The Use of Pre-aggregated Administrative Data Sets in Compiling Population Estimates

Part 2: Gender, Age and Intermediate Zones Using Multiple Linear Regression

1. Introduction

Part one of this report summarised the extent to which census population counts at data zone (DZ) level are consistent with counts taken from the optimum combination of the corresponding counts taken from five administrative data sets (hereafter referred to simply as 'sets'). The optimisation was done using multiple linear regression with census count as the dependent variable and counts from each set as the independent or predictor variables. This second part of the report concentrates on the consistency between census counts and estimates based on counts from sets where population cells are defined not just geographically as in Part one but on the basis of a combination of gender, age and geography. Bringing age and gender into the method is necessary since if sets are to be used in quality assurance of the census, there must be evidence that, over all three of the main demographic dimensions, they can produce cell counts which are sufficiently consistent with the census to serve as a comparator.

In Part one, only five of the six available sets were used. The one omitted was the Customer Information Service (CIS) data from Department of Work and Pensions as it was reported only at Intermediate Zone (IZ) level and therefore did not have the geographical detail required for use in a DZ-based analysis. In this second Part, cells are defined at GAIZ level (referring to Gender, Age and Intermediate Zone) and so CIS can be included. This is important since CIS is only one of two sets (NHS Central Register (NHSCR) being the other) which covers all ages and both genders and it is valuable to assess the extent to which CIS can contribute to GAIZ counts based on sets.

The fact that some sets do not cover all ages poses a particular problem when using them in multiple linear regression. This is that the sets are trying to predict different dependent variables. NHSCR and CIS try to predict all people in a given geographical area; Child Benefit (CB) only those up to approximately age 16; School Census (SC) those between ages 5 and 18; Electoral Register (ER) those aged 16 or above; and Super Older Persons' Database (SOPD) those aged 65 or above. In Part one this problem was bypassed (rather than solved) by treating these partial DZ counts as proxies for the total counts covering all ages. This in effect assumes that the number of people in a given age range is proportional over geographies to the total number of people in that geography. While not wholly correct, there will in practice be a high level of correlation between partial and total population sums. The method used in Part one could be made more sensitive (but would be more onerous to implement) by undertaking different regressions for different age groups covered by different combinations of sets. In this second Part, where age is an explicit variable, this problem cannot be bypassed in this way. Hence separate regressions must be undertaken, involving the partition of the original file into several parts and the running of a regression for each part.

Table 1 shows which sets cover which five year age groups (and hence shows that four different regressions must be undertaken). The age coverage of the sets does not coincide with the five year division so some data is lost. Child Benefit (CB)

coverage of ages 15 and 16 is an example: some research was done on how far CB can predict the census count of age group 4 (ages 15 to 19) but the accuracy was found to be poor and CB was not used as a predictor for this group.

Table 1: Data set coverage of age groups

Regression label	Age group(s)	Ages covered	Sets included
124	1	0 to 4	NHSCR, CIS, CB
1245	2 and 3	5 to 14	NHSCR, CIS, CB, SC
12	4 to 13	15 to 64	NHSCR, CIS
126	14 to 17	65 and above	NHSCR, CIS, SOPD

A further complication concerns the ER set. While this covers the ages from 16 upwards, it does not have a field for age and hence does not supply values for this predictor variable. While this was not insuperable for an analysis summing over age, it is insuperable for the present report. Also, ER does not have a field for gender and so it does not feature in the any of calculations reported below.

Table 1 shows a potential weakness of the analysis for present purposes. This is that for age groups 4 to 13 (corresponding broadly to the period between leaving school and retirement), only the two universal sets can be used. It will therefore be valuable to explore the extent to which accuracy is related to the number of sets used to compile the cell count estimates.

2. Procedure

The procedure will be as similar as possible to that used in Part one of this report. It is summarised here since some changes have been necessary in order to accommodate the introduction of gender and age and the use of a higher level geography. However the basic procedure (referred to in Part one as the red, white and blue model) remains the same. It has the following stages:

a. Identifying the ‘red’ GAIZ cells which cannot be predicted accurately from sets and removing them from the analysis.
b. Identifying from amongst the remaining cells a calibration sample (the ‘blue’ cells) which will be used to calculate regression parameters.
c. Using these parameters to calculate estimates for the remaining ‘white’ cells.

In the following account, it was assumed that census counts were not initially available for any cells but that they became available for the red and blue cells after they had been identified as the two parts of the direct enumeration sample and surveyed. The fourth stage used in Part one (i.e. integrating the red, white and blue cells to provide estimates for all cells) was not used here. Instead, attention will focus on the consistency between actual and estimated census counts for the white cells.

In what follows, cells were defined by two genders, 17 five-year age bands and 1,235 IZs giving a total of 41,990 cells.[1] A source file was compiled with this number of records and, for each record, fields for the un-rebased Small Area Population Estimates (SAPE) count, the 2011 Census count and counts from each of the five sets used in this report which covered that age band.

2A. The red cells

The source file was partitioned into four parts following the rows of table 1 and a separate regression was conducted for each. In all cases, un-rebased SAPE was used as the dependent variable. This which would be available in the absence of a census and can be used as a proxy for it for the purpose of identifying those cells which are hard to predict accurately. The predictor variables were the sets which cover the relevant age group(s) and also gender.[2] The results are reported in table 2. For each variable in each analysis, the table gives the ß regression coefficient (the numbers of asterisks denoting significance at the 5%, 1% and 0.1% levels) and the variance inflation factor.[3] Table 2 also gives, in the final two rows, the Root Mean Square Error (RMSE) between the observed and predicted values and the coefficient of variation, defined as 100 times the RMSE divided by the mean value of the dependent variable.

Table 2 : Regression of unrebased SAPE against five data sets - all 41,900 cells

Regression label	124		1245		12		126	
Variable	ß	VIF	ß	VIF	ß	VIF	ß	VIF
Intercept	3.04***	0	6.78***	0	12.52***	0	8.16***	0
1: NHSCR	0.93***	25.37	0.87***	23.61	0.73***	3.023	0.81***	40.37
2: CIS	-0.03*	15.38	0.10***	14.17	0.20***	3.022	-0.01	10.63
4: CB	0.10***	47.99	-0.01	50.20				
5: SC			-0.03*	17.93				
6: SOPD							0.16***	41.80
gender	-0.02	1.00	0.25	1.00	-6.37***	1.00	-2.18***	1.11
	RMSE	C of V	RMSE	C of V	RMSE	C of V	RMSE	C of V
	7.1	5.2%	10.8	9.6%	25.2	17.8%	9.8	10.9%

Table 2 shows that not all of the independent variables were statistically significant as predictors. In the interests of model parsimony, those which did not reach significance at the 5% level were discarded and the regression run again. The new coefficients for the remaining variables are not reported here as they were very close to those in table 2. However one benefit of discarding non-significant variables was to remove some of the highest VIF values, indicating that one reason for the failure of a ß value to reach significance was a very high level of correlation with another predictor variable.

Footnotes

1) This contrasts with the 6,505 DZs used in part 1 and means that comparisons will not be possible as error components in data tend to make up a smaller proportion of total variance with a smaller number of cells.
2) Gender is available for all five sets and can be used in regression because, as a dichotomous variable, it can be treated as an interval variable.
3) Variance inflation is related to the squared multiple correlation for each predictor variable regressed against all the other predictor variables. High values indicate a high degree of inter-correlation between the predictors, as would be expected in this case, and in extreme cases it can make the regression algorithm unstable. In this case it does not seem to have prevented satisfactory convergence but it is a matter which should be kept in mind when using multiple regression in this way.

The regressions reported in table 2 were used to identify the red cells which are to be enumerated directly due to their lack of predictability on the basis of the sets. It is important to know how far this procedure identifies those cells where the actual census count is not predictable (i.e. the extent to which cells with unpredictable SAPE counts are also those with unpredictable census counts). To this end, the regressions in table 2 were run again, this time using the actual census count as the dependent variable rather than un-rebased SAPE, and the results were compared. All the residuals between predicted and actual values were expressed as a percentage of the actual value. The distribution of each was compiled and symmetric cut-off scores calculated which cut off (at the two tails combined), 5% and 32% of the cells, approximating ±2 and ±1 standard deviations of the normal distribution. These cut-off points were ±28 and ±11 for the SAPE regression and ±22 and ±8 for the census regression. On the basis of these cut-offs, each score was replaced by a number between +2 and -2 denoting which of the five areas of the its distribution it was located in (e.g. zero denotes within one standard deviation of the mean). The census and SAPE results were then cross-tabulated and the results are given in table 3.

Table 3: SAPE and census cell counts (standardised distributions)

		Census					
		-2	-1	0	+1	+2	total
SAPE	-2	135	76	112	66	57	446
	-1	187	1,345	2,575	531	190	4,828
	0	169	4,118	20,857	3,534	715	29,393
	+1	12	425	3,519	1,307	440	5,703
	+2	5	85	654	528	348	1,620
	total	508	6,049	27,717	5,966	1,750	41,990

Table 3 shows a clear degree of correlation between the two sets of regressions, though there are small numbers of cells where the results are markedly different. There are for example 57 cells where the SAPE prediction was two standard deviations or more below the mean but the census prediction was two standard deviations or more above it and five cells where the opposite was the case. As would be expected, a regression using SAPE as dependent variable does not identify all the cells which would be identified using census counts and it includes some which the census would not but it could serve as a reasonable starting point for identifying red cells.

It should perhaps be emphasised at this point that this part of the procedure is rather artificial and is only done in this way due to the manner in which the CIS set is structured. In practice, enumeration, either for the purposes of gathering counts for red cells or for the calibration of blue cells would not be done on the basis of GAIZ cells (e.g. requiring enumerators to count the number of males aged 10 to 14 in a given IZ). Rather, enumeration would aim to cover everyone in a given

geography and DZ would almost certainly be used for this purpose. However to insist on DZ-level enumeration would prevent the CIS set being used in either Part one or Part two of the report and it would be undesirable to make no use at all of one of only two sets which offer coverage of the whole population.

2B. The blue and white cells

The cells whose predicted Small Area Population Estimates (SAPE) counts were more than 28% away from the observed values (i.e. were the least predictable 5% of the cells) were deemed to be red and removed from the file. The remaining cells were sorted by gender, age group and IZ code and every 20th record was deemed to be blue. There were 1,997 of these. The blue file was split into four parts in accordance with table 1 and four multiple regressions were run using census count as the dependent variable. Table 4 gives the results.[4] As before, where a coefficient did not reach significance at the 5% level, it was discarded and the regression was re-run but the revised ß values were not greatly altered and are not reported separately.

Table 4: Regression of 2011 Census against five data sets - 1,997 blue cells only

Regression label	124		1245		12		126	
Variable	ß	VIF	ß	VIF	ß	VIF	ß	VIF
Intercept	0.03	0	4.78	0	21.18***	0	2.33*	0
1: NHSCR	0.38***	38.23	0.60***	23.44	0.78***	4.04	0.54***	39.32
2: CIS	0.07	22.27	0.18*	18.12	0.11***	4.03	0.02	9.87
4: CB	0.59***	77.77	0.22	55.93				
5: SC			-0.01	20.11				
6: SOPD							0.43***	38.98
gender	0.21	1.02	-1.16	1.02	-6.46***	1.00	-0.53	1.14
	RMSE	C of V	RMSE	C of V	RMSE	C of V	RMSE	C of V
	11.48	9.45%	10.63	9.50%	20.31	14.05%	5.20	5.83%

Tables 2 and 4 contain some important information about the relationship between the number of sets on which estimates are based and the accuracy of those estimates. In both cases, the lowest level of accuracy occurred in regression label 12 where only NHSCR and CIS provided coverage. It appears that the age-specific sets which cover younger and older age ranges have an important part to play in improving the accuracy of cell counts in estimating both SAPE and the census.

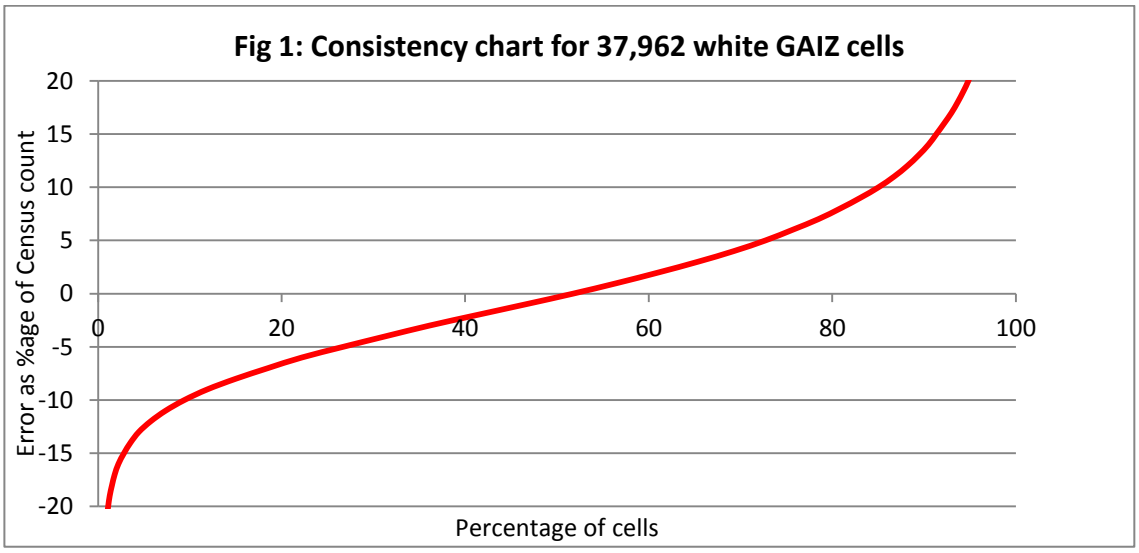
Footnote

4) It is worth noting that a comparison of table 2 and table 4 indicates that administrative data sets are better at predicting SAPE than the census for low ages but are better at predicting the census for higher ages.

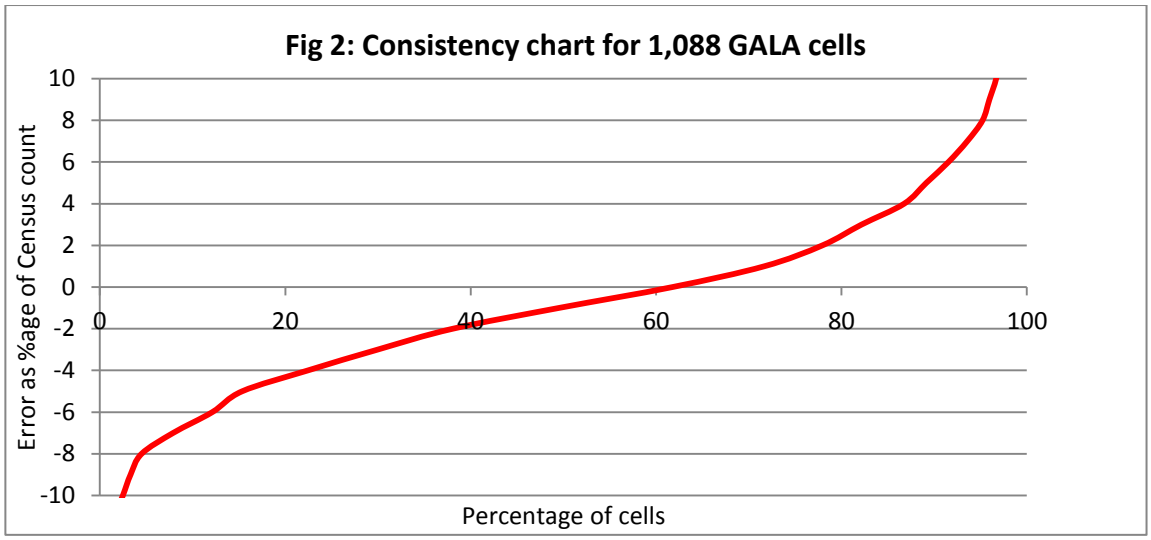
3. Consistency between Census counts and estimates from the sets for the white cells

Using the coefficients in table 4, an estimate of the census counts in each of the 37,962 white cells was calculated. Overall measures of the consistency are given by the root mean square error which was 17.50 and the coefficient of variation which was 13.83%.

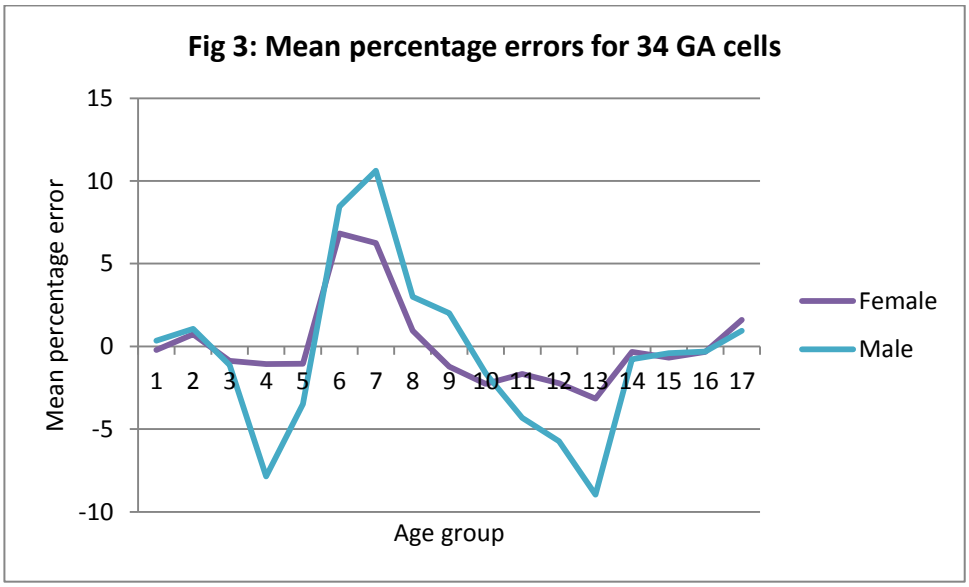
The estimate was then expressed as a percentage of the actual count and a consistency chart was compiled from the results. It is given in fig 1 which shows that 26.7% of the cells are 5% or more below the census count while 27.3% are 5% or more above the census count. The proportions on or outside ±10% are 9.4% below and 15.0% above. The proportions on or outside ±20% are 1.1% below and 5.2% above. However it should be noted that these are very small cells (there are on average over a thousand of them in each local authority) and in general, error components account for a larger proportion of total variance for small cells than they do for larger ones.



To illustrate this, the data charted in fig 1 were summed over all IZs within each LA to give a total of 2 x 17 x 32 = 1,088 cells. The summed data is at GALA (gender / age / LA) level and the consistency chart is given in fig 2. Note that the scale on the vertical axis is only half of that in fig 1 which reflects the much higher level of consistency for the less detailed cell structure. Here, only 6.5% of cells are as much as +10% above or -10% below the census count.

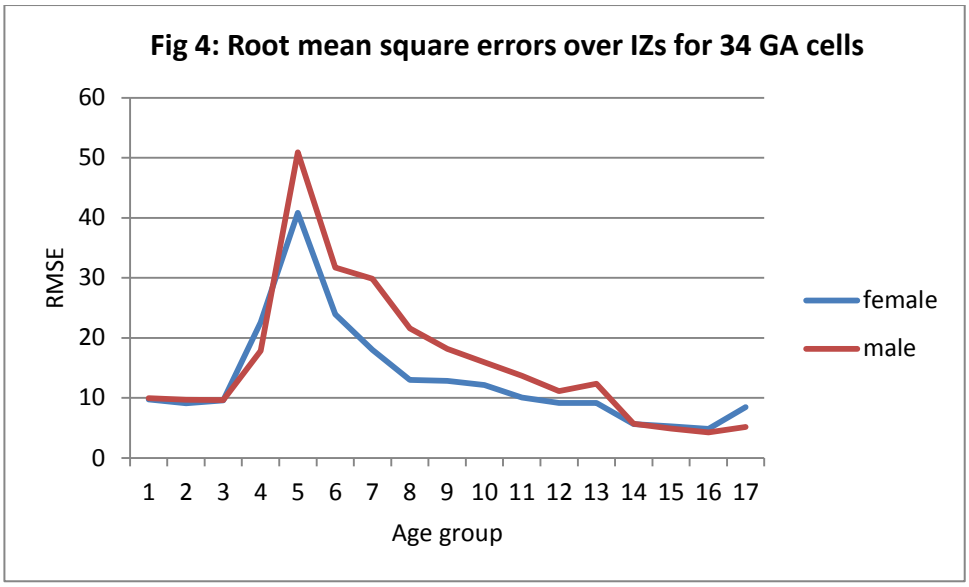


It is also of interest to explore whether the residuals vary systematically over gender and age group. The assess this, the estimates and census counts of the white cells were summed over geography into 34 GA (gender by age group) cells. For each of these the percentage error was calculated and the results are given in fig 3. The sum of the points plotted in fig 3 is close to zero.[5] The pattern shows clear gender / age effects and the method could be made more precise by incorporating these into the calculation, possibly by estimating gender / age effects from the blue cells and assuming that similar effects occurred in the white cells.



Footnote

5) It would be exactly zero if the regression coefficients had been calculated from the white cells rather than the blue. The sum will therefore differ from zero in so far as the blue cells are not representative of the white cells. This, as would be hoped, is not very far.



It was noted above that it would be useful to investigate how the RMSE between the actual and estimated cell counts varies as a function of gender and age for the white cells. This is given in fig 4 which differs from fig 3 mainly in not taking account of the value of the actual census count for each cell within a GA combination or of the direction of the difference. Rather, the squared deviations are simply averaged over all cells in each GA combination. In effect, fig 4 unpackages the overall RMSE value of 17.50 given at the start of section 3 into gender and age components. The results, as would be expected, have the same overall shape as fig 3 with errors being larger for males than for females and a 'spike' starting at age group 4 and tailing off at around age group 10.

4: Conclusion

The above results suggest that the extent to which GAIZ cell counts based on sets are consistent with actual census counts varies with gender, with age and with their interaction. It is most consistent at the two ends of the age range and least consistent over the large period covering working age. This is also the range covered by only two sets but it is likely that consistency would be least in this age range anyway as it is associated with greatest geographical and social mobility. In any case there are strong reasons for finding and adding to the existing sets at least a further one which could supply gender / age / data zone counts for the age ranges from 15 to 64 (Her Majesty's Revenue and Customs (HMRC) being an obvious source), though the barriers to achieving this would be considerable. It would also be very valuable if the Customer Information Service (CIS) counts supplied by Department for Work and Pensions (DWP) could be by gender / five year age band/ DZ level rather than gender / single year / IZ level as the present counts are. This would allow the CIS set to be used with an enumeration model for the red and blue cells based only on DZ rather than on a combination of gender, age and geography.

The figures presented in section 3 above suggest that for many of the cells, the predictions are accurate enough at least to serve as a starting point. Figs 1 and 2 indicate that about three quarters of the smaller GAIZ cells can be predicted to within ±10% and about the same proportion of the larger gender / age / local authority cells to within about ±5%. Accuracy of the less well predicted cells could

be improved by taking account of other relevant information. This would include knowledge of the location, nature and size or communal establishments and other local circumstances known to be relevant to population cell counts and in particular to the extent to which these are likely to be predictable on the basis of administrative data sets.

A final strategic point about the red/white/blue method should be made. The method is used above to assess the consistency between the census and predictions made from data set counts on the assumptions (i) that 5% of red cells are excluded as being insufficiently predictable and (ii) that a calibration sample of 5% (the blue cells) is used to make predictions for the remaining 90% of white cells. It could also be used for QA of an already existing census. If all the non-red cells were to be deemed blue (hence removing white cells altogether from the procedure) then the method would be configured for QA. The second regressions (following the exclusion of the red cells) would cover all remaining cells and the predictions compared to the census counts (which would be available for all cells). As a further special case, the proportion of red cells could be set at zero and all cells included in a single set of regressions. This would cause the tails of the consistency charts in figs 1 and 2 to depart further from the horizontal axis but would provide a picture of all the inconsistency recorded between census counts and those derived from sets.

Alternative Sources
July 2014