Discussant Comments
Concurrent Session 4B: Mortality Projections

W. Ward Kingkade

Presented at the Living to 100 Symposium
Orlando, Fla.
January 4–6, 2017

Shang and Haberman, "Grouped Multivariate and Functional Time Series Forecasting: An Application to Annuity Pricing"

W. WARD KINGKADE: This [Shang and Haberman] paper deals with a problem of immense practical significance: obtaining local-area forecasts and aggregating to higher levels of geography. In practice, local-area forecasts are often in high demand, and the authors make the point that they would be useful for pricing annuities. Those of us who deal with local data regularly encounter the problem that small areas often have small populations and, for this reason, highly volatile mortality statistics. For this reason, we often have to fall back onto data for larger geographic areas and force the local-area statistics into agreement; alternatively, we can let them disagree and leave it at that.

The basic contribution of the paper is the comparison of two methods of forecast reconciliation. The first is a bottom-up approach. In this approach, the series of age-specific death rates (ASDRs) are projected at the most disaggregated level by ARIMA models, and the resulting forecasts are summed up to higher levels of aggregation, using a summing matrix. The elements of the summing matrix indicate how the aggregated and disaggregated series relate. For the bottom-up method, this essentially reflects the distribution of the population. The bottom-up approach is easy to understand and is said to work well when the data have a high signal-to-noise ratio. This seems to be the case most of the time in the Japanese data the authors use, for years 1975–2013.

The bottom-up method is contrasted with an optimal combination method. With this method, all series of ASDRs at all levels of aggregation are independently forecast by ARIMA models.[1] Thereafter, some form of least squares is used to reconcile the aggregated and disaggregated series, minimizing "reconciliation errors." The optimal combination method is recommended for use when the signal-to-noise ratio in the data is low.

The authors find that in the Japanese case with which they are dealing, the bottom-up method generally performs better in rolling window tests of fit against a 15-year holdout sample. This is a commendable method of testing forecast goodness of fit. I also applaud the authors for using multiple principal components (up to five, it appears, in some instances). Finally, the authors present an actuarial example in which their approach is applied to annuity pricing, and I note that the bottom-up method is usually associated with slightly lower annuity prices than other methods to which it is compared in the Japanese case.

If I had to find a bone to pick with the authors, it would be their assumption that exposure to risk remains constant within age cohorts as age advances: $\mathbf{E}_{x+1,\,t+1} = \mathbf{E}_{x,\,t}$. The authors' methods require forecasting a summing matrix, **S**, over time, and this assumption allows them to forecast the **S** matrix for age 60, then apply it to all later ages. This assumption is convenient but seems not to be realistic in practice. I think it would strengthen the paper if there were some exploration of the

[1] The authors employ an automated procedure to fit their ARIMA models.

degree to which **S** matrices vary over age and time. If there is not much of a difference, the assumption can be defended as a practical expedient. If not, cohort survival projections may be a necessary accessory.

Hunt and Villegas, "Mortality Improvement Rates: Modeling and Parameter Uncertainty"

This [Hunt and Villegas] paper deals with an interesting problem—namely, how to model and forecast mortality dynamics in terms of improvement rates. The problem it faces is the volatility of improvement rates. Such rates involve differences over time, and when one differences a time series, the result is often more jagged or noisy than the base series. The authors offer an intriguing way of handling the problem. They reexpress the first difference in the logarithm of the ASDR from time t to $t + 1$ as an expression involving the logarithm of the age-specific mortality rate at time $t + 1$ on the lefthand side and the terms of a generalized linear model (GLM) for improvement rates on the righthand side.[2] In this manner, we obtain a model of improvement rates in terms of model-estimated central death rates, m_x, instead of differences in empirically observed rates.

Imposing the model structure this way leads to smoother series of results, especially with respect to age. However, this comes at a cost of having to live with parameter uncertainty. In the authors' Poisson GLM, uncertainty is inversely proportional to the square root of expected deaths. This can be taken into account by resampling methods such as the bootstrap, and that is what the authors do.

In the spirit of Lee and Carter, the authors use a vector autoregressive model in the first differences (VAR(1)) model to estimate the time-dependent parameter in their model. They also experiment with including a deterministic trend parameter in their model. Moreover, they employ ARIMA models for a cohort parameter that they include in the model.

The authors compare a substantial number of improvement rate models inspired by the existing models for forecasting death rates—e.g., the classic Lee-Carter formulation. They don't examine the goodness of fit of their forecasts, but they do examine parameter and forecast stability. They demonstrate that what they term the "crude" approach—namely, modeling observed first differences in death rates—leads to more instability than their "fitted" approach.

In evaluating their model, the authors encounter some problems of plausibility that often arise in modeling and forecasting rates of change over time in age-specific mortality rates. In particular, the authors find that when forecasting according to model improvement rates by age, they sometimes end up with death rates at younger ages that are higher than death rates at later ages when the forecast is carried far enough ahead in time. This can be seen in the fan charts in Figure

[2] The righthand side also includes the vector of logarithms of the last observed age-specific death rates, which is a commendable approach from the standpoint of forecasting, because it tends to reduce jump-off error.

8. It is clear that for the model with the deterministic trend, ln m_{40} exceeds ln m_{55} sometime after 2050. For the model without the deterministic trend, it appears that this will occur after 2060. I can make my own contribution to the list of problems: the phenomenon whereby improvement rates have different signs at different ages, which is sometimes encountered in practice for specific populations and time periods.

Such mortality crossovers seem to me to be the biggest problem with improvement-rate-based forecasts of mortality. Perhaps they can be handled by imposing a structure of constraints on the parameters of the model(s).

Finally, my recommendation—if one is needed—would be to encourage the authors to try some goodness of fit tests, perhaps with a small selection of their favorite models.

Avraam, Arnold, Vasieva and Vasiev, "On the Heterogeneity of Human Populations as Reflected by Mortality Dynamics"[3]

This [Avraam et al.] paper doesn't fit together well with the two others in the session. It contains no forecasts or projections of mortality. Rather, it fits a number of models of age-specific mortality to data for a few countries in the past and discusses them.

The authors advance a model that belongs to the category of unobserved heterogeneity models. In this kind of a model, one posits the existence of an unobserved factor or set of factors distributed in the population. In the literature on mortality (e.g., Vaupel and Yashin 1983), this is often framed as "frailty"; in the paper, it appears as "fitness genes." The unobserved factor or factors are put to use in explaining mortality patterns in the population. The classic example is the black-white crossover in U.S. mortality, where mortality rates for black males are higher than those of their white counterparts at young ages, but white mortality rates increase more rapidly with age than do black mortality rates, so that by late age, black males exhibit lower mortality rates than white males of the same ages. The heterogeneity explanation of this statistical curiosity is that higher mortality at young ages weeds out all but the hardiest individuals in the black population, who enjoy as a result lower mortality rates at later ages than their less heavily selected and therefore "frailer" white counterparts.

The problem with such explanations is that one can always be created to explain any mortality pattern. As long as the factors remain unobserved, it is hard to prove that heterogeneity is what actually accounts for these patterns. Other factors might just as easily explain the pattern. Without explicit measurement of the presence or absence of frailty, it is difficult to refute the arguments that differences in health-related behavior or age misreporting potentially related to socioeconomic conditions account for the U.S. black-white crossover, for instance. Analogously, alternatives to

[3] This paper was subsequently withdrawn.

unmeasured fitness genes can surely be advanced in explanation of mortality patterns in the countries the authors investigate.

The authors propose a model in which the population is broken down into four genotypes, two of which die out entirely before reaching reproductive age. They show that this model is flexible enough to explain even age patterns in which mortality rises and then declines as age advances. They proceed to fit this model to various countries.

While I'm quite prepared to concede that the four-subpopulation model might possibly explain a wide variety of mortality patterns, including those in Sweden and some other countries, I find it difficult to conclude without more direct and concrete evidence that it actually represents what is underlying the observed mortality patterns or trends. I don't know how easy it would be for the authors to introduce explicit measures of the fitness gene factors supposedly determining the mortality patterns and incorporate them directly into the model. However, I think the paper should move in the direction of being as concrete as possible. A useful step in that direction would be to present the estimated parameters of the model, which are not included in the versions of the paper I have seen.

Another aspect of the analysis which could be improved is the fitting procedure and related statistics. The authors fit death rates in the Human Mortality Database by nonlinear least squares using the Solver in Microsoft Excel. They present a version of the Bayesian information criterion (BIC) due to Priestly (1978), which arose in the field of time series analysis in a context in which it is assumed that estimation errors are independent and are normally distributed with a constant variance. These assumptions will not, in general, hold for age-specific death rates or central death rates. Death rates of either category are bounded, with a lower limit of zero and often an upper limit of one.[4] As a result, disturbances are, in general, not symmetrically distributed with equal variances as the rates approach the asymptote(s). In the case of age-specific death rates, the disturbances are apt to be related to age and hence each other. Problems of this nature are frequently handled by transforming the death rates, taking logarithms (which permits death rates to exceed unity) or logits (where the variable is bounded between zero and unity). There is a well-developed literature and body of statistical software available for such purposes. Using Excel is unconventional, and it would be worthwhile to check the results against those obtained from established statistical software. This is something the authors should certainly be able to do.

[4] It would make no difference if the rates were expressed as percentages, per thousands, etc. There would still be a lower and perhaps (depending on assumptions and computational procedures) an upper asymptote.