Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

EVALUATION OF SMALL AREA ESTIMATION METHODS – AN APPLICATION TO UNEMPLOYMENT ESTIMATES FROM THE UK LFS

Gary Brown[1], Ray Chambers[2], Patrick Heady[1], Dick Heasman[1]

ABSTRACT

This paper describes joint research by the ONS and Southampton University on the evaluation of several different approaches to the local estimation of ILO unemployment. The need to compare estimators with different underlying assumptions has led to a focus on evaluation methods that are (partly at least) model-independent. Model fit diagnostics that have been considered include various residual procedures, cross-validation, predictive validation, consistency with marginals, and consistency with direct estimates within single cells. These have been used to compare different model-based estimators with each other and with direct estimators.

KEY WORDS: Diagnostics; Estimates; Bias; Standard errors; Confidence intervals.

1. INTRODUCTION

A small area estimation methodology can be thought of as a model plus fitting method for the small area values of interest coupled with an estimation method based on the fitted model. Basic properties that we require of such a methodology are:

1. The expected values defined by the model underlying the small area values should be “good”. That is, they should explain a significant proportion of the variation in the small area values of interest. Note that for models that include random effects, these do not contribute to the expected value.
2. The values for the model-based estimates derived from the fitted model should be consistent with the unbiased direct survey estimates, where these are available. That is, they should provide an approximation to the direct estimates that is consistent with these values being "close" to the expected values of the direct estimates.
3. The model-based small area estimates should have mean squared errors significantly lower than the variances of corresponding direct estimates.
4. The changes over time in the model-based estimates for a particular small area should be more stable than the corresponding changes in the direct estimates over the same time.
5. The model-based estimates for a particular small area should be acceptable to informed users from that small area.

This paper does not attempt to cover all of the above agenda. Clearly, standard model-fitting diagnostics can be used to assess property 1 – and we have also restricted the discussion to indicators that relate to a single point in time, thus excluding point 4. The most important omission, however, is point 5. Despite the fact that user-consultation is not discussed here, ONS takes the process of consultation very seriously indeed – both as a way of ensuring public acceptance, and as a valuable input to improving the estimates themselves.

[1] Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ, U.K.
[2] Department of Social Statistics, University of Southampton, SO17 1BJ, U.K.

This paper is about the preliminary internal evaluation work needed to select a suitable small area estimator in situations where there are a number of competing small area models that are not necessarily nested and there is some doubt about the assumptions underpinning all of these models. In particular, we discuss four diagnostics that we have found useful in this regard. These assess the bias and goodness of fit of the estimation method, the coverage of the confidence intervals generated by the method and the calibration error of the method. All are based on the crucial assumption that the direct estimates of the small area values of interest are unbiased (but highly variable) and the confidence intervals associated with these estimates achieve their nominal coverage levels.

These diagnostics have been developed in the process of investigating small area estimators for both unemployment and a range of other socio-economic variables. The theory behind these estimators is described in Ambler et al (2001) and ONS (2001). In the following section we describe the diagnostics in more detail, applying them to a small area unemployment estimator from Ambler et al (2001). In practice several diagnostics are used at each stage of the model selection process.

2. DIAGNOSTICS

2.1 A bias diagnostic

- {*The direct estimates are unbiased estimates of the "truth" – if the truth were known and plotted on the X axis of a graph, with direct estimates as Y, the regression line would fall on 45^o. We plot the model estimates as X, in place of the "truth", and see how close the regression line is to Y=X. This provides a visual illustration of bias, and by comparing the regression line with Y=X, a parametric significance test for the bias of the model estimates*[3].}

The diagnostic is based on the following idea. If the model-based estimates are "close" to the small area values of interest, then unbiased direct estimators should behave like random variables whose expected values correspond to the values of the model-based estimates. That is, the model-based estimates should be unbiased predictors of the direct estimates. As a check for such predictive (i.e. conditional) bias in the model-based estimates, we plot appropriately scaled values of these estimates (X-axis) against similarly scaled direct estimates (Y-axis) and then test whether the OLS (ordinary least squares) regression line fitted to these points is significantly different from the identity line[3].

When there is significant variation in small area sizes this test typically requires an initial transformation of both the direct and model-based estimates so that the homoskedasticity assumption underpinning the OLS fitting method is satisfied. Such a transformation can be identified using standard methods. In our unemployment example below, a square root transformation was used, since the estimates relate to counts of unemployed people in the small areas of interest.

The use of this diagnostic is straightforward when the focus of interest is on small area totals since unbiased direct estimators of such totals are typically available. The use of transformations to stabilise the residual variance in the plot will of course introduce a slight bias, but we feel that this is acceptable. However the issue becomes more complex when the focus of interest is on small area proportions, because the denominator of the direct estimator of such a proportion is typically a random variable and so the proportion is in effect a ratio estimator and hence biased. We have adopted two different strategies in this case:

(I) Concentrate on the numerator of the estimated proportion - the estimated small area total - since this can be estimated without bias.

(II) Compare the direct and model-based estimators of the proportion and accept that the resulting ratio bias may slightly distort the interpretation of the diagnostic.

[3] The calculated significance values do not allow for the fact that the X values are derived from the same data as the Y values. This will often make the true rejection probability of the test lower than its face value – in which case an apparent rejection would be more significant than it seemed.

The relative attractiveness of these two options depends in part on characteristics of the sample and of the population of the small areas of interest. In the case of strategy (I) there is a danger that, if these population sizes vary a great deal, the pattern shown in the scatterplot will owe more to this variability than to the biasedness or otherwise of the model-based estimators of the small area proportions. In the case of strategy (II) the lower the coefficient of variation of the denominator of the small area proportion, the lower the risk of serious bias in the direct estimate of the proportion, and hence the more applicable the diagnostic.
Finally, if the model underlying the small area estimates is actually fitted using proportions, strategy (II) can also be interpreted in the following way. It provides a way of looking for bias due to model misspecification - but cannot be expected to discover any bias in the direct estimates which were used in fitting the model.

Example

In this example the variable of interest is the number of individuals who are unemployed according to the International Labour Organisation definition ("ILO unemployed") in 406 LAD/UAs (Local Authority Districts and Unitary Authorities) in Great Britain. Direct estimates for this variable are available annually from the Local Area Data Base of the UK Labour Force Survey (LFS). A number of logistic models for the probability of being unemployed in a LAD/UA were fitted to these data and used to define model-based estimates for these small areas. See Ambler et al (2001) for further details. Here we focus on the modified Fay-Herriot approach concentrated upon in that paper. The model includes covariates defined by the claimant count in six age by sex cells within each LAD/UA (the claimant count is the number of people who claim unemployment benefits), as well as age by sex effects, LAD/UA regional effects and LAD/UA socio-economic classification effects. In addition, the model includes an extra area level effect, defined by the logit of the total LAD/UA claimant count (as a proportion of the LAD/UA population), which is used a measure of the overall economic activity within the LAD/UA, and consequently reflects an individual's opportunity to obtain employment in that area.

The OLS regression parameters, with standard errors in brackets, from the bias scatterplot for the five years 1995/1996 to 1999/2000 are given in Table 1. None of the regression lines show a significant difference from Y=X. A visual illustration is given in Figure 1 for 1999/2000, where the Y=X and regression lines show very little disparity.

	Intercept	**Slope**
1995/1996	0.463 (1.062)	0.989 (0.014)
1996/1997	-1.354 (1.060)	1.010 (0.014)
1997/1998	-0.832 (1.108)	1.003 (0.016)
1998/1999	-0.615 (1.104)	0.999 (0.017)
1999/2000	-1.927 (1.135)	1.017 (0.018)

Table 1 OLS regression parameters from bias scatterplots

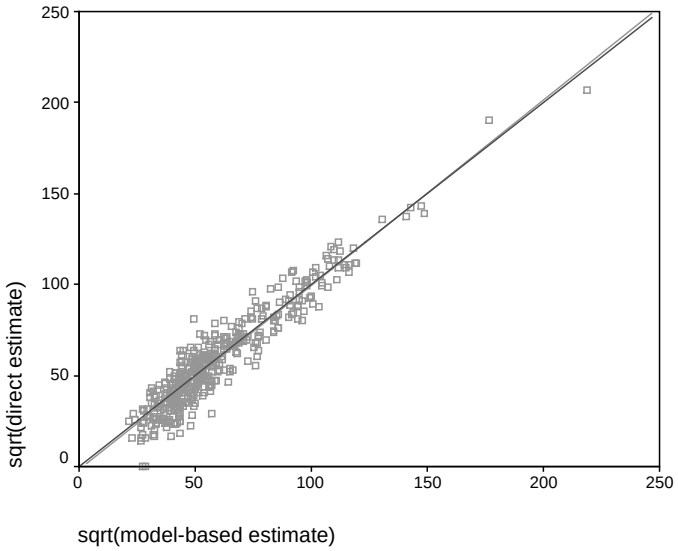

Figure 1 Bias scatterplot for 1999/2000 with Y=X and regression lines fitted

The interpretation changes when scatterplots and regression lines are fitted for the estimated proportions of individuals who are ILO unemployed for LAD/UAs. Figure 2 shows the scatterplot for 1999/2000, with a regression line Y = –0.0145(0.009) + 1.059(0.048)X. This shows more disparity from Y=X and hence more possible bias, although the evidence is still not statistically significant.

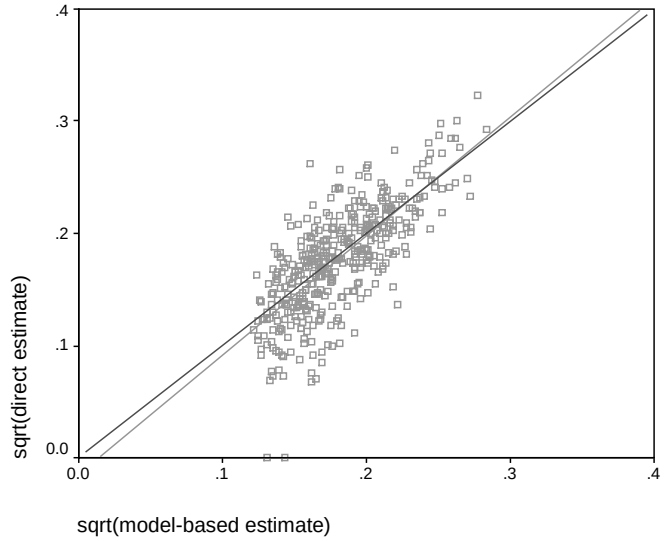

Figure 2 Bias scatterplot for proportions for 1999/2000 with Y=X and regression lines fitted

2.2 A goodness of fit diagnostic

- {*We want our model estimates to be close to the direct estimates when the direct estimates are good. We inversely weight their squared difference by their variance and sum over all areas – this sum gives more weight to differences from good direct estimates than from bad. We test this sum against the* χ^2 *distribution to provide a parametric significance test of bias of model estimates relative to their precision.*}

As a check for unconditional bias in the model-based estimates we use a Wald goodness of fit statistic to test whether there is a significant difference between the expected values of the direct estimates and the model-based estimates.

In order to describe this test, we assume that the variable of interest is unemployment status and the available data consist of direct estimates (from the LFS) of the population proportion unemployed by age and sex in each small area, together with model-based estimates of the population proportion unemployed in each small area. Let i denote age-sex class and j denote small area. We assume that age-sex classes are finely enough defined so that within an age-sex class in an small area there is little or no variation in the sample weights. This allows us to define an "average" weight for all individuals in age-sex class i in small area j,

$$w_{ij} = \frac{\sum_{k \in s(ij)} w_k}{n_{ij}} \tag{1}$$

and to therefore approximate the direct estimate $\hat{z}_{ij}$ of the population proportion unemployed in age-sex class i in small area j by

$$\hat{z}_{ij} \approx \frac{w_{ij} \sum_{k \in s(ij)} U_k}{w_{ij} n_{ij}} = \frac{\sum_{k \in s(ij)} U_k}{n_{ij}} = \hat{p}_{ij} \tag{2}$$

where s(ij) denotes the individuals in age-sex class i who are in the survey sample in small area j, U_k takes the value 1 if individual k is unemployed and zero otherwise, and w_k is the survey sample weight attached to individual k. Note that $\hat{p}_{ij}$ is just the sample proportion of interest in age-sex class i and small area j. The corresponding approximation to the direct estimate of the proportion of unemployed in the small area is then

$$\hat{z}_j \approx \frac{\sum_i w_{ij} n_{ij} \hat{p}_{ij}}{\sum_i w_{ij} n_{ij}} = \frac{\sum_i w_{ij} u_{ij}}{\sum_i w_{ij} n_{ij}} \tag{3}$$

where u_{ij} denotes the total number of unemployed in sample in age-sex class i in small area j. Since the sample design of the LFS is essentially simple random sampling within a small area, we can model the sample counts $\{u_{ij}\}$ and $\{n_{ij}\}$ as realisations of correlated multinomial random variables. In particular, let π_{ij} be the probability that a randomly chosen individual in age-sex class i and small area j has the characteristic of interest and ϕ_{ij} be the probability that a randomly chosen individual in small area j is in age-sex class i. Then

$$E(u_{ij}) = n_j \pi_{ij} \phi_{ij},\ \mathrm{var}(u_{ij}) = n_j \pi_{ij} \phi_{ij} (1 - \pi_{ij} \phi_{ij}),\ \mathrm{cov}(u_{ij}, u_{i'j}) = -n_j \pi_{ij} \phi_{ij} \pi_{i'j} \phi_{i'j} \tag{4}$$

$$E(n_{ij}) = n_j \phi_{ij},\ \mathrm{var}(n_{ij}) = n_j \phi_{ij} (1 - \phi_{ij}),\ \mathrm{cov}(n_{ij}, n_{i'j}) = -n_j \phi_{ij} \phi_{i'j} \tag{5}$$

and

$$\mathrm{cov}(u_{ij}, n_{ij}) = n_j \pi_{ij} \phi_{ij} (1 - \phi_{ij}),\ \mathrm{cov}(u_{ij}, n_{i'j}) = -n_j \pi_{ij} \phi_{ij} \phi_{i'j}. \tag{6}$$

First order approximations to the expected value and variance of $\hat{z}_j$ are

$$(7) \qquad E(\hat{z}_j) \approx \frac{\sum_i w_{ij}\phi_{ij}\pi_{ij}}{\sum_i w_{ij}\phi_{ij}} = \zeta_j,\ var(\hat{z}_j) \approx \frac{\underline{w}_j^T var(\underline{u}_j - \zeta_j \underline{n}_j)\underline{w}_j}{(\underline{w}_j^T E(\underline{n}_j))^2}$$

where $\underline{u}_j$, $\underline{n}_j$ and $\underline{w}_j$ are the vectors with components $\{u_{ij}\}$, $\{n_{ij}\}$ and $\{w_{ij}\}$ respectively. Furthermore, the components of $var(\underline{u}_j - \zeta_j \underline{n}_j)$ are given by

$$(8) \qquad var(u_{ij} - \zeta_j n_{ij}) = n_j \phi_{ij}(1-\phi_{ij})\left[\frac{\pi_{ij}(1-\pi_{ij}\phi_{ij})}{1-\phi_{ij}} - 2\zeta_j \pi_{ij} + \zeta_j^2\right]$$

and

$$(9) \qquad cov(u_{ij} - \zeta_j n_{ij}, u_{i'j} - \zeta_j n_{i'j}) = -n_j \phi_{ij}\phi_{i'j}(\pi_{ij} - \zeta_j)(\pi_{i'j} - \zeta_j).$$

Typically, small area-level model-based and direct survey estimates will be approximately uncorrelated. Consequently, a Wald statistic for testing the small area-level goodness-of-fit of a model-based set of estimates of interest is

$$(10) \qquad W = \sum_j \frac{(\hat{z}_j - \hat{\zeta}_j)^2}{\hat{V}(\hat{z}_j) + \hat{V}(\hat{\zeta}_j)}$$

where $\hat{\zeta}_j$ is the model-based estimate of the proportion of small area j population that are unemployed, $\hat{V}(\hat{\zeta}_j)$ is its estimated variance and

$$(11) \qquad \hat{V}(\hat{z}_j) \approx \frac{\underline{w}_j^T \hat{V}(\underline{u}_j - \zeta_j \underline{n}_j)\underline{w}_j}{(\underline{w}_j^T \hat{E}(\underline{n}_j))^2} = \frac{\underline{w}_j^T \hat{V}(\underline{u}_j - \zeta_j \underline{n}_j)\underline{w}_j}{\hat{N}_j^2}$$

where $\hat{N}_j$ is the survey estimate of the population of the small area and $\hat{V}(\underline{u}_j - \zeta_j \underline{n}_j)$ is a matrix with diagonal components

$$(12) \qquad n_{ij}\left(1 - \frac{n_{ij}}{n_j}\right)\left[\frac{\hat{\pi}_{ij}(n_j - \hat{\pi}_{ij} n_{ij})}{n_j - n_{ij}} - 2\hat{\zeta}_j \hat{\pi}_{ij} + \hat{\zeta}_j^2\right]$$

and off-diagonal components

$$(13) \qquad -\frac{n_{ij} n_{i'j}}{n_j}(\hat{\pi}_{ij} - \hat{\zeta}_j)(\hat{\pi}_{i'j} - \hat{\zeta}_j).$$

Here $\hat{\pi}_{ij}$ is the model-based estimate of the proportion unemployed in age-sex class i in small area j.

Under the hypothesis that the model-based estimates are equal to the expected values of the direct estimates, and provided the sample sizes in the small areas are sufficient to justify central limit assumptions, W will then have a χ^2 distribution with degrees of freedom equal to the number of small areas in the population.

Example

We continue with the model-based approach introduced earlier for the same five years. The goodness of fit statistics are in Table 2. None of the statistics show evidence to reject a χ^2 distribution, if fact the fit seems almost too good. This may be an artefact of including estimated between LAD/UA variance in the mean squared error of the model-based estimates, which errs on the side of caution.

1995/1996	1996/1997	1997/1998	1998/1999	1999/2000
349.84 [p-value 0.98]	358.80 [p-value 0.96]	376.21 [p-value 0.85]	349.59 [p-value 0.98]	377.85 [p-value 0.84]

Table 2 Goodness of fit statistic values with [p-values]

2.3 A coverage diagnostic

- {*95% Confidence intervals for the direct estimates should contain the "truth" 95% of the time. So should the confidence intervals surrounding model-based estimates. We adjust both sets of intervals, so that their chance of overlapping should be 95% and count how often they actually do overlap. Assuming that the estimated coverage of the direct confidence intervals is correct, comparing the counts to the Binomial distribution provides a non-parametric significance test of the bias of model estimates relative to their precision.*}

This diagnostic evaluates the validity of the confidence intervals generated by the model-based small area estimation procedure. It assumes that valid 95 percent confidence intervals for the small area values of interest can be generated from the direct estimates. The basic idea then is to measure the overlap between these direct confidence intervals and corresponding 95 percent confidence intervals generated by the model-based estimation procedure. However, since the degree of overlap between two independent 95 percent confidence intervals for the same quantity will be higher than 95 percent, it is necessary to first modify the nominal coverage levels of the confidence intervals being compared in order to ensure a nominal 95 percent overlap.

This modification is based on the fact that if X and Y are two independent normal random variables, with the same mean but with different standard deviations, σ_X and σ_Y respectively, and if z(α) is such that the probability that a standard normal variable takes values greater than z(α) is $\alpha/2$, then a sufficient condition for there to be probability of α that the two intervals X ± z(β)σ_X and Y ± z(β)σ_Y do not overlap is when

$$z(\beta) = z(\alpha)\left(1+\frac{\sigma_X}{\sigma_Y}\right)^{-1}\sqrt{1+\frac{\sigma_X^2}{\sigma_Y^2}}\,. \tag{14}$$

Consequently, this diagnostic takes z(α) = 1.96, calculates z(β) using the above formula, with σ_X replaced by the estimated standard error of the model-based estimate and σ_Y replaced by the estimated standard error of the direct estimate and then computes the overlap proportion between the corresponding

z(β)-based confidence intervals generated by the two estimation methodologies. Nominally, for z(α) = 1.96, this overlap proportion should be 95 percent. Note that z(β) = z(α) when σ_X = 0.

This diagnostic can also be used to assess the need to include a small area random effect in the model, by just looking at the proportion of direct estimate-based confidence intervals that cover the model-based estimates of the expected values of the small area quantities of interest. Ideally, if the model-based estimator is essentially the small area quantity of interest, then around 5% of the small areas will record such noncoverage. However, if small area level random effects are present (i.e. a multilevel model is more appropriate than a single level model) then more than 5% of small areas will necessarily show noncoverage. Used in this way, this diagnostic can be interpreted in two ways, as a test for bias in a single level model, or as a test for whether a multilevel model is needed – the interpretation depending on whether a single level model is known to be sufficient or not.

Example

We continue with the model-based approach introduced earlier for the same five years. Non-coverage totals and percentages are shown in Table 3 (we filter out zero direct estimates of unemployment). For 1997/1998 and 1999/2000 there is significant evidence to reject 5% non-coverage. However, this means we have overcoverage, and the mean squared error of the model-based estimates is too large. As this is erring towards giving conservative confidence intervals it is not a major cause for concern.

1995/1996	1996/1997	1997/1998	1998/1999	1999/2000
11 out of 406 (2.7%) [p-value 0.03]	13 out of 406 (3.2%) [p-value 0.11]	17 out of 406 (4.2%) [p-value 0.54]	11 out of 406 (2.7%) [p-value 0.03]	13 out of 406 (3.2%) [p-value 0.11]

Table 3 Non-coverage totals with (percentages) and [p-values]

2.4 A calibration diagnostic

- {*Calculating how much modelled estimates differ from direct estimates when aggregated to larger domains shows us whether any particular larger domain is estimated worse than any other. For example this may show how a model may be poorly estimating large urban areas, whilst estimating large rural areas well. This provides some evidence regarding spatial bias/autocorrelation of model estimates. However, the value of the evidence depends on the size of domains in question.*}

The final diagnostic we consider is the amount of scaling required to calibration a set of model-based small area estimates. This measure is based on what is typically a key requirement for small area estimates - that they sum to direct estimates at appropriate levels of aggregation. We refer to this property as calibration. The basis for this requirement is simple. Large sample sizes at higher levels of aggregation mean that the direct estimates can be considered to accurate at these levels. Consequently, given two sets of model-based estimates, one that agrees with the direct estimates under appropriate aggregation and one that does not, we prefer the former. In practice, since model-based small area estimates are calibrated, usually by appropriate scaling, checking calibration after such scaling is irrelevant. However, by calculating the relative difference between the aggregated model-based estimates prior to this calibration and the aggregated direct estimates we obtain a measure of how accurate the aggregated model-based estimates are, and provides a means to compare different models.

An interesting issue to consider when using this diagnostic is deciding the calibration level. Since the aggregated direct estimates to which the aggregated model-based estimates are being compared are themselves subject to sampling variation, it is inappropriate to calibrate at too low a level of aggregation. It is important to identify this "cut-off" size when considering what calibration to perform.

Example

We continue with the model-based approach introduced earlier for the same five years. Our model-based estimates are required to be consistent with direct estimates at three margins: 6 National age-sex breakdowns; 12 Government Office Regions; 7 Socio-Economic classifications. We calculate how deviant the uncalibrated model-based estimates are from these margins, i.e. how much calibration is required to achieve consistency. The results are in Table 4, in terms of the percentage increases needed to model-based estimates in each margin. Although none of the percentages are major, the 5th category in the National age-sex margin consistently requires the largest amount of calibration - this category is women aged 50+, for whom the relationship between ILO unemployment and claimant count is known to be different from other age-sex categories. Overall the calibration required is increasing over time (as can be seen by counting the number of values over 1% per year). Clearly, future performance of the model will need to be monitored, although at present these percentage differences are not excessive.

	National age-sex	**Government Office Region**	**Socio-Economic Classification**
1995/1996	0.4 -0.6 0.3 0.3 1.8 -1.1	0.9 0.4 -0.1 1.1 1.2 -0.1 0.6 -0.1 0.7 0.3 0.2 0.3	0.3 -0.3 0.5 0.4 0.6 0.8 -0.5
1996/1997	-0.1 -0.4 0.5 0.3 2.2 -0.4	-0.8 0.5 -0.1 1.1 -0.5 1.4 1.5 0.8 -0.8 0.5 0.5 0.6	0.2 0.1 0.6 0.6 0.6 0.6 -0.3
1997/1998	-0.3 -1.2 0.0 0.6 2.8 0.8	0.7 0.5 -0.3 1.5 0.4 1.0 0.5 -0.8 0.0 1.3 0.2 0.0	0.6 0.3 0.0 -0.2 0.6 1.0 -2.1
1998/1999	-0.1 -0.4 0.7 1.1 2.2 -0.6	0.0 0.6 1.3 0.5 0.3 1.0 0.6 -0.2 1.6 1.2 0.1 0.0	1.3 1.0 1.2 -0.6 0.1 0.9 1.7
1999/2000	0.1 -0.9 0.7 0.8 3.9 1.6	1.8 1.1 1.2 0.7 0.0 0.7 0.0 1.0 0.9 2.1 -0.3 0.7	1.0 1.3 0.7 1.0 0.4 -0.2 2.3

Table 4 Percentage increases needed to model-based estimates, by margin, to achieve consistency with direct estimates

3. CONCLUDING REMARKS

In the previous section we have presented four diagnostics that we have found useful for both assessing the "fit" of a set of model-based small area estimates as well as comparing competing estimation methods (and models). However, there are a number of other diagnostics that are currently under development. The most relevant is a test of the robustness of the small area model to slight changes in the sample data. One approach is via cross-validation, splitting the sample data into smaller subsets, fitting the same model to each, and deriving a corresponding set of small area estimates. If the subsets are large enough to be representative of the population we would expect similar models to result and similar estimates to be obtained from each subset. The major problem here is deciding how to split the original sample data, and whether reweighting is appropriate. With unit level sample data from each of the small areas of interest, this should be reasonably straightforward. Unfortunately however, this is not always the case (e.g. the LFS example, where the data consist of direct estimates by age, sex and small area). We would welcome comments both on the methods we are currently using, and on ways in which we could add to our diagnostic repertoire.

REFERENCES

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. and Wang, S. (2001), "Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method", *Proceedings of the International Association of Survey Statisticians, Meeting of the International Statistical Institute, Seoul, August 2001*.

ONS (2001), "Small Area Estimation Project", unpublished report, London, U.K.: Office for National Statistics.