Small Domain Estimates Including Use of Administrative Data

Discussion: Tim Holt

The need for sound methodology for small area estimation continues to be of great importance. For official statisticians the demand for statistics for ever lower levels of geography is increasing. I congratulate the authors for three very interesting papers.

The primary focus of Bell and Jiang et al is to provide measures of the use of small area estimates rather than the estimates themselves. This is a very important problem. Bell works within the Fay-Herriott model that links the variable of interest to auxiliary variables at the small area level. Jiang et al provide results within the class of full mixed models that allow for multi-level models with random regression parameters. The paper by Attal-Toubert & Sautory is different. The emphasis is more directed towards the small area estimates than measures of MSE for these. This papers echoes the Fay-Herriott model in that the statistical model applies at the small area level. It focuses on the joint distribution of a vector of small area means, with no auxiliary variables and uses the Principle Components framework.

It is striking that Bell and Attal-Toubert et al provide empirical results at very high levels of aggregations (21 regions in France, 51 'States' in USA). Much larger direct sample sizes are available than occurs in many other practical applications.

Should we use the Fay-Herriott model or a full hierarchical model with auxiliary information at area and individual unit levels? Often there may be no choice because the availability of auxiliary information may only exist at the small area level. It is true also that small area auxiliary information can often contain good predictive power for the small area estimates. Nevertheless there are, in my view, benefits to using information in a full hierarchical model, not just for the small area estimates themselves but in terms of providing a basis for MSE estimation. My experience is that random coefficients for all auxiliary variables (and not just the intercept) have important effects on both estimation and MSE estimation. However there are many examples where auxiliary information is available only at the small area level, because it is derived from a different source for example. In these cases the full hierarchical model cannot be used.

Turning to the papers separately I congratulate Jiang et al on a very thorough presentation of the Jackknife approach to estimating MSE within a full hierarchical mixed model framework. I believe that the model needs a block diagonal structure for PSU's but apart from this it is very general. I would like to see empirical assessments of the robustness of the MSE estimates and, in particular, the extent to which these were affected by non-normal error distributions.

Bell provides MSE estimates under various approaches for the Fay-Herriott model and argues for Bayesian or other methods that lead to more plausible MSE estimates. I agree with him that the apparent reduction in MSE for many states is very large and implies that the predictive power of the auxiliary variables is very strong. This is because the small area components of variance are very small (zero in some cases). When the variance component estimate is very small (or zero) this leads to the apparently very large reductions in MSE for state estimates. This occurs because of the predictive power of the anxiliary information in the absence of a state level component of variance.

Bell's results confirm others in the literature that it is very important to include uncertainty not only from the prediction for each small area given the regression parameters and components of variance

but the uncertainty from estimating the regression parameters and components of variance too. His empirical results confirms others that, in particular, the uncertainty introduced by estimating the components of variance can be an important component of the MSE. Empirical results that capture this in some way are much more plausible.

Attal-Toubert and Sautory have no auxiliary data and instead use the vector of small area means within a principal components framework. The essential idea here is that the covariance matrix of the vector of unemployment rates by age and sex allow for a multivariate smoothing of the individual component estimates at the regional level. In some circumstances this may be all that can be done. Nevertheless I would be concerned that the estimates produced are too smooth and do not allow for sufficient variation between regions. Their results show very large reductions in variance for the regional estimates using their approach. Intuitively one would feel that if good auxiliary information was available it would help to preserve some of the regional differences in unemployment rates.