Danny Pfeffermann, Hebrew University, Jerusalem

The paper by Hidiroglou and Sarndal (HS) discusses several important aspects of small area estimation, a topic of great concern in recent years. The major problem in small area estimation is the small sample sizes realized within these areas. As a result, estimates of areas characteristics such as means and totals based on only the separate samples are often unstable and information has to be borrowed from other small areas to get more reliable estimates. Several such estimators, some of them new are discussed in the paper and their performance is compared by re-sampling from an actual data set.

Once appropriate estimators have been selected, the question arises as to how to assess their properties and use them for further inference. The authors focus on conditional inference, that is, calculating the bias, variance and confidence intervals with respect to the conditional distribution of the estimators, given realized values of certain sample statistics, in this case the achieved sample sizes within the small areas. I believe such an approach should be most welcome. Once the sample has been selected and the sample sizes within the small areas are known, it makes little sense to assess the errors of estimators taking into account all possible samples with all possible sample sizes for the small area samples. The papers by Holt and Smith (1979) and Rao (1985) referenced by HS elaborate on this issue. Distinction should be made, however, between pre-sampling and post-sampling inference. When planning the design for example, one would usually have to use unconditional distributions for the simple reason that the sample sizes within the areas are at that stage unknown. (This comment applies to situations where the within area sample sizes are not under the sampler's control as assumed by HS.)

Having decided on a conditional approach, the question still remaining is whether to use the randomization distribution or whether to use the model (superpopulation) distribution. The authors chose to work in the design framework which is less tied to model assumptions (a model is assumed for the construction of the estimators) and hence more robust. The use of the conditional approach in a design-based framework is a refreshing idea that will hopefully stimulate further research. Unfortunately, at the present state of art, the use of this approach is restricted in terms of the sample designs that permit use of a conditional approach and the sample statistics to condition on. Thus, it seems that as for now, the more one likes to condition on the more one has to employ the model distribution in the inference process.

My next comments are more specific to the estimators and confidence intervals discussed in the paper. The new estimator proposed by HS is $\hat{t}_{dDRE}$ (eq. 3.4) which modifies the estimator $\hat{t}_{dMRE}$ (eq. 3.3) by multiplying the correction term $c_d = \Sigma_{s_d} e_k/\Pi_k$ by a dampening factor in the range where the estimated area size $\hat{N}_d$ is smaller than the actual size N_d. The reason for using the dampening factor as presented by HS is that it controls the volatility of the correction term in situations where the area sample size is very small. However, following the authors recommendation of using h=2, the estimator $\hat{t}_{dDRE}$ can be written as

$$\hat{t}_{dDRE} = \Sigma_{U_d} \hat{Y}_k + a\, c_d \qquad a=\min\left[\frac{N_d}{\hat{N}_d}, \frac{\hat{N}_d}{N_d}\right]$$

This formula suggests that the two cases $N_d > \hat{N}_d$, $N_d < \hat{N}_d$ are actually treated in a quite symmetric way in the sense that whether N_d is two times as large as $\hat{N}_d$ or two times as small, in both cases the correction factor is multiplied by 1/2. Whereas the symmetry between the two cases could be examined using other criterions (e.g., looking at the absolute difference $|N_d - \hat{N}_d|$ which leads to different conclusions depending on whether N_d or $\hat{N}_d$ are held fixed) it is generally true that the correction factor c_d introduced originally to correct for a possible bias of the synthetic estimator $\Sigma_{U_d} \hat{Y}_k$, is actually never used to its full extent. This obviously reduces the variance of the estimator but on the other hand creates a bias and the trade off between the two components of the MSE in situations where the postulated regression model does not hold has to be further investigated.

As regards the computation of the confidence intervals, HS use as pivot the function $\tilde{Z}_d = (\hat{t}_d - t_d)/\sqrt{\hat{V}_c(\hat{t}_d)}$ postulating that $\hat{V}_c(\hat{t}_d)$ are consistent estimators of the conditional variance of $\hat{t}_d$ and hence that $\tilde{Z}_d$ is approximately N(0,1). However, the estimators they use for the variances are based essentially only on the small area samples (e.g., see eq. 6.15). If consistency of the variance estimators assumes

that the sample sizes within the small areas are allowed to increase then this clearly conflicts with the basic problem of small area estimation which is that the sample sizes within these areas are very small. Asymptotic analysis in small area estimation can be formulated in terms of increasing the number of areas and hence the overall sample size - not in terms of increasing the sample sizes within these areas. One is forced to conclude that in practice, since the sample sizes within the small areas are generally small, the estimators of the unknown variances should borrow information from other areas, similar to the borrowing of information for estimating means.