Working Paper No. 15
ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE-EUROSTAT Work Session on Registers and Administrative Records for Social and Demographic Statistics
(Geneva, 9-11 December 2002)

Supporting paper – SESSION 1

ANALYSIS OF ADMINISTRATIVE DATA SETS FOR POSSIBLE USE IN SMALL AREA POPULATION ESTIMATES

Submitted by Cal Ghee, Jonathan Chappell, Denise Williams & Andy Bates
Office for National Statistics, United Kingdom

ABSTRACT

The Office for National Statistics in the United Kingdom (ONS) has set up a research project to investigate the production of postcensal small area population estimates for England and Wales.

As the United Kingdom does not have population registers, administrative data sets are being considered for their usefulness in producing small area population estimates.

Results from the United Kingdom 2001 Census will provide ONS with the ideal opportunity to evaluate possible data sources and methods which might be suitable for producing small area population estimates. A number of different administrative data sets have already been acquired prior to receipt of 2001 Census data, and compared. This paper considers the initial analyses of these administrative data sets at different geographic levels.

I. INTRODUCTION

1. An initial analysis of the following data sets has been carried out:

- Patient Register (PR) data for mid-2000, Census day in April 2001, mid-2001, by single year of age and sex, and by postcode;
- Department for Work and Pensions (DWP) data on children in receipt of child benefit (CB) for Census day in April 2001, by broad age groups (0-4, 5-10, 11-15, 16+) and sex, on December 1998 ward1 boundaries;
- DWP data on persons aged 65+ in receipt of benefits for mid-2001 – "Super Older Persons Database" (SOPD), by single year of age and sex, on December 1998 ward boundaries;
- Electoral Registers (ER) data on persons aged 18+ who register to vote, for October 1998, October 1999 and October 2000, on December 2000 ward boundaries.
- Foreign armed forces data, both military and their dependants, for mid-2000 on December 1998 boundaries

II. LIMITATIONS ON ANALYSIS POSSIBLE AND DATA QUALITY ISSUES

2. Data quality issues include aspects of the administrative systems the data sets are part of. These include the practice of "deadwooding" registers infrequently, or on an area by area basis; the time lag involved in registering births or immigrants; re-registering of internal migrants; removal of emigrants and dead people. These affect different data sources in different ways and cannot be fully evaluated until the small area census data are published.

II.1 Patient Register data

3. Data evaluation reports accompanying the PR data record that, for each year, around 0.4 per cent of postcodes are missing from the raw data; 0.1 per cent of postcodes are invalid, and 0 per cent of records contain range errors. The quality of this data set has improved considerably in recent years.

II.2 Geographical coding in Patient Register data

4. In the mid-2000 data provided by the data supplier, 959 postcodes (relating to 6,184 persons) did not have a ward code allocated to them. This relates to only 0.07 per cent of the total number of postcodes, or 0.01 per cent of the total population. On a national, or Local Authority District (LAD) level this is probably negligible. However, this may have a substantial impact at small area level if the postcodes were clustered in particular areas. The original coding in the April 2001 extract is much improved: just 507 postcodes had no ward code, relating to only 1,720 persons.

5. In order to compare the PR data against the other data sources received on particular boundaries, it was necessary for the Project team to re-allocate the postcoded data supplied onto 1998 and 2000 boundaries. This was done using postcode to ward geographic lookup files. The recoded data for mid-2000 now only has 1,834 persons not allocated to a ward (just over 0.003 per cent of the total number of patients), and in the April 2001 data only 1,460 persons (just under 0.003 per cent of total patients).

6. It has also been found that some postcodes may cross ward boundaries. It is estimated that on average each ward may have 2 per cent of postcodes that overlap two or more wards. Until data are available on individual addresses, and more accurate lookup files used, the extent of potential allocations to the wrong wards from averaged postcode locations will not be known.

II.3 DWP Child Benefit (CB) data and Super Older Persons Database (SOPD)

7. Under the current agreement with DWP, the majority of the checking of their data has been done externally, and a report on the quality of the raw data has been requested. Data are available on the people who could not be assigned to an area in the SOPD - approximately 0.5 per cent of the total for England and Wales.

8. DWP have provided child benefit and older persons data at ward level by age group with some cells suppressed for "disclosure control" purposes. The basis of this suppression was any cell containing fewer than ten individuals for the CB data, and fewer than 5 in the SOPD. The total number of cells suppressed in the CB data is small (0.07 per cent), partly because the data have been provided in age groups. The number of cells suppressed in the SOPD, however, is much larger (18 per cent) due to the potentially small numbers of elderly people at ward level.

II.4 Electoral Register

9. The ER data are compiled into electronic form from returns provided by LADs which include ward aggregates. We have found several limitations with this data. For instance, some LADs made returns for ward boundaries in October that would come into effect the following December, while others only included those that were effective the previous December. In order to look at the time series of data over the three years, data returned on out of date or future boundaries are just compared at LAD level.

10. A further limitation of the electoral data is that it contains no age or sex details. The level of participation by the eligible population is known not to be 100 per cent, and there are issues about whether tendency to register varies across time and between areas.

11. The electoral register is compiled as persons living at the residence in question as at 15 October each year, and these registers are then valid for voting during the next calendar year. In order to estimate a "mid-year" electorate, it was necessary to interpolate between two registers.

12. "Attainers" are included in the registers: anyone aged 16 or 17 on 15 October. The numbers of attainers are included in this data and may affect the comparisons. Similarly, the electronic data received do not include European citizens who are eligible to vote in European Parliamentary elections but not UK Parliamentary elections, so comparisons may be affected.

II.5 Foreign Armed Forces and their Dependants

13. The foreign armed forces data at ward level can only be used as an indicator of the possible number of foreign armed forces personnel and dependants. The data for the US Navy and Air Force were only given by place name or postcode sector. The data included in the analysis below should therefore be considered as an initial estimate of the location of foreign armed forces to aid the comparisons between data sets where they are not included in the records.

II.6 Summary

14. The validation has identified considerable difficulties with the postcoded data and differences in the methods used to assign the postcoded data to wards. Some have been missing or incorrect or they straddle ward boundaries. This has significant impact at small area level as large numbers of people could be "missed" or assigned to the wrong ward. However the situation should improve with the improvements to geographic lookup files being implemented by ONS.

15. The following analysis comprises an initial look at the data as received. It is limited for the reasons given above but gives a useful indication of potential data errors or inconsistencies, both between sources and between years. Data will be more comparable once they are subject to the same geographical treatment, and more meaningful comparisons between data sources can be made when the small area census data are available.

III. INITIAL ANALYSIS OF DATA SETS

16. The following basic analysis of the available data sets can be carried out:

- mid-2000 PR data for totals aged 18+ against ER data interpolated to mid-2000, at national, LAD and ward level, on 2000 wards (including Home and Foreign Armed Forces, foreign armed forces dependants, and prisoners in the LAD comparison);
- April 2001 PR data against April 2001 CB data for 0-15 year olds, at national, LAD and ward level, on 1998 wards;
- mid-2001 PR data against mid-2001 SOPD for ages 65+, at national, LAD and ward level, on 1998 wards;
- comparison of ER growth between 1999-2000 and 2000-2001;
- investigation of the growth in PR data in the year mid-2000 to mid-2001, at national, LAD and ward level;
- comparison between number of 0 year olds in mid-2000 PR with 1 year olds in mid-2001 PR, in order to try and identify a presence of a time lag in registering newborn babies on doctors' registers.

III.1 Comparisons at National level

17. The figures in Table 1 show that at a national (England and Wales) level, the number of people on patient registers exceed those on the other data sets. This may be caused by list inflation, but the true extent of this cannot be identified until the data sets are compared against the census. The largest differences exist between the patient registers and the electorate; and the patient registers and the SOPD for males aged over 65. The difference against child benefit was smallest at under 4 per cent. The difference between the PR and ER data would be even larger (at around 8.3 per cent) if home armed forces were added into the PR data, and Europeans added into and attainers subtracted from the ER data.

Table 1: Comparisons between data sets at national level

Difference between:		**difference as % of PR**
PR and ER	mid-2000, totals aged 18+	+7.8%
PR and SOPD	mid-2001, males 65+	+10.0%
	mid-2001, females 65+	+4.7%
PR and CB	census-2001, males 0-15	+3.8%
	census-2001, females 0-15	+4.0%

III.2 Comparison between Patient Register and Electoral Register at LAD and Ward level

18. It was found that patient register counts were much larger than the electorate in many Inner London areas where there are likely to be large numbers on non-UK nationals resident who are not eligible to vote. It was also noted that in some areas with high numbers of students, patient register counts were higher than the electorate. This could be caused by the patient registers being inflated in these areas due to students not re-registering when moving, compounded by students not bothering registering to vote. As is expected, the electorate was higher than patient register counts mainly in areas where there are expected to be a high number of home armed forces.

Table 2 Percentage difference between PR and ER data at LAD and Ward level

	LAD	Ward
Maximum	41.7%	65.0%
75th percentile	7.3%	7.0%
Median	4.2%	3.3%
25th percentile	1.9%	0.6%
Minimum	-3.0%	-101.1%

19. At national level the difference between the two data sets was around 8 per cent. At local authority level the median percentage difference is over 4 per cent, however at ward level the median percentage difference is just over 3 per cent. This, and the relatively small interquartile range of the LAD and ward comparisons, indicates that a large overall difference between the numbers of patients and electors at national level is distributed as small differences in many areas, and a few larger differences. The extent to which some of these differences can be explained by the different coverage of the data will be examined in the next stage of the analysis.

20. It was not possible to examine sex ratios for the electoral roll data because the data is only given as a total population for each ward.

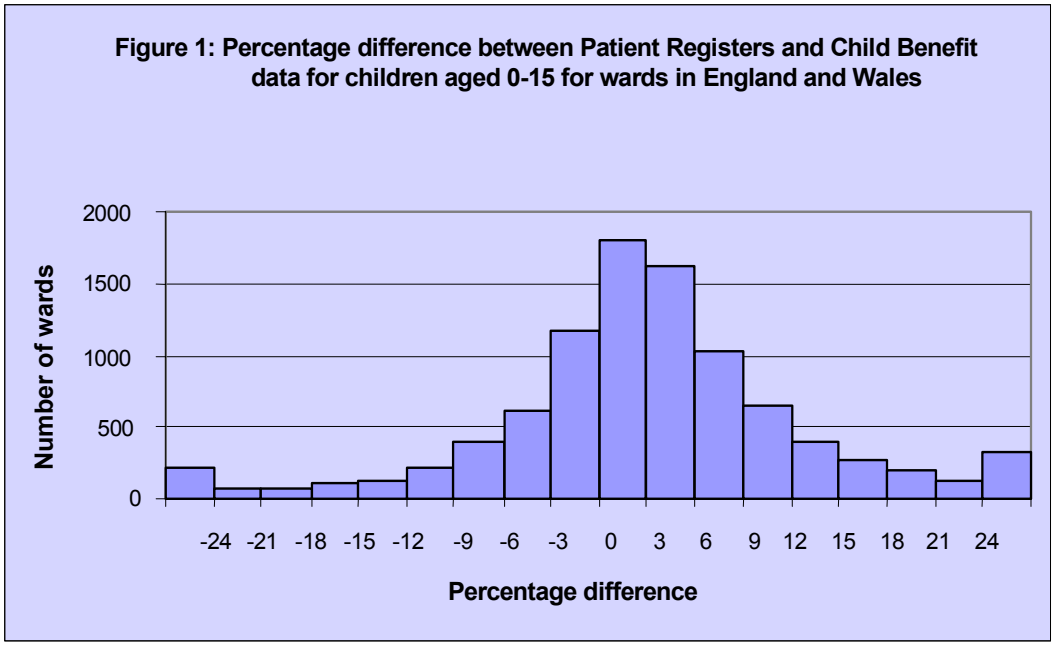
III.3 Comparison of Patient Register and Child Benefit data

21. There were 3.7 per cent more persons registered on the patient registers aged 15 and under than in the child benefit data for England and Wales. This percentage is similar for males and females. The number of children on the patient registers is slightly higher compared with child benefit data for older age groups with a 2.7 per cent difference for the 0-4 year olds compared with 4.5 per cent for the 11-15 year olds. The number of boys to every 100 girls is similar in both data sets at about 105, which is what would be expected for this age group.

Table 3 Percentage difference between patient registers and DWP child benefit for 0-15 year-olds for local authority districts and wards

	LAD	Ward
Maximum	22.9%	93.4%
75th percentile	4.4%	7.4%
Median	2.8%	2.8%
25th percentile	1.7%	-1.2%
Minimum	-2.8%	-325.0%

22. The differences between these data sets at ward and local authority district level are not very different for males compared with females, or for the age groups 0-4, 5-10 and 11-15. Generally, data compare very well between these two sources as can be seen from Figure 1. However there are a few extreme outliers at ward level to be investigated.



23. Although the two data sources are very close in most areas there are quite a few wards with very large differences, both positive and negative. In some wards where patient register counts are much higher than child benefit there are boarding schools. Interestingly a majority of the wards where child benefit counts are much higher than patient register counts contained considerable numbers of armed forces at the LAD level, ward counts not being available for armed forces at present. This may imply either that armed forces dependants are being treated by forces medical care rather than GPs, or that CB records do not capture the migration of such populations as adequately as the patient registers. An element of the differences between the two data sets may also be due to different versions of the geographic lookup file being used.

III.4 Comparison of Patient Register data and Super Older Persons Database

24. There were 6.4 per cent more persons registered on the patient registers aged 65 and over than in the SOPD for England and Wales. Nationally the percentage difference is larger for males than females, 9 per cent compared with 4.5 per cent. The difference is higher for those aged 80 or older (more than 9 per cent) compared with those aged under 80 (less than 5 per cent).

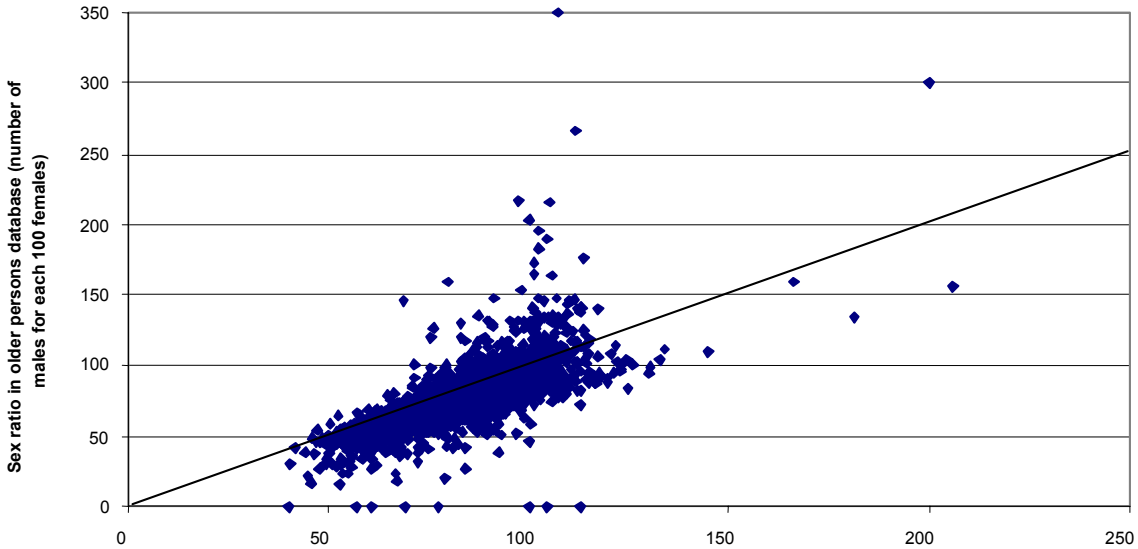
Table 4 Percentage difference between patient registers and older persons database for wards and local authority districts

	LAD	Ward
maximum	30.8%	100.0%
75th percentile	8.4%	13.5%
median	6.2%	6.5%
25th percentile	4.0%	2.5%
minimum	1.1%	-85.6%

25. The percentage difference between the two data sets at LAD level is far less variable than that at ward level (see Table 4). This would be expected because some of the results for wards will have been effected by the disclosure measures used on the data before receipt. The average difference is slightly higher in wards than in local authority districts.

26. Figure 2 compares the sex ratios for wards calculated using these two different data sources. The sex ratios calculated using patient register data tend to be higher than those calculated using the SOPD.

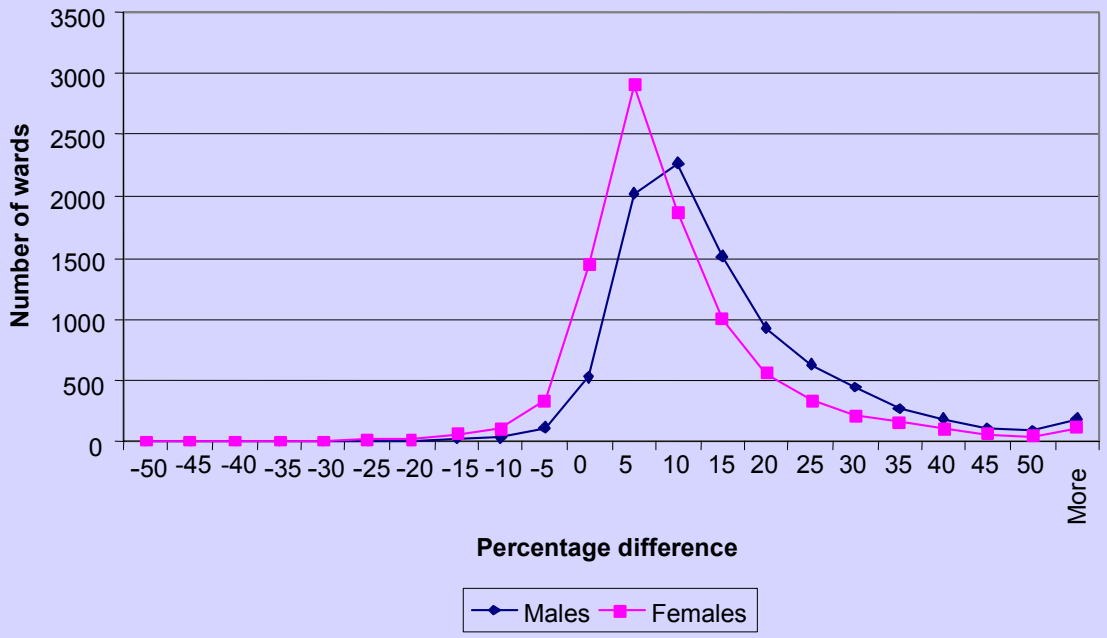
Figure 2: Sex Ratios compared for wards calculated using patient registers and the older persons database, mid-2001 data for wards in England and Wales



However, there are a number of extreme outliers which have been marked, where the sex ratio is much higher in the SOPD than in the patient registers. The quality of the data would need to be investigated in these areas to check validity.

27. At ward level the percentage difference between the datasets for males tends to be higher than that for females. One possible reason for this could be that women are under represented on the SOPD (for instance if they could not claim certain benefits), another could be if men are for some reason over represented on the patient registers. Figure 3 shows the number of wards by percentage difference between the datasets for men and women.

Figure 3: Percentage difference between patient registers and older persons database for males and females aged 65 and over in wards in England and Wales, mid-2001



28. There are some noticeable differences by age group with differences for those aged 80 and over being much greater than for those under 75. The greater variability in differences will be to an extent be because this age group is smaller, they will also be effected to a greater extent by the disclosure measures applied to the SOPD.

29. There are far more extreme values for positive differences between patient register counts and the SOPD than negative ones. Ignoring those where counts are small there are some wards where the differences go against the trend of wards around them. For example in one ward, patient register counts were more than 40 per cent smaller than those on the SOPD, but in a neighbouring ward they were more than 50 per cent higher, this indicates that there may be problems with allocation of data to wards in one of the datasets. This was also observed in some other areas.

III.5 Comparison of 0 year olds in mid-2000 PR against mid-2001 PR

30. The number of 0-year-olds in the July-2001 patient register has been compared to the number of 1-year-olds in the July 2001 patient register. This analysis was undertaken in order to examine whether there is any evidence of a time lag in registering new-borns with doctors. A positive percentage difference

indicates a growth from 2000 to 2001. For both males and females the growth from 2000 to 2001 is 3 per cent at local authority level.

31. The only reason for the number of 1-year-olds to be higher than the number of 0-year-olds, other than a delay in registering births, is in-migration. This is unlikely to account for a median value increase in the population of 1-year-olds between 3 and 4 per cent, as the total change in population in England and Wales is only estimated to be around 0.5 per cent per year. In conclusion it is likely that the percentage difference is due to a delay in the registering of births on the patient registers. The only difference at ward level is that the spread is much larger, with some areas showing an increase of up to 200 per cent, and some a decline of up to 75 per cent, due to the presence of small figures in the data. Table 5 shows the percentage difference at ward and LAD level.

Table 5 Percentage difference between 0-year-olds in mid-2000 and 1-year-olds in mid-2001 Patient Register data at LAD and ward level

	LAD	Ward
Maximum	12.5%	200.0%
75th percentile	5.6%	11.3%
Median	3.5%	3.9%
25th percentile	2.0%	-2.3%
Minimum	-3.8%	-75.0%

IV. CONCLUSIONS AND FURTHER WORK

IV.1 General

32. The analysis has shown interesting differences between the various data sets. However at this stage it is not possible to say which data sets are closest to the true population we are trying to measure.
As would be expected, some wards consistently appear to be outliers on each data set. This would imply that they would be difficult to estimate using these administrative data sets. In order to identify the type of areas involved, a categorisation of area types using the census and other information will be compiled.

IV.2 Specific to data sets

33. Patient register figures are far larger than other data sets. When census data become available it will be necessary to identify where list inflation exists and its extent.

34. Large differences were found between PR and electorate in some wards. It will be necessary to investigate these and identify why they occur. The aim is to try and make the data sets as comparable as possible by adding in foreign (European) nationals and removing attainers from the electorate data. Also under investigation is whether postcoded electorate data are available from commercial companies, and their potential.

35. Only limited analysis of the DWP data has been possible because only ward level data was available by age group and the data had been suppressed in some cases due to disclosure control, and only one extract of each data set being available at present.

36. The Child benefit and patient register figures were closest and generally represent the age groups similarly. However some large outliers will need to be investigated further.

37. For those aged 65 and over, there were more people on PR than on SOPD. Differences were largest in the older age groups and for males. Possible reasons for this will be sought.

38. More sensitivity analysis needs to be undertaken on the data. In particular investigation of time series of data will be done when more years of data are available. This will aid investigation on whether the fluctuations in the data sets at small area level are due to real changes in population or anomalies with the data. It will then also be possible to see whether the assumption of consistency across time with the ratio/additive change population estimation method, or the assumption of consistency across areas with the apportionment population estimation method, applies best to the data.

39. The Patient Register data may not provide good estimates of 0-year-olds due to delay in registering births. This will need to be investigated further.

- - - - -