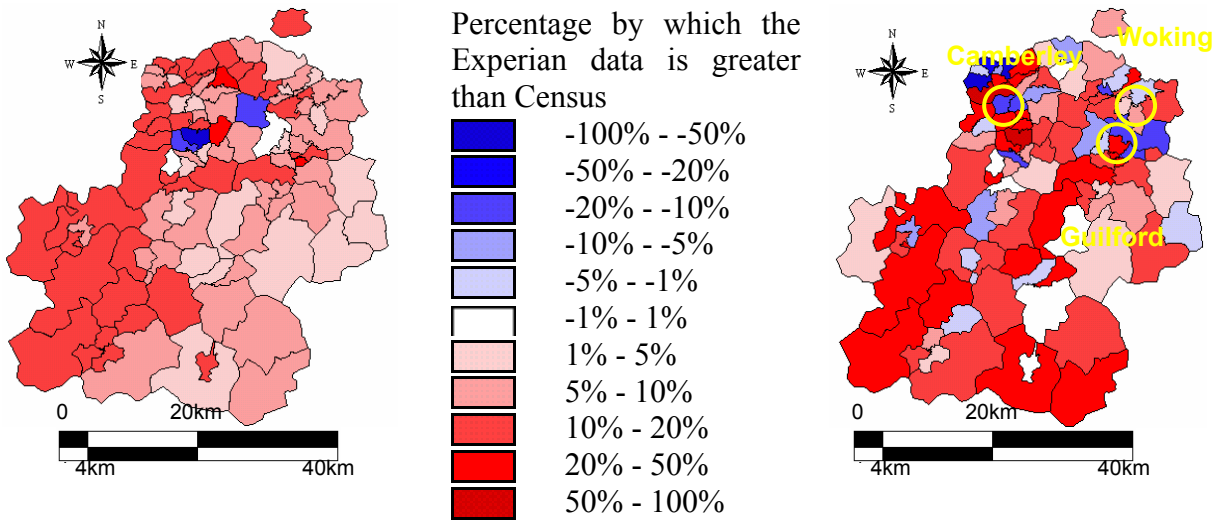


Figure 22: Postal Area GU (Guilford) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

Figure 23 shows how Experian data for postal area NG appears to underestimate cars the nearer to the main urban area of Nottingham, further away from the city the Experian data seems to be higher than the census. This reinforces the patterns observed in the AB (figure 17) and EH (figure 21) postal areas. A pattern seems to be appearing for the car variable, the Experian data seems to underestimate the number of cars in urban regions and overestimate the number of cars in the more rural locations.

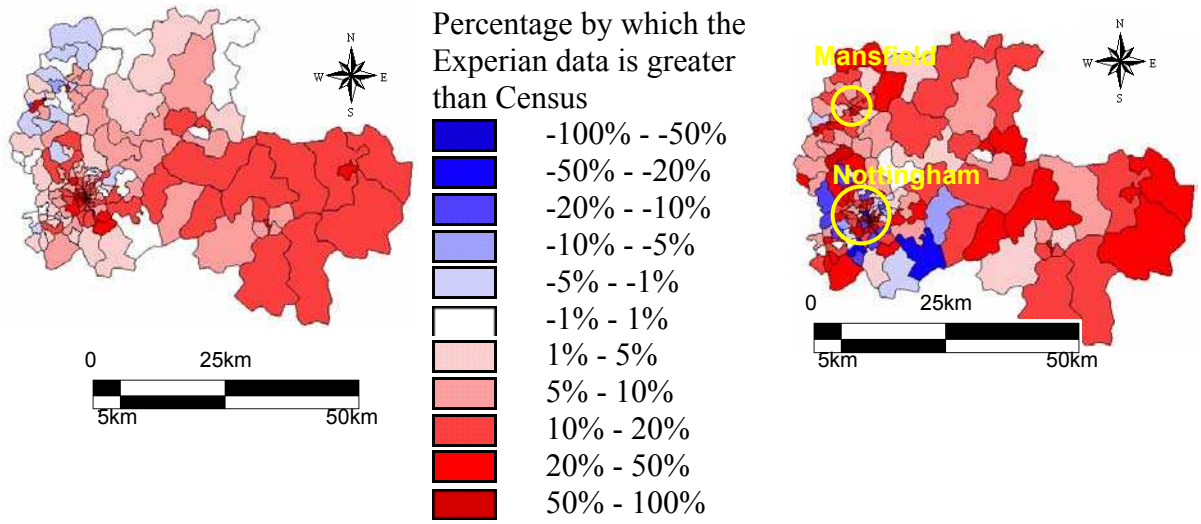


Figure 23: Postal Area NG (Nottingham) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

The pattern of underestimation by Experian in urban areas can be seen again in figure 24. This pattern can been clearly in areas which contain both urban and rural areas such as postal area AB (figure 17), EH (figure 21), NG (figure 23), NR (figure 24) and especially YO (figure 27).

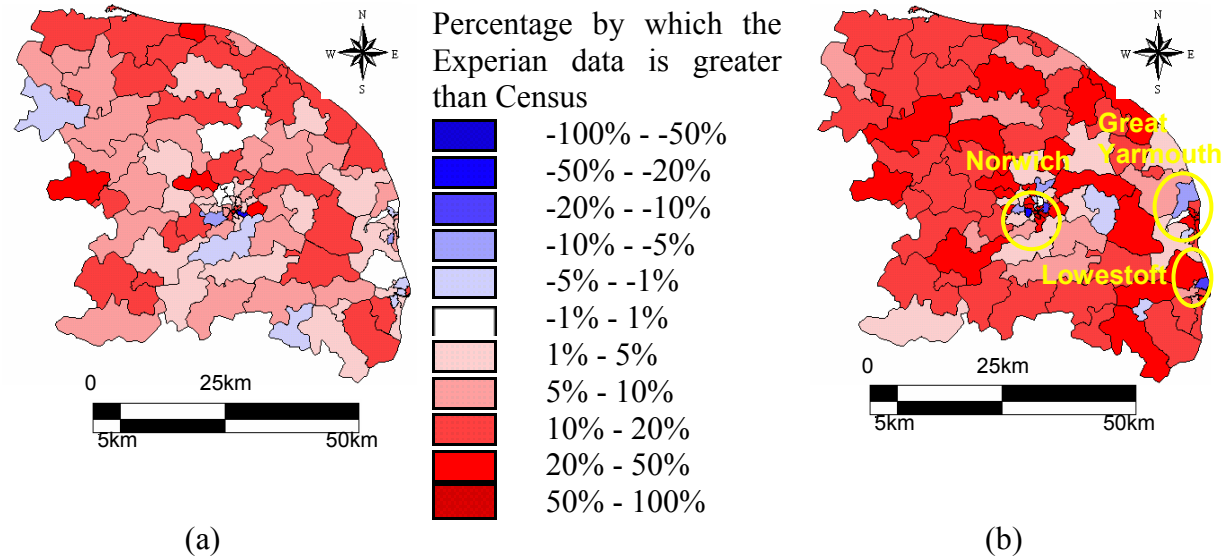


Figure 24: Postal Area NR (Norwich) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

Postal area TQ (figure 25) exhibits only very tangible evidence of any significant geographic pattern. In contrast to the pattern showing underestimation by the Experian dataset in urban areas for some postal areas. When the area is mainly rural such as CA (figure 18) or when relatively urban throughout such as E (figure 20), GU (figure 22), and WR (figure 26), the pattern showing the census value being higher than the Experian value in urban areas cannot be seen to any degree of clarity as these areas do not have a rural/urban contrast.

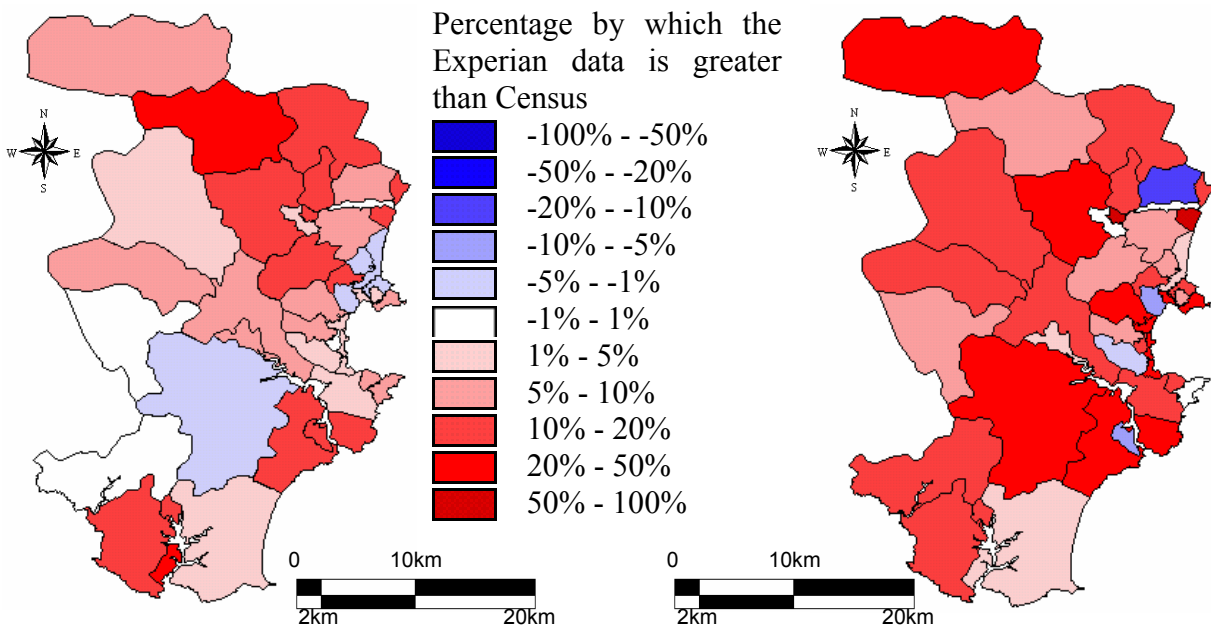
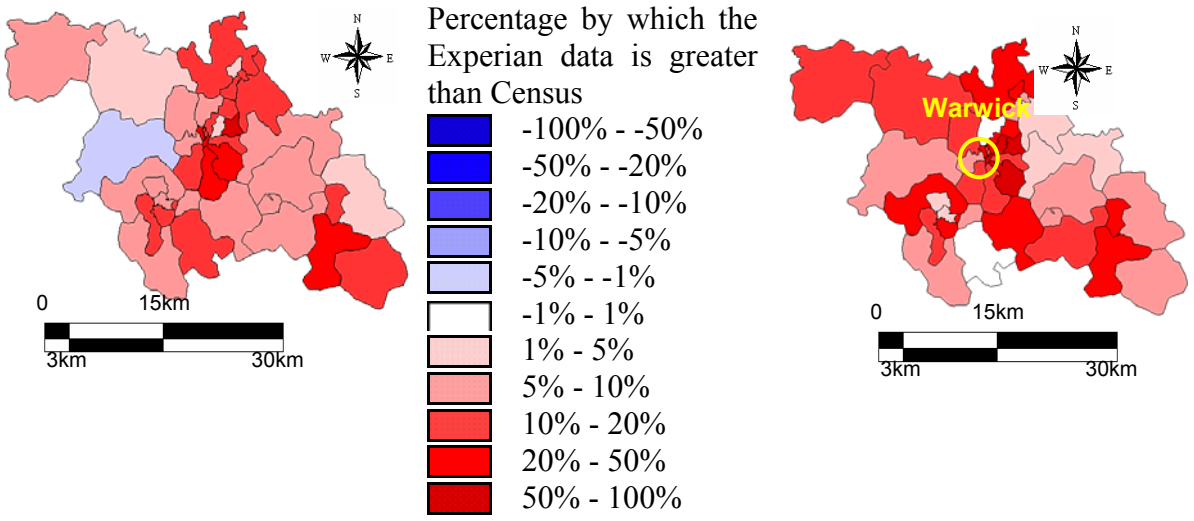


Figure 25: Postal Area TQ (Torquay) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

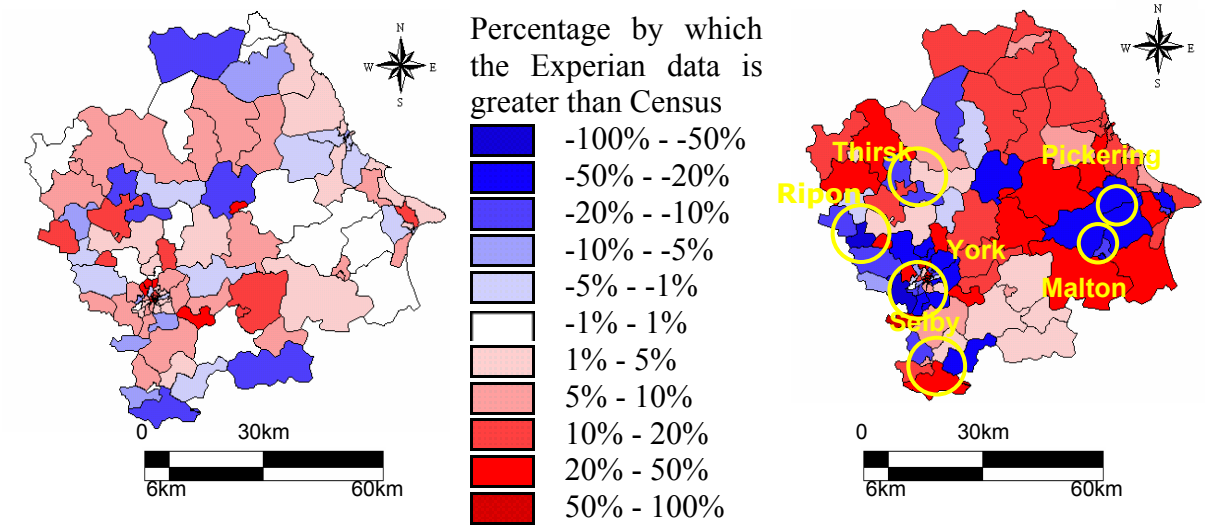
Postal area WR (figure 26) appears to contradict the trend of the census data having a higher value than the Experian data in urban areas. The sector, which so the Experian data to be greater that the census data to the greatest degree are in the main urban centre of Warwick.



(a) (b)

Figure 26: Postal Area WR (Warwick) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

Figure 27 shows postal area YO to have perhaps the largest degree of underestimation by the Experian dataset in urban areas. Many of the postal sectors that have a higher value in the census dataset than the Experian dataset can be attributed to some of the areas largest urban areas.



(a) (b)

Figure 27: Postal Area YO (York) the percentage difference in the value of postal sectors between the Experian and Census datasets using the look-up table to apply Census data to postal sectors (a) Population (b) Number of cars

The population figures seem to show the same pattern of underestimation in urban areas and over estimation in rural areas but not to such a large or obvious extent. It is not known how much of the difference between the two datasets is due to growth over time, differences within the Experian dataset, or differences in the way that the census data links to the postal sectors. Therefore the difference that is measured is the suitability of the Experian and census data to be linked together as an added-value dataset.

One thing that is clear is that any sectors that appear blue in figures 17 – 27 are particularly important, as they rule out growth over time as a reason for the difference between the two datasets. Therefore sectors coloured blue in figures 17 - 27 point to differences in the Experian data or the Experian look-up table, which links the census EDs to postal sectors. The geographic variation in difference demonstrates that the difference between the two datasets is not geographically random. It also backs up evidence from the large number of sectors observed with more cars than people as seen in (table 5) that the cars variable displays more variation from the census dataset than the population variable.

It was found that when the number of cars per person was calculated the two datasets showed significant differences. This raises the question whether these differences show any geographic patterns. Figures 28 - 38 show how the percentage difference between the number of cars per person in the two datasets varies geographically. The maps showing the number of cars per person is not affected by the Experian dataset containing more cars and people than the census, as they are rates not real numbers.

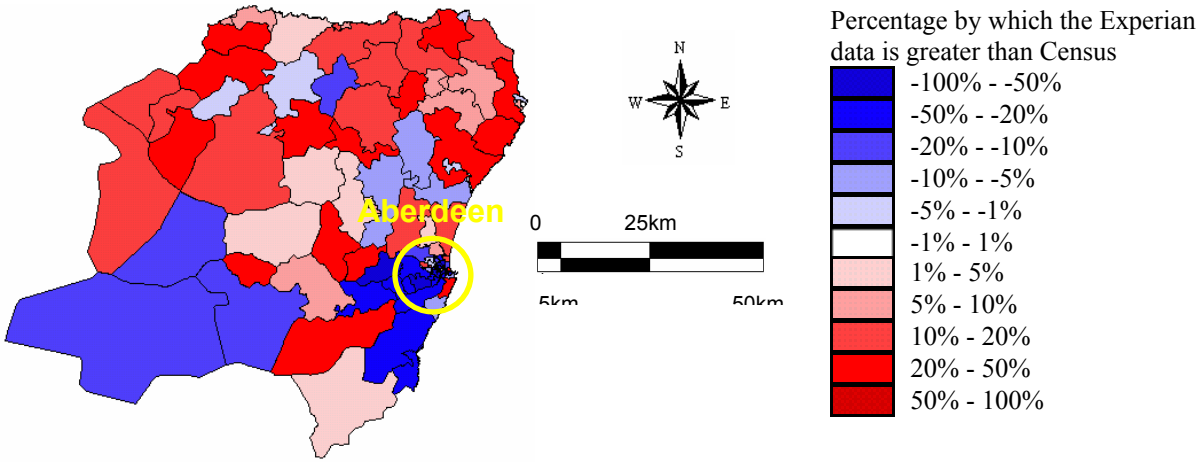


Figure 28: Postal Area AB (Aberdeen) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area AB (figure 28) shows very clearly the urban rural disparity between the two datasets, Aberdeen the main urban centre in the region displays a much higher level of car ownership in census data than the Experian data. This reflects the pattern shown in (figure 17b) showing the number of cars in the AB postal area.

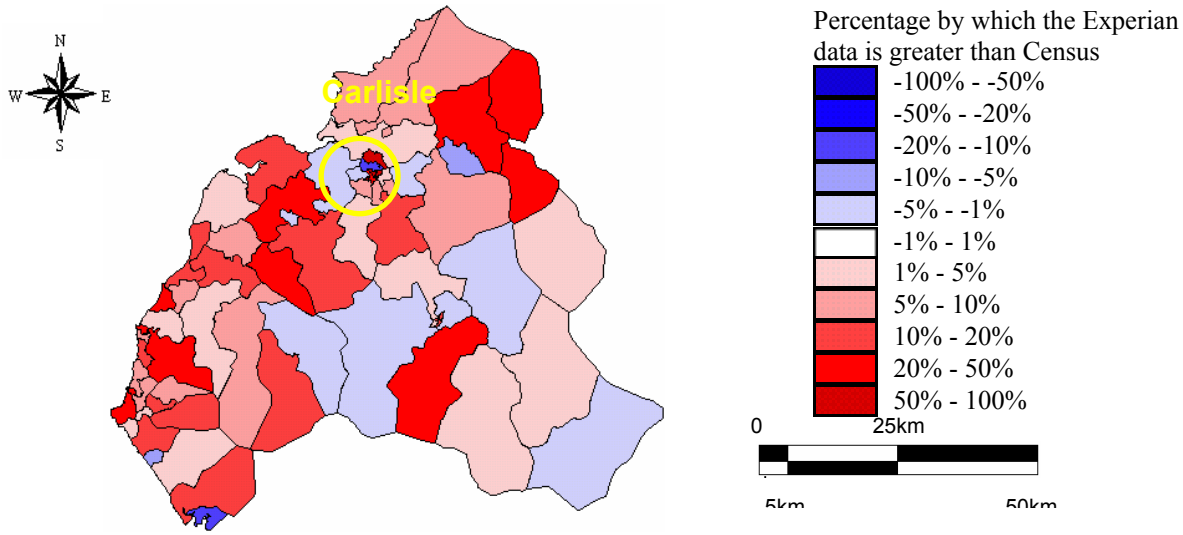


Figure 29: Postal Area CA (Carlisle) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Figure 29 displays geographic variation in car ownership between the two datasets for postal area CA. There is not the large cluster of blue coloured sectors as seen for postal area AB (figure 27), however there is one postal sector in the centre of Carlisle, for which the Census dataset has a much high rate of car owner ship than the Experian dataset.

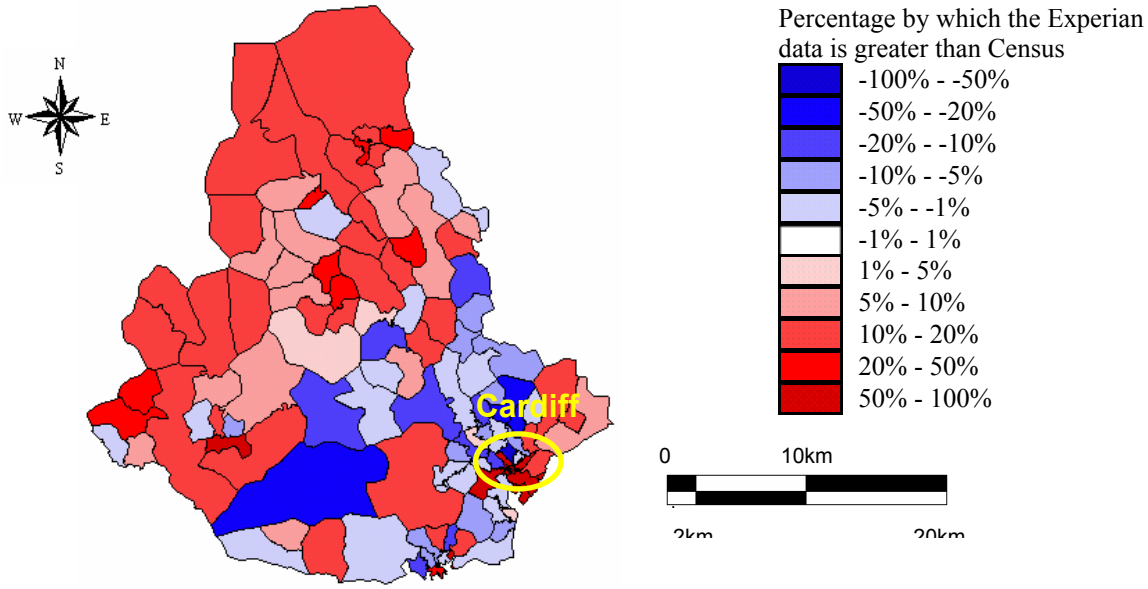


Figure 30: Postal Area CF (Cardiff) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area CF (figure 30) shows a distinct rural/urban contrast in which dataset shows the highest rate of car ownership. In the more rural parts of the postal area to the north, the Experian dataset shows a higher rate of car ownership than the census data. In contrast in the more urban areas of the postal sector towards the south, in and around the city of Cardiff the census dataset shows a higher rate of car ownership.

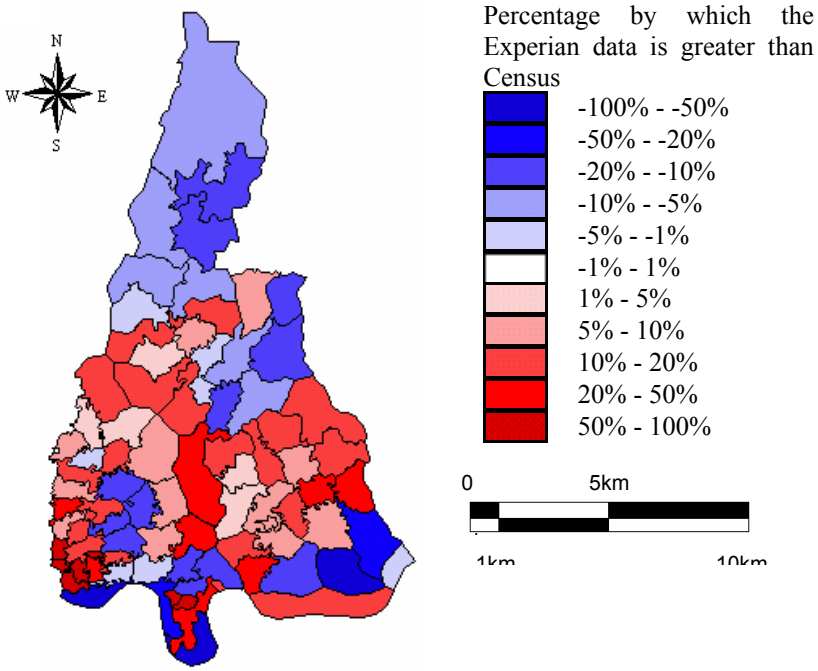


Figure 31: Postal Area E (London East) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area E (figure 31) shows distinct clusters of areas where the Census data shows a higher rate of car of car ownership than the Experian dataset. Unlike in previous examples the differences in postal area E are not evidence of a urban/rural difference between the two datasets, as postal area E is in inner London and equally urban throughout. However the fact that clustering that is shown suggests that the differences between the two datasets are not random.

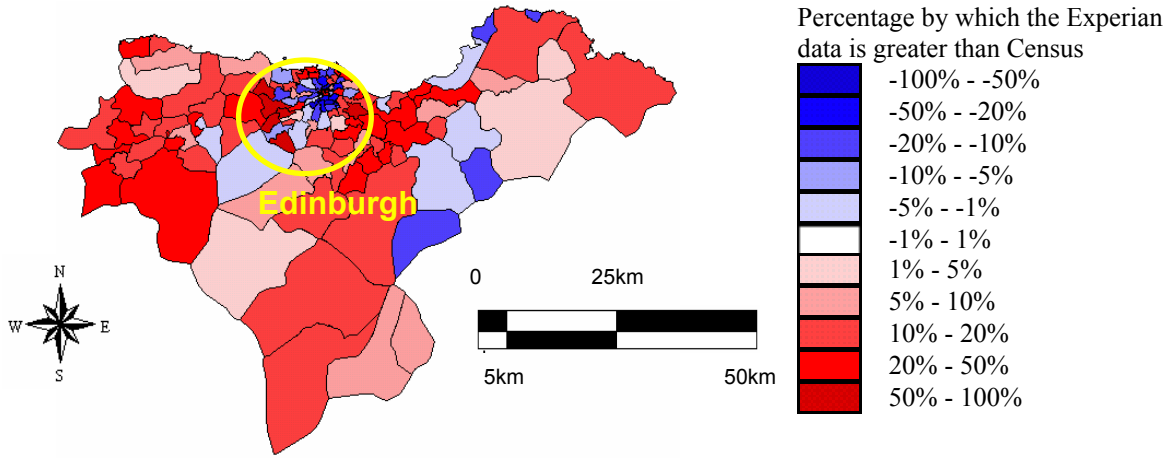


Figure 32: Postal Area EH (Edinburgh) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area EH (figure 32) displays the rural/urban differences as seen previously the census values are higher than the Experian values in the main urban area of Edinburgh whereas the Experian values are higher than the census in most other postal sectors.

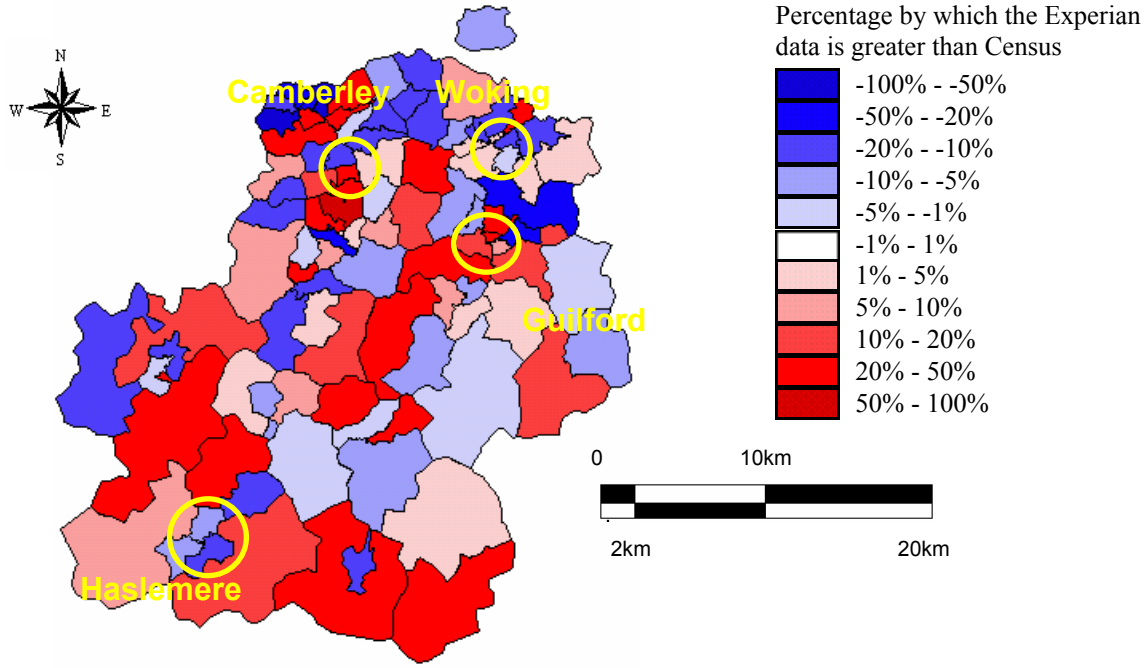


Figure 33: Postal Area GU (Guildford) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area GU (figure 33) shows some small groupings of areas where the census value is higher than the Experian value but no real clustering can be seen. This is perhaps not surprising as postal area GU covers parts of Outer London and Surrey, which mainly suburban, therefore not showing an urban rural/urban contrast. However, some of the postal sectors that have a higher Census value than Experian correspond to some sizeable settlements such as, Camberley, Woking and Haslemere, but there are many postal sectors with a higher value from census data than Experian, which do not correspond to urban areas.

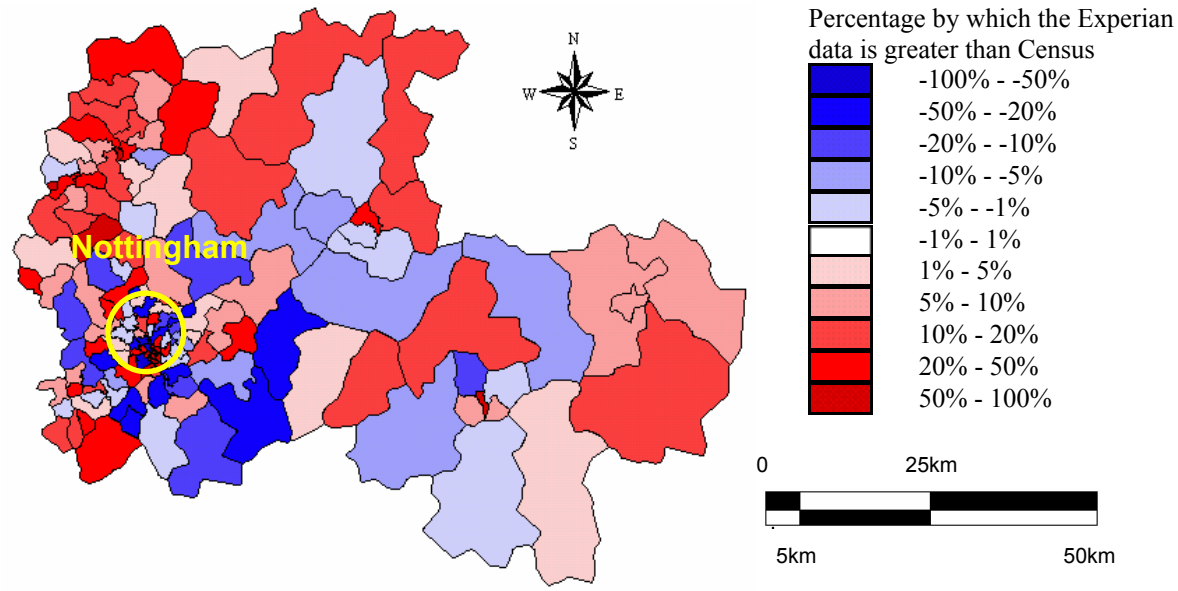


Figure 34: Postal Area NG (Nottingham) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area NG (figure 34) shows a distinct trend. In and around the main urban centre of Nottingham there are many postal sectors for which the census value is significantly higher than the Experian value. Further away from Nottingham the amount by which the census values are greater than the Experian values is greatly reduced, in most cases the Experian value is greater than that of the census.

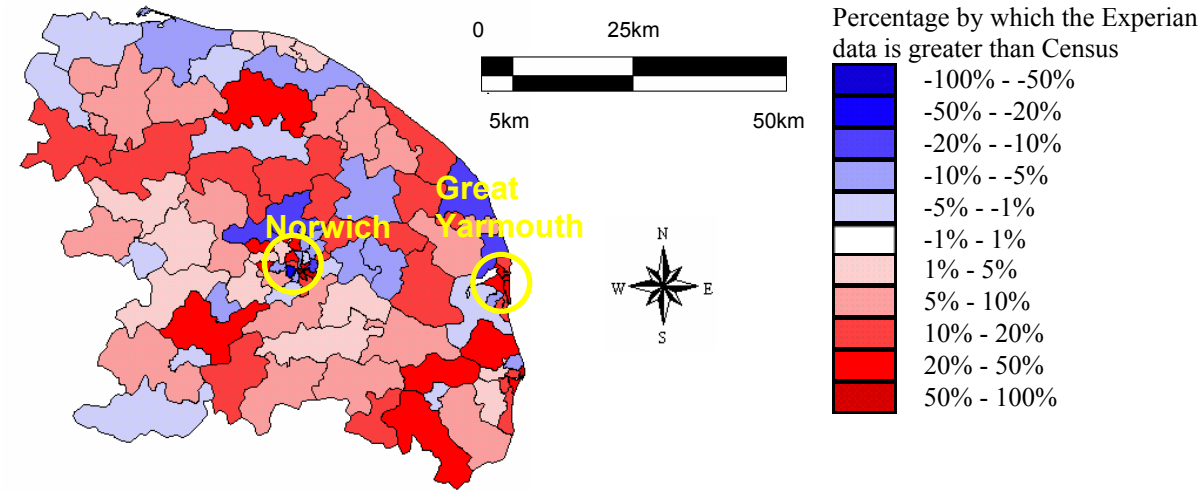


Figure 35: Postal Area NR (Norwich) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area NR shows the pattern of the census data being greater than the Experian value in urban areas this can be seen in Norwich the main urban centre of the region and the coastal town of Great Yarmouth.

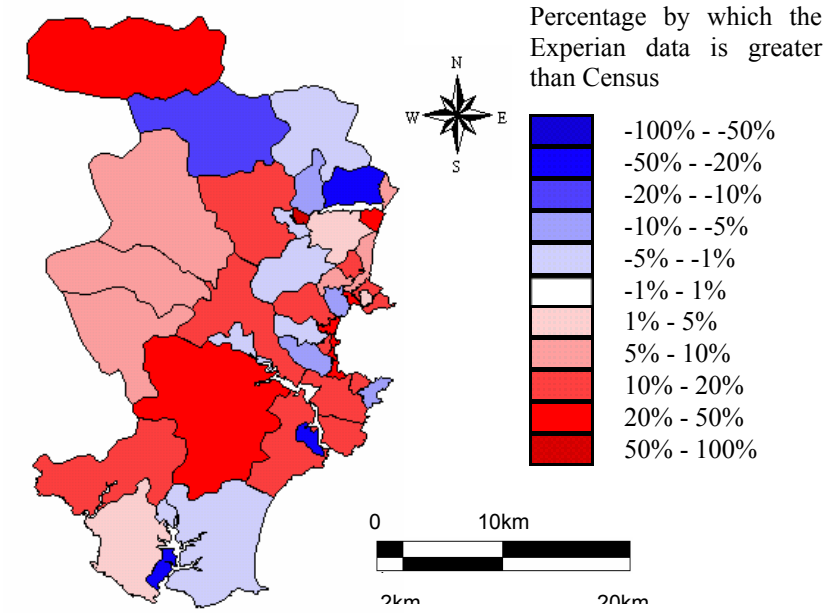


Figure 36: Postal Area TQ (Torquay) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Figure 36 illustrating postal area TQ does not show the rural urban pattern that other areas show, the sectors where the Census data is greater than the Experian data appears to be randomly distributed. The main urban centres in this area are on the east coast.

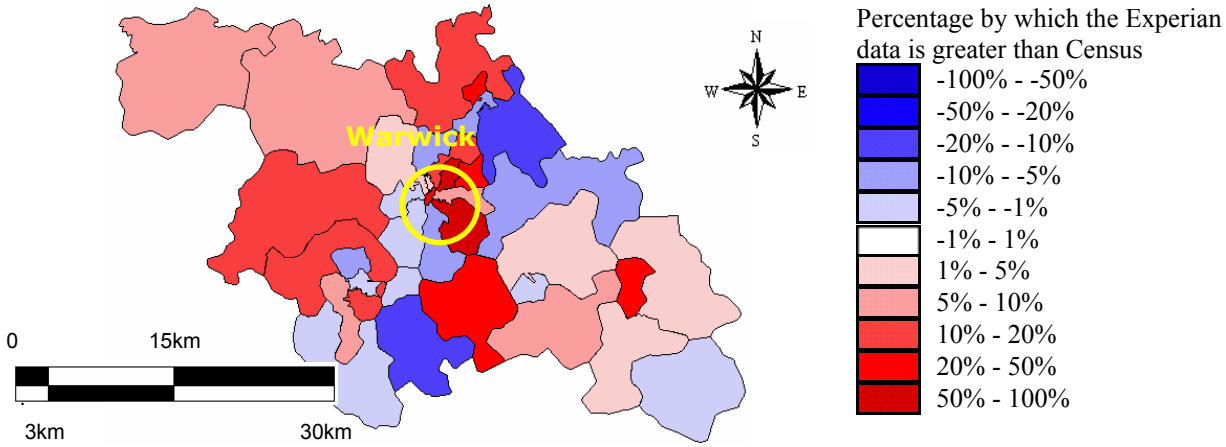


Figure 37: Postal Area WR (Warwick) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area WR (figure 37) appears random in terms of how the values of the sectors are distributed. Warwick the main urban centre does not show the trend of having higher census values, which are displayed, in urban areas in other areas.

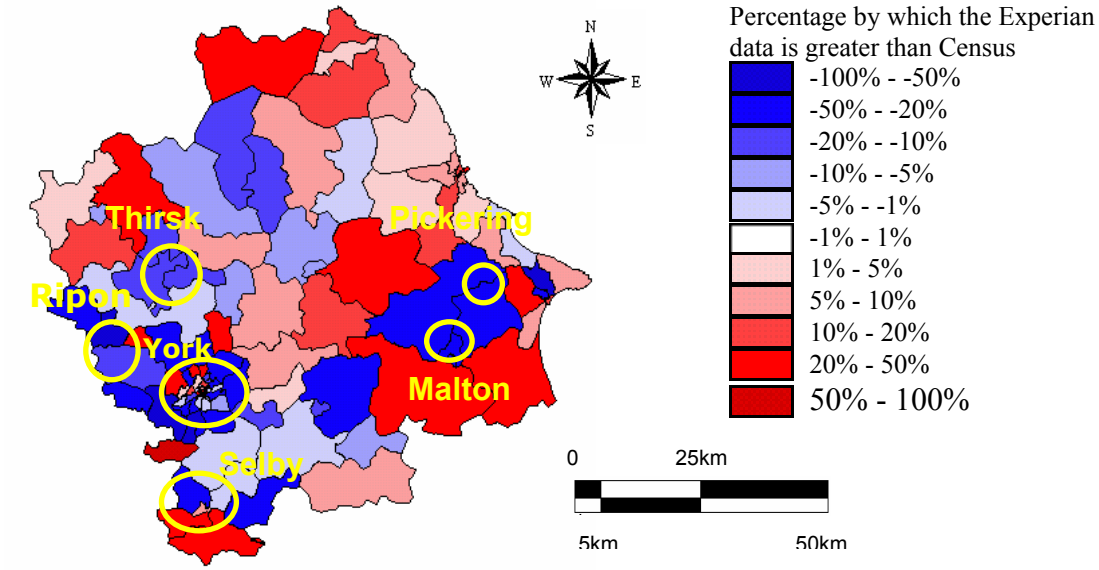


Figure 38: Postal Area YO (York) the percentage difference between the rate of car ownership in the Experian and Census datasets using the look-up table to apply Census data to postal sectors

Postal area YO (figure 38) perhaps more than any other clearly displays the pattern of the Experian data underestimating the rate of car ownership in rural areas. Clusters of the sectors where the Census value is greater than the Experian value can be seen to relate to urban areas.

As in the maps of differences in number of cars, the variation in the level of car ownership appears to show a clear geographic pattern. The census data having a higher level of car ownership in urban areas and the Experian dataset having the highest level of car ownership in more rural areas. This suggests that the differences between the two data sets are not geographically random. It has already been established that the cars variable in the Experian dataset contains differences. It has now been recognised that these differences are affected by geography. This would suggest that when the Experian car values were worked out, the process used in some way underestimated in urban areas and overestimated in more rural areas actual values. Another possibility is that because the urban postal sectors and EDs are much smaller than their rural counterparts, this could create larger differences for urban sectors when they are linked via the look-up table. This is a phenomenon that is recognised to have taken place in the OPCS postcode to ED directory (see section 3.1).

5.4 The accuracy of the Experian look up table

For some postal sectors the values in the Experian data vary greatly from those for the postal sectors when using the census data using the look-up table. But are these differences caused mainly by differences in the data set or is the look-up table unreliable and responsible for most of the differences between the values shown by the two datasets?

The look up table that is provided with the Experian postal sector data enables the users to link the postal sector data to census EDs. This enables the user to either apply the census data or other data at ED level to postal sectors, or alternatively the Experian data currently at postal sector level can be applied to census EDs. This is indeed a very valuable tool if you want to link two data sets at different spatial scales, however the accuracy of data that is linked in this way will rely heavily on the accuracy and precision of the look-up table.

There are several different ways in which the look-up table could have been made; however no information is provided as to which method has been employed. Different forms of Interpolation between areal units include:

- Percentage of individual EDs covered
- Centriods
- Population weighted centriods

The poor way in which the look-up table has been put together illustrated by 0 weightings, which have no practical use for inclusion also suggest unreliability and indecision. The EDs used in the look-up table are a very poor illustration of the EDs which the postal sectors correspond to geographically. This is evident by the fact that some EDs are included which do not correspond to the postal sectors and others EDs that the postal sectors correspond to are not included.

The county number is missing from the start of the ED number in the look-up table e.g. 37 for North Yorkshire. This could be seen as trivial, however some postal areas stretch over several counties that all have a different number, which is needed before the ED can be joined to the correct postal sector. This can become a very fiddly process and makes data analysis take

much longer. As much as anything else this is another example of the apparent lack of care in the creation of the look-up table.

The use of look-up table has been generally been discussed as poor way of linking datasets due to several inherent sources of difference (such as the assumption of uniformity) within its makeup. With difference in the look-up table being unavoidable it is important that it is produced as accurately and reliably as possible. Figures 39 – 41 show examples of how the Experian look-up table matches the geography of the postal sectors and EDs. Figure 39, (a) demonstrates that the EDs that the look-up table links to the postal sector, (b) shows all the EDs that intersect the postal sector. Figure 39 reveals that the Experian look-up table links postal sector YO8 9 to 25 census EDs, while Postal sector YO8 9 actually encompasses 36 census EDs. Also two of the EDs that the look-up table links sector YO8 9 to do not intersect that postal sector. Figure 40 exhibits that the Experian look-up table links postal sector YO10 4 to 18 Census EDs, whereas postal sector YO10 4 actually overlaps 25 census EDs. Figure 41 reveals that the Experian look-up table links postal sector NG17 6 to 2 census EDs, however postal sector NG17 6 actually overlaps 3 census EDs although one is by a small amount. Figures 39 – 41 illustrate that not only does the look-up table naturally contain differences, but it also seems to have been produced inaccurately and hence is unreliable. To test how reliable the look-up table is new weightings were produced for the three sectors in figures 39 – 41.

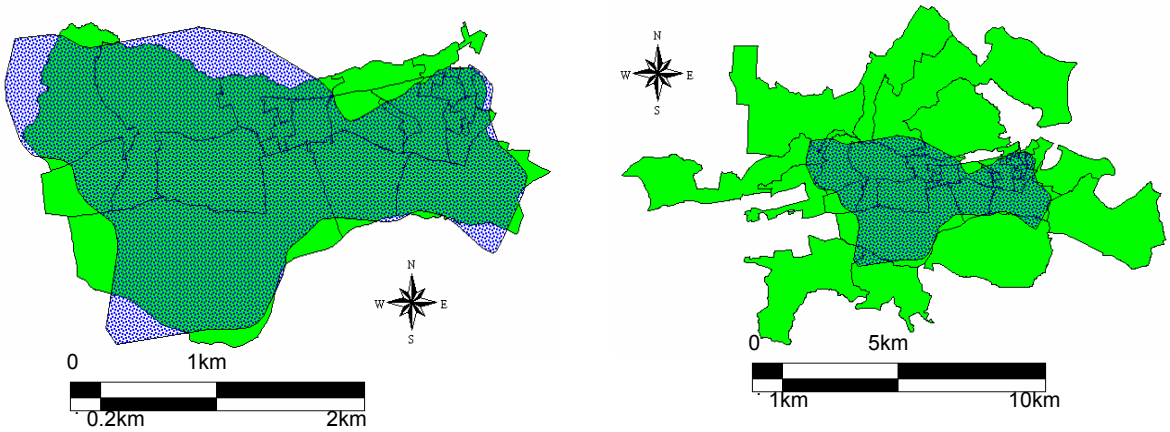


Figure 39: Postal sector YO8 9 the overlap of postal sectors and census EDs, (a) EDs that are linked to the postal sector in the Experian look-up table. (b) EDs that the postal actually intersects.

Table 7 presents the values produced for the postal sector YO8 9 using the look-up table equivalent to what is seen in figure 39 (a). Table 8 presents the values produced for the postal sector YO8 9 using the EDs the sector intersect table equivalent to what is seen in Figure 39 (b). Not only does the Experian look-up table not link each sector to the correct number of EDs. When tables 7 and 8 are compared it can be observed that the weightings given to each ED do not relate accurately to the amount by which the postal sectors and the EDs intersect. The comparison between the two different sets of weightings in tables 9 and 10 show three contrasting sets of values for the same postal sectors. Despite linking to more EDs the value for both the population and cars variables is lower for the intersection weightings than both the Experian value and the census value through the Experian look-up table. Examination of tables 7 and 8 reveal significant differences in the weightings to each ED.

Table 7: Experian Weightings look-up table weightings for postal sector YO8 9

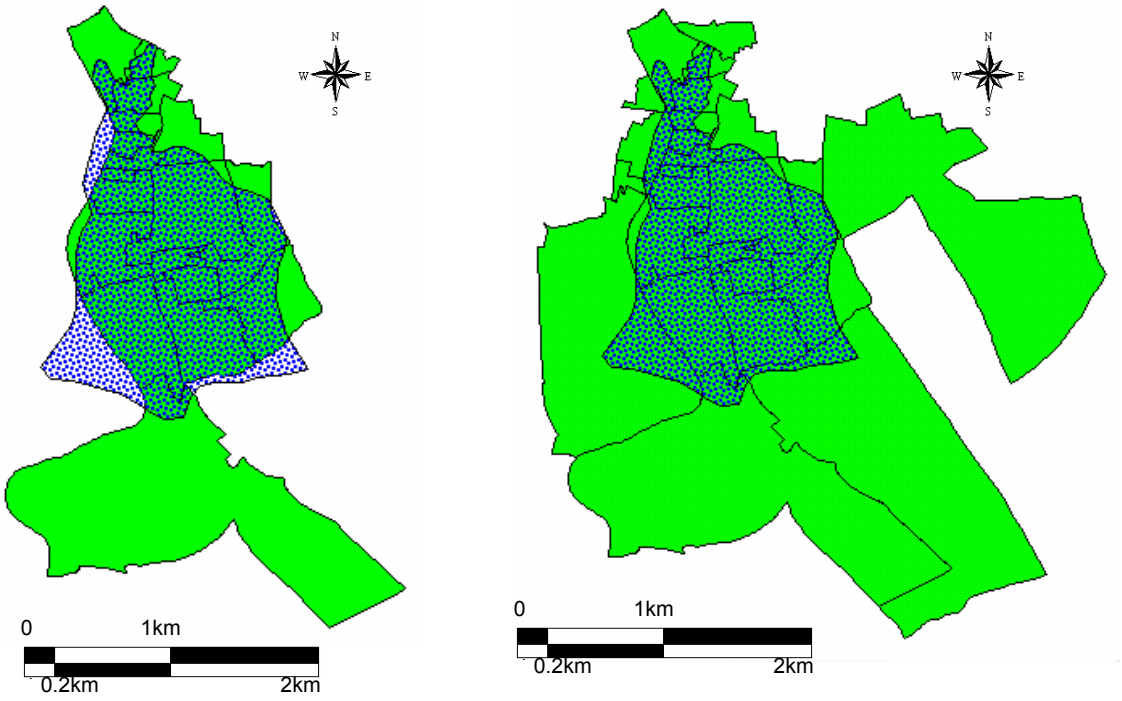
ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
37PDFF02	1	638	280	638	280
37PDFF03	1	519	227	519	227
37PDFF04	1	702	281	702	281
37PDFF05	1	354	113	354	113
37PDFF06	1	692	286	692	286
37PDFF07	1	524	218	524	218
37PDFF08	1	496	223	496	223
37PDFF09	1	603	369	603	369
37PDFT02	1	171	113	171	113
37PDFT03	1	462	229	462	229
37PDFT04	1	433	183	433	183
37PDFT05	1	515	293	515	293
37PDFT06	1	119	60	119	60
37PDGA03	0.1257	358	116	45.0006	14.5812
37PDGD01	0.0038	538	290	2.0444	1.102
37PDGD02	0.1282	536	219	68.7152	28.0758
37PDGD03	1	490	207	490	207
37PDGD04	1	515	189	515	189
37PDGD05	1	573	187	573	187
37PDGD06	1	559	219	559	219
37PDGK01	1	582	289	582	289
37PDGK02	1	610	240	610	240
37PDGK03	1	487	224	487	224
37PDGK04	1	566	301	566	301
37PDGK05	1	582	251	582	251
Total				**11307.76**	**5025.759**

Table 8: Weightings produced from the actual amount the postal sector covers each ED for postal sector YO8 9

ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
37PDFD03	0.0195	111	54	2.1645	1.053
37PDFF01	0.0012	489	241	0.5868	0.2892
37PDFF02	0.6769	638	280	431.8622	189.532
37PDFF03	0.8343	519	227	433.0017	189.3861
37PDFF04	1	702	281	702	281
37PDFF05	0.9365	354	113	331.521	105.8245
37PDFF06	1	692	286	692	286
37PDFF07	0.9616	524	218	503.8784	209.6288
37PDFF08	0.9462	496	223	469.3152	211.0026
37PDFF09	0.8913	603	369	537.4539	328.8897
37PDFF10	0.0284	450	243	12.78	6.9012
37PDFK03	0.0072	402	219	2.8944	1.5768
37PDFK05	0.0584	355	180	20.732	10.512
37PDFT02	0.8152	171	113	139.3992	92.1176
37PDFT03	0.8771	462	229	405.2202	200.8559
37PDFT04	0.9996	433	183	432.8268	182.9268
37PDFT05	0.5969	515	293	307.4035	174.8917
37PDFT06	1	119	60	119	60
37PDFT07	0.0813	153	80	12.4389	6.504
37PDFX03	0.0963	336	174	32.3568	16.7562
37PDGA05	0.0328	398	114	13.0544	3.7392
37PDGA06	0.1283	362	138	46.4446	17.7054
37PDGB01	0.0618	258	148	15.9444	9.1464
37PDGC02	0.0119	503	116	5.9857	1.3804
37PDGD02	0.1748	536	219	93.6928	38.2812
37PDGD03	0.9038	490	207	442.862	187.0866
37PDGD04	1	515	189	515	189
37PDGD05	0.9916	573	187	568.1868	185.4292
37PDGD06	1	559	219	559	219
37PDGE03	0.0147	207	109	3.0429	1.6023
37PDGF02	0.059	247	153	14.573	9.027
37PDGK01	0.9423	582	289	548.4186	272.3247
37PDGK02	1	610	240	610	240
37PDGK03	1	487	224	487	224
37PDGK04	1	566	301	566	301
37PDGK05	1	582	251	582	251
Total				**10660.04**	**4705.371**

Table 9: Differences in population using different weightings for postal sector YO8 9	Experian Population **11039**	Population table 7 **11307.76**	population table 8 **10660.04**
Experian Population **11039**		**+2.38%**	**-3.55%**
Population table 7 **11307.76**	**+2.43%**		**+6.08%**
population table 8 **10660.04**	**-3.43%**	**-5.73%**	

Table 10: Differences in the number of cars using different weightings for postal sector YO8 9	Experian cars **6309**	Cars using table 7 **5025.76**	Cars using table 8 **4705.37**
Experian cars **6309**		**+25.53%**	**+34.08%**
Cars using table 7 **5025.76**	**-20.34%**		**+6.81%**
Cars using table 8 **4705.37**	**-25.42%**	**-6.38%**	



(a) (b)

Figure 40: Postal sector YO10 4 the overlap of postal sectors and census EDs, (a) EDs that are linked to the postal sector in the Experian look-up table. (b) EDs that the postal actually intersects.

Tables 11 - 14 show the differences in the weightings for YO10 4 as shown in Figure 40. Tables 13 and 14 show once again that the value for both the population and cars variables is lower for the intersection weightings than both the Experian value and the census value through the Experian look-up table. The difference for sector YO10 4 is even greater than that shown for sector YO8 9.

Table 11: Experian Weightings as in the look-up table weightings for postal sector YO10 4

ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
37PDFS01	1	444	210	444	210
37PDFS02	1	449	185	449	185
37PDFS03	0.98	570	229	558.6	224.42
37PDFS04	0.02	457	210	9.14	4.2
37PEFF03	0.72	282	117	203.04	84.24
37PEFF04	0.29	433	118	125.57	34.22
37PEFF06	0.82	340	132	278.8	108.24
37PEFF07	1	390	145	390	145
37PEFF08	1	328	112	328	112
37PEFF09	1	374	125	374	125
37PEFF10	1	734	246	734	246
37PEFF11	1	521	187	521	187
37PEFF12	1	504	220	504	220
37PEFF13	1	437	143	437	143
37PEFF14	1	404	132	404	132
37PEFH12	0.2	204	65	40.8	13
37PEFH14	0.71	445	72	315.95	51.12
37PEFH17	0.58	433	147	251.14	85.26
Total				**6368.04**	**2309.7**

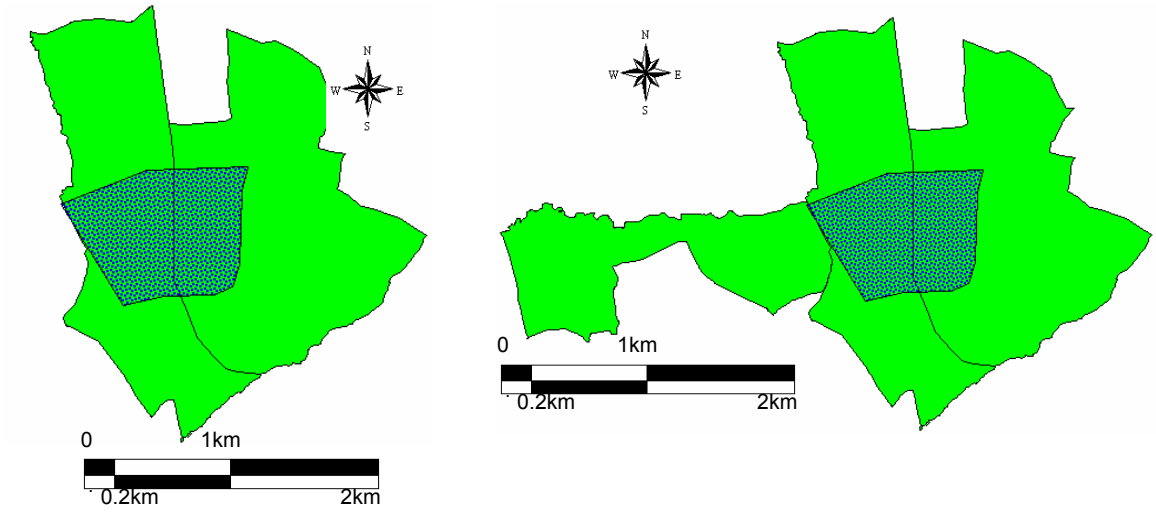
Table 12: Weightings produced from the actual amount the postal sector covers each ED for postal sector YO10 4

ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
37PDFS01	1	444	210	444	210
37PDFS02	1	449	185	449	185
37PDFS03	0.6258	570	229	356.706	143.3082
37PDFS04	0.0325	457	210	14.8525	6.825
37PDFS05	0.0423	429	163	18.1467	6.8949
37PDFW01	0.0021	337	155	0.7077	0.3255
37PEFC09	0.0087	343	124	2.9841	1.0788
37PEFF03	0.5466	282	117	154.1412	63.9522
37PEFF04	0.2008	433	118	86.9464	23.6944
37PEFF06	0.9436	340	132	320.824	124.5552
37PEFF07	1	390	145	390	145
37PEFF08	1	328	112	328	112
37PEFF09	1	374	125	374	125
37PEFF10	0.9696	734	246	711.6864	238.5216
37PEFF11	0.7509	521	187	391.2189	140.4183
37PEFF12	0.8647	504	220	435.8088	190.234
37PEFF13	1	437	143	437	143
37PEFF14	1	404	132	404	132
37PEFH12	0.2059	204	65	42.0036	13.3835
37PEFH13	0	389	78	0	0
37PEFH14	0.4552	445	72	202.564	32.7744
37PEFH17	0.6001	433	147	259.8433	88.2147
37PEFL05	0.376	435	133	163.56	50.008
37PEFL09	0.0904	359	113	32.4536	10.2152
37PEFL12	0.1536	425	180	65.28	27.648
Total				**6085.727**	**2214.052**

Table 13: Differences in population using different weightings for postal sector YO10 4

	Experian Population **7141**	Population using table 11 **6368.04**	Population using table 12 **6085.73**
Experian Population **7141**		**+12.14%**	**+17.34%**
Population using table 11 **6368.04**	**-10.82%**		**+4.64%**
Population using table 12 **6085.73**	**-14.78%**	**-4.43%**	

Table 14: Differences in the number of cars using different weightings for postal sector YO10 4	Experian cars **2523**	Cars using table 11 **2309.7**	Cars using table 12 **2214.05**
Experian cars **2523**		**+9.23%**	**+13.95%**
Cars using table 11 **2309.7**	**-8.45%**		**+4.32%**
Cars using table 12 **2214.05**	**-12.24%**	**-4.15%**	



(a) (b)

Figure 41: Postal sector NG17 6 the overlap of postal sectors and census EDs, (a) EDs that are linked to the postal sector in the Experian look-up table. (b) EDs that the postal actually intersects.

Tables 15 - 18 show the differences in the weightings for NG17 6 as shown in figure 41. Tables 17 and 18 (sector NG17 6) show the greatest difference of the three examples given, with the differences between the three ways of establishing values for the sector showing differences into the thousands of percent. The large difference here is because NG17 6 is a sector for which the original Experian data has an difference with only 4 people but 1213 cars.

Table 15: Experian Weightings as in the look-up table weightings for postal sector NG17 6

ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
38PFFN13	0.01	475	198	4.75	1.98
38PFFN18	0	372	182	0	0
Total				**4.75**	**1.98**

Table 16: Weightings produced from the actual amount the postal sector covers each ED for postal sector NG17 6

ED	*Weight*	*Population*	*Cars*	*Weighted Population*	*Weighted Cars*
38PFFN13	0.159806	475	198	75.90808	31.64168
38PFFN18	0.246398	372	182	91.65999	44.8444
18FMFX01	0.0007	692	277	0.4844	0.1939
Total				**168.0525**	**76.67999**

Table 17: Differences in population using different weightings for postal sector NG17 6

	Experian Population **4**	Population using table 15 **4.75**	Population using table 16 **168.05**
Experian Population **4**		**-15.79%**	**-97.62%**
Population using table 15 **4.75**	**+18.75%**		**-97.18%**
Population using table 16 **168.05**	**+4201%**	**+3536%**	

Table 18: Differences in the number of cars using different weightings for postal sector NG17 6

	Experian cars **1213**	Cars using table 15 **1.98**	Cars using table 16 **76.68**
Experian cars **1213**		**+61262%**	**+1581%**
Cars using table 15 **1.98**	**-99.84%**		**-97.42%**
Cars using table 16 **76.68**	**-93.68%**	**+3872%**	

Tables 7 - 18 show that the look-up table does not relate the postal sectors successfully to the EDs, which they overlie, in terms of the actual EDs, to which they are linked and the weightings used. The three examples shown in tables 7 - 18 produce an average difference of

30.6% between the weighting in the Experian look-up table and the values that the weightings should have to the EDs which the sectors overlie.

The difference between the Experian data and the census data linked to postal sectors by the Experian look-up table disclose an average difference of 9.95% per sector for the population variable and 15.89% for the cars variable. The effect of mismatching of postcodes to EDs causes blunting of socio-economic inequalities but not to the extent that the difference in geographical location suggests, this is because neighbouring EDs display similar socio-economic characteristics (Reading & Openshaw 1993). Although the effect of similar sized differences to those found in the linking of the census data to postal sectors has been shown to be relatively small in the linking of EDs to postcodes. This is due to the relatively small size of EDs, which gives similarity of neighbouring areas of that size. Differences in linking data to postal sectors will not be blunted by the similarity of neighbouring output areas due to the relatively large size of postal sectors. Therefore differences in the linking of data to postal sectors are likely to have a significant effect on the value returned for the individual postal sector.

It is clear that the look-up table linking the census EDs to postal sectors does not correspond well in reality. However, the look-up table contains two sets of weightings, the other weighting (the fourth column in the look-up table figure 3) links postal sectors to EDs. Tables 19 and 20 show that this weighting should be treated carefully as the postal sectors are significantly larger than the census EDs so linking of data in that order assumes total uniformity of data within the postal sector. This is even less likely than the uniformity within the EDs, which the first set of weightings assumes. Even with the use of weightings, linking

datasets in this way from large areas too smaller ones will cause ecological fallacy (see section 4). Tables 19 and 20 show random differences that range between +100% and -246%, any use of these weightings to join datasets will create meaningless and inaccurate results.

Table 19: Differences in 'reverse look-up' table from postal sectors to EDs, population example

ED	Sector	Weighting	ED Pop	Sector Pop	Weighted ED Pop	Difference	% Difference
37PDFE05	YO23 2	0.2	506	3856	771.2	-265.2	-52.4111
37PDFE06	YO23 2	0.15	711	3856	578.4	132.6	18.64979
37PDFF01	YO 8 8	0.08	489	5745	459.6	29.4	6.01227
37PDFF02	YO 8 9	0.0539	638	10903	587.6717	50.3283	7.888448
37PDFF03	YO 8 9	0.0722	519	10903	787.1966	-268.197	-51.6756
37PDFF04	YO 8 9	0.0727	702	10903	792.6481	-90.6481	-12.9128
37PDFF05	YO 8 9	0.0191	354	10903	208.2473	145.7527	41.17308
37PDFF06	YO 8 9	0.0546	692	10903	595.3038	96.6962	13.97344
37PDFF07	YO 8 9	0.0333	524	10903	363.0699	160.9301	30.71185
37PDFF08	YO 8 9	0.0541	496	10903	589.8523	-93.8523	-18.9218
37PDFF09	YO 8 8	0	603	5745	0	603	100
37PDFF09	YO 8 9	0.0512	603	10903	558.2336	44.7664	7.423947
37PDFF10	YO 8 8	0.07	450	5745	402.15	47.85	10.63333
37PDFH01	YO 8 8	0.03	234	5745	172.35	61.65	26.34615

Table 20: Differences in 'reverse look-up' table from postal sectors to EDs, Cars example

ED	Sector	Weighting	ED Cars	Sector Cars	Weighted ED Cars	Difference	% Difference
37PDFE05	YO23 2	0.2	229	1208	241.6	-542.2	-236.769
37PDFE06	YO23 2	0.15	340	1208	181.2	-238.4	-70.1176
37PDFF01	YO 8 8	0.08	241	3903	312.24	-218.6	-90.7054
37PDFF02	YO 8 9	0.0539	280	6309	340.0551	-307.672	-109.883
37PDFF03	YO 8 9	0.0722	227	6309	455.5098	-560.197	-246.783
37PDFF04	YO 8 9	0.0727	281	6309	458.6643	-511.648	-182.081
37PDFF05	YO 8 9	0.0191	113	6309	120.5019	-95.2473	-84.2896
37PDFF06	YO 8 9	0.0546	286	6309	344.4714	-309.304	-108.148
37PDFF07	YO 8 9	0.0333	218	6309	210.0897	-145.07	-66.5458
37PDFF08	YO 8 9	0.0541	223	6309	341.3169	-366.852	-164.508
37PDFF09	YO 8 8	0	369	3903	0	369	100
37PDFF09	YO 8 9	0.0512	369	6309	323.0208	-189.234	-51.2828
37PDFF10	YO 8 8	0.07	243	3903	273.21	-159.15	-65.4938
37PDFH01	YO 8 8	0.03	119	3903	117.09	-53.35	-44.8319

The weightings in the Experian look-up table have been found not to equate completely with the EDs below each postal sector, there are several possible reasons for this. One possibility is digitising difference; it is sometimes easy to forget when comparing the figures in each dataset that they represent geographical areas. All digitised boundaries contain random differences due to factors such as the quality of map used for digitising, the person who digitised the boundaries and how much the boundaries have been smoothed to make them look more aesthetically pleasing. A digitising difference of a few millimetres could represent several hundred metres in reality. If the boundaries from which the look-up table was created, or from which they have been checked were digitised poorly or rounded significantly this could explain why the look-up table provides a poor link between the two datasets (Duke-Williams & Rees 1998). This seems like a very plausible reason for such a large inaccuracy in the look-up table however one feature of how the postal sectors link to the EDs is that every sector tested seems to link to less EDs than it intersects. If digitising difference was the main cause of inaccuracy in the construction of the look-up table some sectors would be linked to many EDs and some to not enough, as an ED not assigned to the correct sector would still be assigned to another wrongly.

Another probable cause of inaccuracy in the look-up table is that a large amount of rounding was used in its creation. The reason for the inclusion of 1938 EDs that have a 0 weighting to postal sectors may be because they have been rounded down either intentionally or accidentally by software used in the creation of the look-up table. However on close examination of the look-up table there does not seem to be a great deal of rounding the weightings are specified to five decimal places for the ED to postal sector look-up. The postal sector to ED 'reverse look-up' table is less accurate (to four decimal places). Upon closer

scrutiny there does not seem to be any obvious reason for the look-up table to link as poorly as it does. The look-up table does not appear to suffer from digitising difference or rounding to any great extent, which brings about the conclusion that the table was unreliably constructed resulting in significant omissions from the table.

One source of unavoidable difference within the look-up table is due to some EDs having 0 values in all data fields; this is because of reasons of statistical disclosure. Census data cannot be published at a level where information about individuals could be derived. When this problem occurs the data for any ED that falls below the disclosure level, is added to a neighbouring ED, so it is still included in the dataset (Martin 2002). Although this only occurs for relatively few EDs (3694 or 3.46% for England in 1991) the linking of data using the look-up table does not take into account this feature in one of the datasets.

4,840 special EDs in the 1991 census such as prisons and communal establishments have no geographical area (Martin 2002). Therefore special EDs cannot be linked by the look-up table and will be missing if data is linked in this way.

It can only concluded that the look-up table is inaccurate both inherently and because it has been poorly constructed. If the look-up table were used to bring census data into the dataset this would add further difference to the dataset, however there is no way of knowing how much data if any was brought into the dataset in this way.

6 CONCLUSIONS

There appears to be a visible geographic difference between the values of the car variable in the census and Experian data. The Experian data tends to be much higher than the census data in rural areas, whereas in urban areas this difference is not so high or in many cases the census value is higher than that in the Experian data. This suggests that differences in the Experian data set are endemic to there source data rather than them being differences during input.

Look-up tables are a poor way of linking two different of aggregations of spatial at best they assume uniformity within each areal unit. If produced poorly or imprecisely as in the case of the Experian look-up table, they can link to the wrong information and give a poorly constructed and unreliable answer.

There are three main reasons for observed differences between the Experian Postal sector data and the census data linked to postal sectors via the look-up table:

1. Change/growth over time. The time difference between the creation of the two datasets enables significant growth in population to have taken place, therefore providing a likely source of difference between the two datasets.

2. Differences with the Experian dataset. There is little doubt that the Experian dataset contains substantial differences, such as 106,000 cars for 5,000 people, this raises concern over the reliability of these variables, and therefore due to the lack of transparency in its construction and documentation the reliability of the whole dataset is called into question.

3. Differences in the look-up table that links the census data to postal sectors. Linking of Experian and census data through the Experian look-up table exhibits significant variations between the two datasets, despite factors such as population growth concerning being taken into account. The look-up table is poorly constructed in both the EDs that it links to and the weightings they are given, as illustrated by figures 39 - 41.

However, it is very difficult to assess the amount of difference each point listed above is responsible for.

REFERENCES

1. Bryman, A. & Cramer, D. B. (1994), ***Quantitative Data Analysis for Social Scientists revised edition***, Routledge, London.

2. Collins, S. E. Haining, R. P. Bowns, I. R. Crofts, D. J. Williams, T. S. Rigby, A. S. & Hall, D. M. B. (1998), ***Errors in postcode to enumeration district mapping and their effect on small area analyses of health data***, in Journal of Public Health Medicine Vol. 20 No. 3 pp. 325-330, Oxford University Press.

3. Dale, A. & Marsh, C. (1993), ***The 1991 Census User's Guide***, HMSO, London.

4. Department of the Environment, (1987) ***Handling geographic information: The report of the committee of Enquiry chaired by Lord Chorley***, HMSO, London.

5. Duke-Williams, O & Rees, P. (1998), ***Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure***, in The International Journal of Geographical Information Science, Vol. 12 No. 6 pp 579-605.

6. Haynes, R. M. Lovett, A. A. Bentham, G. Brainard, & Gale, S. H. (1995), ***Comparison of ward population estimates from FHSA patient registers with the 1991 Census***, in Environment and Planning A, Vol. 27, pp 1849-1858, Pion.

7. Holt, D. Steel, D. G. Tranmer, M. & Wrigley, N. ***Aggregation and Ecological Effects in Geographically Based Data***, in Geographical Analysis Vol. 28 No. 3 pp 244 - 261

8. Martin, D. (1992), ***Postcodes and the 1991 Census of Population issues, problems and prospects***, in Transactions of the Institute of British Geographers Vol. 17 350 -7

9. Martin, D. (1997) ***From enumeration districts to output areas: experiments in the automated creation of a census output geography***, working paper No. 38, Statistical

Commission and Economic Commission for Europe Conference of European Statisticians.

10. Martin, D. (1998) ***Optimising census geography: the separation of collection and output geographies***, in International Journal of Geographical Information Science, Vol. 12, No. 7, pp 673 - 685, Taylor and Francis.

11. Martin, D. (2000), ***Towards the geographies of the 2001 UK Census of population***, in the Transactions of the Institute of British Geographers Vol. 25 pp 321 - 332.

12. Martin, D. (2002), ***Geography for the 2001 Census in England and Wales***, in Population Trends 108 pp 7 - 15, Office for National Statistics.

13. Martin, D. Nolan, A & Tranmer, M. (2001), ***The application of zone-design methodology in the 2001 UK Census***, in Environment and Planning A, Vol. 33, pp 1949-1962, Pion.

14. Monmonier, M. S. (1996), ***How to lie with maps 2nd Ed***, University of Chicago Press, Chicago.

15. Openshaw, S. (1984a), ***The Modifiable Areal Unit Problem***, CATMOG 38, Geobooks, Norwich.

16. Openshaw, S. (1984b), ***Ecological fallacies and the analysis of areal census data***, in Environment and Planning A, Vol. 16, pp 17-31, Pion.

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

18. Openshaw, S. & Taylor, P. J. (1981), ***The modifiable areal unit problem***, in Wrigley, N. Quantitative geography: A British View pp 60 - 70, Routledge, London.

19. Raper, J. Rhind, D. & Shepherd, J. (1992) ***Postcodes: The New Geography***, Longman, Harlow.

20. Reading, R. & Openshaw, S. (1993), ***Do inaccuracies in small area deprivation analyses matter?*** In Journal of Epidemiology and Community health Vol. 47 pp. 238-241

21. Rees, P. (1996), ***Access to population census data for research purposes in the uk*****,** Paper presented at the Annual Conference of the Australian Population Association, held at the University of Adelaide, South Australia, 5 December 1996

22. Rees, P Martin, D and Williamson, P. (2002) ***The Census Data System***, Wiley, Chichester.

23. Tranmer, M. and Steel, D. G. (1998) ***Using census data to investigate the causes of the ecological fallacy***, in Environment and Planning A, Vol. 30, pp 817-831, Pion.

24. Wrigley, N. (1995), ***Revisiting the Modifiable Areal Unit Problem and the Ecological Fallacy***, in Cliff, A. Gould, P. Hoare, A. and Trift, N. (eds.) Diffusing Geography: Essays for Peter Haggett, Blackwell Oxford.