SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

DISCUSSION OF PAPERS BY DURBIN/QUENNEVILLE AND BELL

Avinash C. Singh[1]

It is my pleasure to discuss these two excellent papers. Both papers make important contributions in modelling survey data and in some sense complement each other. The first paper on benchmarking of time series by Durbin and Quenneville provides an elegant solution using state space models and applies the methodology to Statistics Canada's data series on retail trade. In finding the MSE of the resulting estimates, they however do not take into account the uncertainty in second-order parameter estimates. This happens to be a special case of a more general problem addressed in the second paper on modelling survey errors by Bell. He puts forward a Bayesian solution which is conceptually simple but somewhat complex computationally. A frequentist solution, on the other hand, may be even more complicated. Bell also illustrates the methodology with a time series example of 5+ unit housing starts data of the U.S. Census Bureau.

Specific comments are given below separately for each paper.

DURBIN AND QUENNEVILLE'S PAPER

The authors consider the problem of benchmarking a time series of survey estimates $\{y_t : 1 \le t \le T\}$ such that the estimates of signals $\{\eta_t\}$ satisfy benchmarks $\{x_s : 1 \le s \le S\}$. The signal at time t is the parameter of interest appropriately defined, e.g., it may correspond to the expectation of the seasonally adjusted y_t. There are two stages of the problem: (i) Initial signal extraction, i.e., finding $\tilde{\eta}_t$ suitably, and (ii) Adjusting $\{\tilde{\eta}_t\}$ to get $\{\hat{\eta}_t\}$ such that benchmarks are satisfied.

Given $\{\tilde{\eta}_t\}$, the benchmarking problem can be viewed simply as that of constrained regression. We have

$$\tilde{\eta} = \eta + \delta \ , \ \delta \sim (0, \Omega),$$
$$x = L\eta + e \ , \ e \sim (0, \Sigma_e),$$

where δ, e are uncorrelated. The optimal (in the sense of BLUE) $\hat{\eta}$ is given by

$$\hat{\eta} = \tilde{\eta} + \Omega L'(L\Omega L' + \Sigma_e)^{-1}(x - L\tilde{\eta})$$

This is equation (3.1) of the paper. If the benchmarks are nonrandom, then $\Sigma_e = 0$. If they are random, then for the binding case, $\hat{\eta}$ is obtained by setting Σ_e equal to 0. This gives rise to nonoptimal estimates. For the nonbinding case, Σ_e is not set to 0 and then the benchmark constraints are clearly not exactly satisfied. The amount of adjustment in $L\tilde{\eta}$ depends on the MSE $L\Omega L'$ relative to Σ_e.

In practice, the initial signal estimates $\{\tilde{\eta}_t\}$ are extracted using a model which may treat signal parameters either as random or nonrandom (such as X11-ARIMA for seasonal adjustment). When a model with nonrandom signals is used for $\{\tilde{\eta}_t\}$, the corresponding covariance matrix Ω may be (approximately) estimated using sampling design considerations. In this case, the benchmarking problem reduces to that of constrained regression. This was the approach taken by Cholette and Dagum (1994). They also showed that the traditional method of Denton (1971), a numerical procedure based on minimizing a distance function subject to benchmark constraints, can also be viewed as a problem of constrained regression by using an appropriate "working" covariance matrix Ω. Here, the distance is defined by $(\tilde{\eta} - \eta)'\Omega^{-1}(\tilde{\eta} - \eta)$. Such non-random signal-based methods for benchmarking are useful if the object is to perturb the already extracted signals $\{\tilde{\eta}_t\}$ only a little in order to satisfy the benchmarks. On the other hand, if the object is to extract the random unobserved components (trend, seasonal etc.) by fitting a model to $\{y_t\}$, then a model with random signal extraction with benchmarking would be appropriate. Here, the signals $\{\eta_t\}$ are random because they will be functions of the random unobserved components. Note that under this approach the benchmarked signal estimates would tend to look smoother in general.

For the random signal case Hillmer and Trabelsi (1987) used ARIMA modelling to extract $\{\tilde{\eta}_t\}$ and to specify the corresponding covariance structure Ω, while Durbin and Quenneville use structural modelling to

[1] Avinash C. Singh, Survey and Analysis Methods Development Section, HSMD, Statistics Canada, Ottawa, K1A 0T6, and Department of Mathematics and Statistics, Carleton University, Ottawa, K1S 5B6. This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

extract $\{\tilde{\eta}_t\}$ and to specify the corresponding Ω. Basic structural modelling seems to give them more flexibility. They allow for more general nonstationary time series, heteroscedasticity in survey errors and bias in survey estimates. They also consider the nonlinear case when the basic model is multiplicative in components while the benchmarks are additive in components. They consider two approaches to model fitting under the state space framework. First is the familiar two-stage approach which was also used by Hillmer-Trabelsi. Here, in the first stage, signal estimates $\tilde{\eta}$ and the corresponding covariance matrix are extracted without benchmarks, and in the second stage, benchmarked estimates $\hat{\eta}$ are obtained via constrained regression. The second is the proposed alternative approach termed single stage. Here a new time series is constructed by inserting the benchmarks at appropriate places and then this series is cast in a state space framework under the assumption that the benchmarks are independent. It is claimed that the single stage approach provides more efficient estimates of model parameters because full data (including the benchmarks) are used.

For estimating model parameters, the two-stage approach can be modified to make it equivalent to the single stage approach by viewing benchmark information as generating another time series $\{x_s\}$ which is augmented at the end of the original time series. Now the smoothed estimates $\tilde{\eta}$ obtained after the first stage (i.e., without benchmarks) can be used to create innovations (via orthogonalization) from the benchmark segment of the extended time series. This orthogonalization is done recursively like in the Kalman filter and, therefore, the likelihood can be extended to include benchmark information for parameter estimation. Note that the benchmarked estimates $\hat{\eta}$ will be automatically produced at the end of these two stages of Kalman filtering. This is a special case of the approach taken in Singh and Kovacevic (1995), who proposed a Segmented Kalman Filter (SKF) for benchmarking. In their approach, each segment of Kalman filtering is with respect to a suitable segment of the extended time series, the first segment being the usual Kalman filtering (and smoothing) for the original time series similar to the first stage of the two-stage approach. The SKF is especially useful for a multidimensional benchmarking problem involving multivariate (or a system of) series because the number of benchmarks may be quite large as they involve not only the usual (temporal) benchmarks for each component series, but also the (contemporaneous) benchmarks across components for each time. To do this with a two-stage approach may be computationally difficult because of the high dimension of matrix inversion involved in the second stage, the dimension being the number of benchmarks. However, if the benchmarks are uncorrelated, a recursive least squares can be used at the second stage to avoid matrix inversion. Also, with SKF, it is easy to revise later parts of the series in light of new benchmarks without revising the earlier parts, when the new benchmarks do not involve signals from earlier parts; this feature may be quite attractive in practice. As regards the single stage approach of Durbin and Quenneville, it is defined for a unidimensional benchmarking problem and its generalization to the multidimensional case seems computationally tedious.

The work of Durbin and Quenneville represents a considerable enhancement of the existing methodology. However, there are areas for further improvement. For example, when random benchmarks are treated as binding, then the MSE of the estimated state vector (and hence the signal) from the usual KF is not correct. The KF can be easily modified along the lines of Pfeffermann and Burck (1990) to correct for this situation. When dealing with possible bias in survey estimates, a constant bias is assumed which sometimes may not be reasonable in practice. Again the KF can be modified to allow for random bias evolving over time. Another area that is of concern but may not have a simple solution, is the use of approximate likelihood for estimating hyperparameters, for the nonlinear (or multiplicative) case; the approximation in the likelihood seems hard to justify. Finally, there is the problem of downward bias in MSE when variability due to estimation of second order parameters is not accounted for; see e.g. Singh, Stukel, and Pfeffermann (1995) for a review of available methods for non-time series problems. For time series problems, there exist methods due to Ansley and Kohn (1986) and Hamilton (1986). However, they do not give the desired order of correction; these methods can be improved and are currently being investigated in Singh and Quenneville (1996).

BELL'S PAPER

Bell makes a bold and commendable attempt to answer an often neglected question in the context of modelling survey data. This has to do with how to account for extra variability due to estimation of parameters in modelling survey errors. Although the general problem does not appear to have been considered before, an important special case is fairly well understood. This is the case of a (pure) model-based approach at the element level where the effect of the survey design on the elementary observations is modelled by incorporating available design variables. Here, it is well known that the MSE of estimates of first-order parameters are biased downward when estimates of second-order (or hyper) parameters are plugged in the corresponding MSE expressions. Such problems arise quite often in practice such as in small area estimation. There exist several methods to correct MSE; Singh, Stukel, and Pfeffermann (1995) provide a recent account and some new developments. The question raised by Bell is more difficult but realistic when dealing with survey estimates at the aggregate or domain level. As a matter of fact, in practice, it is generally preferable and relatively easier to account for survey design by modelling domain-level survey estimates. However, one then needs to deal also with uncertainty arising from estimating extra hyperparameters involved in modelling the survey error, i.e., in describing the covariance structure of survey errors. It may be noted that by addressing this concern due to estimation of hyperparameters, Bell's paper complements in a nice way the paper of Durbin and Quenneville.

Bell outlines a Bayesian solution which can be put in simple terms as follows. First assume that the prior information (such as the design-based estimate of the covariance matrix of survey estimates) used in survey error modelling is independent of the survey estimates and the estimated covariance matrix has a distribution with known form such as Wishart. Now, under a hierarchical Bayesian set-up for the population parameters, in which the prior information about the parameters of the survey error model can be easily incorporated, posterior variance would automatically take into account the extra uncertainty caused by not knowing the hyperparameters in the prior distribution. However, exact computation of the posterior mean and variance may be difficult in general. Since these involve integrals, the usual recourse is to use Monte Carlo Integration by simulating from the exact or the asymptotic posterior distribution. An alternative would be to simplify the integrand by the delta method under suitable asymptotics and then compute the integral with respect to the posterior distribution.

When information in the data about parameters of the survey error model is not independent of the survey estimates (this is the realistic situation), it may not be feasible in general to create a hierarchical framework for obtaining the posterior mean and variance. However, as suggested by Bell, a Bayesian solution can still be obtained by first computing the posterior mean and variance assuming that survey error model parameters are known, and then integrating them out using the prior distribution for these parameters. This should provide a reasonable solution although it won't be optimal. A frequentist solution, on the other hand, may be quite complicated. Based on the discussion in Singh, Stukel, and Pfeffermann (1995), it is, however, conjectured that under suitable asymptotics both Bayesian and frequentist approaches should give similar answers.

Bell provides a simple illustration of his proposed methodology using a real time series of data on 5+ unit housing starts. It seems that in order to convince practitioners about using the methodology, the main burden of proof would depend on developing model diagnostics, and then establishing robustness with respect to departures from "working" assumptions. I am sure this important contribution by Bell will generate interest among researchers and I look forward to further developments in this area.

ADDITIONAL REFERENCES

Ansley, C.F. and Kohn, R. (1986). "Prediction mean squared error for state space models with estimated parameters", *Biometrika*, 73, 467-473.

Denton, F.T. (1971). "Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization", *Journal of the American Statistical Association*, 53, 372-375.

Hamilton, J.D. (1986). "A standard error for the estimated state vector of a state-space model", *Journal of Econometrics*, 33, 387-397.

Pfeffermann, D. and Burck, L. (1990). "Robust small area estimation combining time series and cross-sectional data", *Survey Methodology*, 16, 217-237.

Singh, A.C. and Kovacevic, M. (1995). "Benchmarking of time series by segmented Kalman filtering", *Technical Report, Household Survey Methods Division*, Statistics Canada, Ottawa.

Singh, A.C. and Quenneville, B. (1996). "The estimation of mean squared error for state space models with estimated parameters". To be presented at Les XXVIII[e] Journées de statistiques, Québec City, Québec, May 27-30.

Singh, A.C., Stukel, D.M, and Pfeffermann, D. (1995). "Bayesian versus frequentist measures of error in small area estimation", *Technical Report No. 273, Laboratory for Research in Statistics and Probability*, Carleton University, Ottawa.