Journal of Forecasting, Vol. 11, 157–167 (1992)

Estimating the Strength of Expert Judgement: The Case of US Mortality Forecasts

JUHA M. ALHO
University of Illinois at Urbana-Champaign, IL, U.S.A.

ABSTRACT

The use of expert judgement is an important part of demographic forecasting. However, because judgement enters into the forecasting process in an informal way, it has been very difficult to assess its role relative to the analysis of past data. The use of targets in demographic forecasts permits us to embed the subjective forecasting process into a simple time-series regression model, in which expert judgement is incorporated via mixed estimation. The strength of expert judgement is defined, and estimated using the official forecasts of cause-specific mortality in the United States. We show that the weight given to judgement varies in an improbable manner by age. Overall, the weight given to judgement appears too high. An alternative approach to combining expert judgement and past data is suggested.

KEY WORDS Demography Mixed estimation Combination of forecasts

Expert judgement has been an integral part of demographic forecasting at least since the pioneering work of Pascal Whelpton. Whelpton (1928) noted that several mathematical models often fit past data well but yield widely differing forecasts. He suggested that one use all available data to formulate the most likely target values for the vital rates, and connect the current, or *jump-off*, values to the targets via a smooth mathematical curve (Thompson and Whelpton, 1933, pp. 312–13; Whelpton *et al.*, 1974, pp. 10, 29). Despite the subsequent development of 'data-driven' time-series techniques (e.g. Box and Jenkins, 1976), model choice remains largely judgemental (cf. Lee, 1947, p. 614), and depends, in part, on how reasonable are the forecasts it produces. Indeed, the US Office of the Actuary still uses the targeting approach in its demographic forecasts (e.g. Wade, 1987).

The role of judgement in the accuracy of forecasts is difficult to assess, however, because it enters into the forecasting process in an informal way. We have argued earlier (Alho, 1990) that the US Office of the Actuary might have produced more accurate forecasts of mortality during the past decades had it put less weight on expert judgement. Briefly, the analysis showed that simple trend extrapolations of the *standardized mortality rate* would have been more accurate than the actual forecasts, with the actual forecast and the trend extrapolation typically

0277–6693/92/020157–11$05.50

bracketing the realized value. A corresponding analysis of life expectancies produced less clear-cut results.

The targets of official mortality forecasts are formulated based on extensive data analysis of age-, sex-, and cause-specific mortality rates (Wade, 1987). The information contained in trend extrapolations is one ingredient. However, since the targets do not coincide with the extrapolated values, we can think of the difference as being attributable to other sources of information, the effect of which is not completely reflected in the past trend of the time series. Sources of such information include changes in lifestyle, environmental changes, and imminent medical advances. We refer to the use of such outside information as *expert judgement* (Alho and Spencer, 1980, p. 210).

We shall think of the forecast as a weighted average of simple, data-dependent, trend extrapolations, and of expert judgement. The purpose of this paper is to formalize these ideas within a statistical model that incorporates expert judgement, and to estimate the weights used by the US Office of the Actuary in its recent forecast (Wade, 1987). The embedding of actual forecasting practice into a formal statistical model provides at least two benefits. First, it enables us to determine which types of expert judgement are compatible with the actual forecasts and past data. In many instances, the compatible judgements appear not to be very attractive, suggesting that the forecast itself may not be as credible as one might hope. Second, a formalized representation provides a basis for reconsidering the way data analysis and expert judgement are used in official forecasts in the future.

It is emphasized from the outset that we do not attempt to specify the weights that would have yielded the most accurate forecasts in the past, because comparable annual age-, sex-, and cause-specific data for the USA are only available from the year 1968. Instead, we will study the strength and internal coherence of expert judgement. One would expect that neighboring age groups would receive smoothly varying weights within a cause, because the factors influencing cause-specific mortality (such as lifestyle or diet) would change gradually as a function of age. The fact that the weights sometimes vary in ways that cannot be easily justified suggests that more care is needed in the exercise of expert judgement in official forecasts. We also present further evidence that the weight given to expert judgement appears to be too high in many cases.

In the next section a simple time-series regression model that will be used to carry out the analysis will be defined. The concept *strength of expert opinion* will also be explained and a formula for its estimation derived. The connection between the mixed-prediction approach represented by the model, the corresponding Bayesian analysis, and the problem of combining forecasts will also be pointed out. The following section contains analysis of the mortality data. In particular, the coherence of expert opinion over age and cause is investigated. We will describe how the impact of expert judgement varies as a function of forecast year and consider factors influencing the weight given to judgement in practice. An alternative analysis that differs in the way data analysis is allowed to influence forecasts will be presented. The modified formulation mimics closely the way forecasts are currently viewed, but it has a peculiar, counter-intuitive property and, as such, may not provide as good a basis for future improvements as the first analysis. The Discussion concludes by suggesting how the official forecasting procedures could be made more coherent.

THE RELATIVE STRENGTHS OF DATA ANALYSIS AND EXPERT JUDGEMENT

Consider a time series $y(t)$, $t \in \mathbf{Z}$, of the form

$$y(t) = f(t) + \varepsilon(t)$$

where $E[\varepsilon(t)] = 0$. The mean, or the *trend*, $f(t)$ is of the form

$$f(t) = \beta_0 f_0(t) + \cdots + \beta_k f_k(t)$$

where $f_0, \ldots, f_k$ are known functions and $\beta_0, \ldots, \beta_k$ are unknown parameters to be estimated. Suppose we have observed $y(t)$ at $t = 1, \ldots, n$. Define $Y_0 = (y(1), \ldots, y(n))^{\mathrm{T}}$, $\varepsilon_0 = (\varepsilon(1), \ldots, \varepsilon(n))^{\mathrm{T}}$, and let X be an $n \times (k+1)$ matrix with the (i, j) element equal to $f_j(i)$. We assume that $n > k$ and that X is of full rank. We assume also that $\mathrm{Cov}(\varepsilon_0) = \sigma^2 \Sigma_0$, where Σ_0 is a known positive definite matrix. The usual GLS estimator of $f(n+m) = A^{\mathrm{T}}\beta$, where $A = (f_0(n+m), \ldots, f_k(n+m))^{\mathrm{T}}$, is $\hat{f}(n+m) = A^{\mathrm{T}}\hat{\beta}$ with $\hat{\beta} = (X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}X^{\mathrm{T}}\Sigma_0^{-1}Y_0$ and $\mathrm{Var}(\hat{f}(n+m)) \equiv \phi = A^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}A$. As is well known, the GLS estimator is the minimum variance linear unbiased estimator. Correspondingly, $\hat{y}(n+m) = \hat{f}(n+m) + \Sigma_{2,\mathrm{p}}^{\mathrm{T}}\Sigma_0^{-1}(Y_0 - X\hat{\beta})$ is the minimum MSE prediction of $y(n+m)$, if $\sigma^2\Sigma_{2,m} = \mathrm{Cov}(\varepsilon(n+m), \varepsilon_0)$ is known up to the constant σ^2 (cf. Rao, 1973, (v), p. 522).

Now we incorporate expert opinion into the prediction of $y(n+m)$ by using *mixed estimation* (Alho and Spencer, 1985; Theil, 1971). Suppose we have available a guess $y^* = f(n+m) + \varepsilon^* = A^{\mathrm{T}}\beta + \varepsilon^*$, where $E[\varepsilon^*] = 0$ and $\mathrm{Var}(\varepsilon^*) = \varkappa\sigma^2$. Assume y^* is *independent* of Y_0. Then we can simply augment the observation vector Y_0 by y^* and augment X and Σ_0 correspondingly. We call the resulting estimate a *mixed forecast*, and denote it by $\tilde{y}(n+m)$. Numerically, the mixed forecast can readily be calculated. However, its analytical form is rather complex (see the Appendix). Fortunately, in the case of $\Sigma_{2,m} = 0$ (which we will consider in our applications) the formulas simplify to

$$\tilde{y}(n+m) = [\varkappa\hat{f}(n+m) + \phi y^*]/(\varkappa + \phi) \tag{1}$$

Furthermore,

$$\mathrm{Var}(\tilde{y}(n+m)) = \varkappa\phi\sigma^2/(\varkappa + \phi) \tag{2}$$

Using these results we define the *strength of expert opinion at target year* as $\zeta \equiv \phi/(\varkappa + \phi)$, i.e. it is the relative weight given to y^* in the mixed forecast of $y(n+m)$. A more general definition that permits $\Sigma_{2,m} \neq 0$ is given below in equation (5). For now, we continue to assume that $\Sigma_{2,m} = 0$.

In our application we will take $y(t)$ to be the log-transform of age-, sex-, and cause-specific mortality rate in the USA for 1968–85 ($n = 18$). Models based on second-degree polynomials ($k = 2$; $f_j(t) = t^j$) will be fitted to the past data. The target year will be 2010 ($m = 25$). We will interpret the middle mortality targets of the US Office of the Actuary (Wade, 1987, pp. 9–10) as a mixed forecast $\tilde{y}(n+m)$. The high and low target values for mortality will be used in the following section to derive an estimate of $\mathrm{Var}(\tilde{y}(n+m))$. Equations (1) and (2) can be solved for the two unknowns y^* and $\varkappa$, with

$$y^* = \tilde{y}(n+m) - [\hat{f}(n+m) - \tilde{y}(n+m)]\,\varkappa/\sigma \tag{3}$$

where

$$\varkappa = \phi\ \mathrm{Var}(\tilde{y}(n+m))[\phi\sigma^2 - \mathrm{Var}(\tilde{y}(n+m))] \tag{4}$$

We see that $\varkappa$ depends on ϕ, σ^2, and $\mathrm{Var}(\tilde{y}(n+m))$, but not on $\tilde{y}(n+m)$ or $\hat{f}(n+m)$. The usual estimator of residual variance will be employed to estimate the unknown σ^2.

Note also that no solution for $\varkappa$ exists if $\phi\sigma^2 \leqslant \mathrm{Var}(\tilde{y}(n+m))$. In that case, the addition of y^* would not decrease the variance of the target value, indicating that expert judgement is exercised in a manner that is incompatible with the trend extrapolation model chosen or with past data. In principle, we can account for modeling error if we replace the time-series regression model by an *approximately linear model*, as in Alho and Spencer (1985, pp. 308–9).

This inflates ϕ by the square of modeling bias at the target year. If an adequate allowance for bias is made, then $\varkappa$ can always be made estimable. However, the data-based specification of a bound for modeling bias is a non-trivial task (*ibid.*, p. 311) that has not been attempted with the mortality data we use. Hence, we continue to assume that the time-series regression model is correctly specified.

So far, our definition of the strength of expert judgement covers only the case $\Sigma_{2,m} = 0$ at the target year. A more general definition can be given as follows. Consider forecast years $p \neq m$. Note that adding a new observation y^* mades $\mathrm{Var}(\tilde{y}(n+p))$ always smaller than $\mathrm{Var}(\hat{y}(n+p))$. Since both $\hat{y}(n+p)$ and $\tilde{y}(n+p)$ are unbiased predictors under our assumptions, we propose to measure the effect of y^* by the *strength of expert opinion* at year $n+p$:

$$\zeta(n+p) \equiv [\mathrm{Var}(\hat{y}(n+p)) - \mathrm{Var}(\tilde{y}(n+p))]/\mathrm{Var}(\hat{y}(n+p)) \tag{5}$$

This is the relative decrease in the variance of the predictor that is due to expert judgement. Note that equation (5) is applicable under arbitrary correlation structures.

Define $K = [f_0(n+p), \ldots, f_k(n+p)]^{\mathrm{T}}$. When $\mathrm{Cov}[\varepsilon(n+p), \varepsilon_0] = \sigma^2\Sigma_{2,p} = 0$, expression (5) reduces to

$$\zeta(n+p) \equiv [\mathrm{Var}(K^{\mathrm{T}}\hat{\beta}) - \mathrm{Var}(K^{\mathrm{T}}\tilde{\beta})]/\mathrm{Var}(K^{\mathrm{T}}\hat{\beta}) \tag{6}$$

where $\tilde{\beta}$ is the mixed estimator of β (see the Appendix). A simple calculation shows that $\zeta(n+m) = \zeta$, so we see that the definition given below equation (2) is a special case of equation (5). However, we cannot, in general, write $\tilde{y}(n+p)$ in the form of equation (1). The reason for this is that, as shown in the Appendix, y^* influences different components of $\tilde{\beta}$ with a different weight.

To compute $\zeta(n+p)$ given in equation (6), note that $\mathrm{Var}(K^{\mathrm{T}}\hat{\beta}) = K^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}K\sigma^2$. Write $\phi(K) = K^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}A$, for short. Then, $\phi(A) = \phi$. With some algebra, we can show that

$$\mathrm{Var}(K^{\mathrm{T}}\tilde{\beta}) = \mathrm{Var}(K^{\mathrm{T}}\hat{\beta}) - \phi(K)^2\sigma^2/(\varkappa + \phi) \tag{7}$$

Mixed estimation provides a computational framework for incorporating expert judgement into forecasting. However, for the purpose of choosing the strength of expert judgement in practice (a problem to be discussed below), it is useful to view the procedure as an example of combining forecasts (cf. Granger, 1989). In fact, Pankratz (1989, pp. 77–8) has recently used essentially a mixed estimation approach to the combination of forecasts, and Trabelsi and Hillmer (1989, Theorem 1, p. 355) have derived a general result on how to obtain a minimum MSE forecast based on two possibly correlated forecasts. What we call expert judgement can be viewed as an alternative forecast, so our results complement those mentioned above. (Strictly speaking, this interpretation holds only for the forecast year $t = n + m$, because for other years t there does not exist any forecast of $y(t)$ based on y^* alone.).

We conclude the presentation of the statistical model by commenting on the relationship between mixed estimation and a fully fledged Bayesian analysis. Suppose our *prior beliefs* satisfy the two conditions, $E[A^{\mathrm{T}}\beta] = y^*$ and $\mathrm{Var}(A^{\mathrm{T}}\beta) = \varkappa\sigma^2$; but nothing else is assumed known about β, such as the exact distributional form, or the first two moments of other linear combinations. Then, a straightforward calculation shows (details omitted) that the affine estimator of $A^{\mathrm{T}}\beta$ that minimizes the prior quadratic risk is given by the mixed prediction estimator. This result is related to Theorem 1 of Anandalingam and Lian Chen (1989, p. 203), who derive the corresponding result assuming a normal prior. The result of Anandalingam and Lian Chen requires the specification of a full prior distribution, not just the first and second moments, for the *whole* parameter vector. However, even the specification of the moments seems difficult in practice, since the functions f_j do not have any substantive interpretation by themselves.

One reason for taking y^* to be independent of Y_0 is to allow for this Bayesian interpretation, in which expert judgement puts an incompletely specified prior distribution on the parameters β. However, the substantive interpretation of the independence assumption needs some care. Much of what we might consider as outside information has already been reflected in the past of the mortality time series. In our model we will attribute all of that to trend extrapolation. Only the 'orthogonal' remainder that is truly independent of the trend forecast is considered to be expert judgement. Such a convention is necessary for the identifiability of expert judgement. Psychologically, it may lead to a situation in which a forecaster might not readily recognize what we call expert judgement as being his or her own expert input.

ANALYSIS OF US MORTALITY DATA FROM 1968 TO 1985

The US Office of the Actuary bases its forecasts on cause-specific mortality data. Based on the ninth revision of the International Classification of Diseases (ICD), the following ten classes are defined (ICD numbers in parentheses):

I. Diseases of the Heart (390–398, 402, 404–429);
II. Malignant Neoplasms (140–208);
III. Vascular Diseases (400–401, 403, 430–457, 582–583, 587);
IV. Accidents, Suicide, Homicide (E800–E989);
V. Diseases of the Respiratory System (460–519);
VI. Congenital Malformations and Diseases of Early Infancy (740–779);
VII. Diseases of the Digestive System (520–570, 572–579);
VIII. Diabetes Mellitus (250);
IX. Cirrhosis of the Liver (571);
X. All other causes.

Death rates are calculated by age (0, 1–4, 5–9,..., 90–94, 95 +) and sex and their *rate of decline* is determined by fitting a linear trend to the log-transformed rates. The negative of the slope is the *jump-off* rate of decline, which is connected to the *target* rate of decline by a smooth curve (cf. Alho and Spencer, 1990, p. 211).

We will use the 1968–85 data as a basis for analysis. Results for the first two causes will be used in illustrations. A recent official forecast (Wade, 1987) will be employed to calculate the high, middle, and low targets for the mortality rates. The calculation of the three targets is based on the jump-off values estimated from past data (Wade, 1987, Table 7, p. 9), the assumed ultimate annual percentage reductions (*ibid.*, p. 10), and the projection formula of Andrews and Beekman (1987, p. 21) with modifications as indicated in Wade (1987, p. 10). The projection formula causes initially slow change from the jump-off value. Then there is rapid change, followed by a smooth approach to a plateau at the target year. It was shown in Alho and Spencer (1990, pp. 214–15) that this nonlinear curve is well approximated by a *linear rate of decline model.* If the rate of decline is linear, then the decline itself is quadratic. This is the basis for assuming a second-degree polynomial for the trend function. Although there is some evidence that the regression residuals have positive autocorrelations (Alho, in press, Table 1), the scarcity of the data ($n = 18$) prevents a reliable analysis of the correlation structure. Therefore, $\Sigma_0 = I$ was assumed.

Since the official forecasts are produced using informal non-stochastic methods, there does not exist an official interpretation of what is the intended confidence level of the high–low interval. We will first interpret the official high–low target values for mortality in 2010 as a 90% prediction interval. Although subjective, this assumption is not an unreasonable one, as we have shown earlier that the official high–low intervals appear to be sometimes narrower, sometimes wider than the empirically estimated 90% prediction intervals (cf. Alho and Spencer, 1985, p. 313; Alho and Spencer, 1990).

Suppose that the width of the high–low interval for the logarithm of the mortality rate in the year 2010 has the width L. Then, assuming that a normal distribution can be used to represent our subjective uncertainty about the likelihood that the high–low interval will capture the target year mortality, we have that $L = 2 \cdot 1.65 \cdot \mathrm{Var}(y(n+m) - \tilde{y}(n+m))^{1/2}$. Alternatively, by the independence of $y(n+m)$ and $\tilde{y}(n+m)$ we have that $\mathrm{Var}(\tilde{y}(n+m)) = (L/3.3)^2 - \mathrm{Var}(y(n+m))$. Using equation (4) in the previous section, this yields an estimate of κ. Taking the middle forecast target as $\tilde{y}(n+m)$, we can deduce y^* from equation (3).

Note that if the target interval is interpreted as, say, a 67% or a 95% prediction interval, rather than as a 90% interval, then we should replace 1.65 above by 1 or 1.96. Hence, all values of $\mathrm{Var}(\tilde{y}(n+m))$ are adjusted by an additive constant. However, the corresponding κ- and ζ-values will be nonlinearly affected. Therefore, we present all three calculations to indicate the robustness of the results.

COHERENCE OF EXPERT OPINION BY AGE AND CAUSE

Table I contains three estimates of the strength of expert judgement in the official forecasts of male heart disease and cancer, the two dominant causes of death. The estimates correspond to three levels of coverage for the official high–low interval for the year 2010 (67%, 90%, and 95%). The results for females and for the remaining eight causes of death are qualitatively very similar.

As expected, the strength of expert opinion varies fairly smoothly from one age to the next. However, we note that the 67% assumption does not yield a solution for ages 45–54 and 60–64 for the heart disease data, and for ages 15–19, 65–69, and 75–79 for the cancer data. In all these cases mortality has evolved in an almost linear manner, leaving very small residuals from a time-series regression fit. Consequently, a purely empirical 67% prediction interval for the target year is narrower than the one assumed in the official forecast. This may be an indication that some amount of modeling bias has implicitly been recognized in the official high–low intervals.

A converse phenomenon occurs in certain young ages for heart disease (for example, ages 0–24 with the 95% assumption). In these cases the variance of the prediction error, $\mathrm{Var}[y(n+m) - \tilde{y}(n+m)]$, is *smaller* than the estimate of σ^2, or prediction error is assumed to be smaller than the observed lack of fit! In these cases the observed rates are quite low, so specification errors on the part of the forecasters may easily occur. As these cases represent *overconfidence* on the part of forecasters, they have been assigned the value $\zeta = 1.00^*$ in Table I, with the asterisk representing the fact that a formal calculation would actually have produced a value greater than 1.

The three alternative estimates all lead to a similar conclusion as to the coherence of the use of expert judgement. Frequently, the strength of expert judgement appears to be incompatible between ages (for example, that the existing information about changes in the diet of the US

Table I. Estimated strength of expert opinion at target year 2010 for US male heart disease and cancer mortality in the forecast of the US Office of the Actuary (Wade, 1987) corresponding to 67%, 90%, and 95% coverage probabilities

	Heart disease			Cancer		
Age	67%	90%	95%	67%	90%	95%
0	1.00	1.00*	1.00*	0.96	0.99	1.00
1– 4	1.00	1.00*	1.00*	0.77	0.92	0.95
5– 9	1.00*	1.00*	1.00*	0.83	0.94	0.96
10–14	1.00*	1.00*	1.00*	0.76	0.92	0.94
15–19	1.00	1.00*	1.00*	—	0.59	0.71
20–24	0.98	1.00	1.00*	0.82	0.94	0.96
25–29	0.88	0.96	0.97	0.68	0.89	0.92
30–34	0.42	0.79	0.86	0.71	0.90	0.93
35–39	0.26	0.73	0.81	0.64	0.87	0.91
40–44	0.24	0.73	0.81	0.66	0.88	0.92
45–49	—	0.54	0.68	0.84	0.95	0.97
50–54	—	0.04	0.32	0.93	0.98	0.99
55–59	0.08	0.67	0.77	0.89	0.97	0.98
60–64	—	0.55	0.68	0.82	0.94	0.96
65–69	—	0.41	0.58	—	0.61	0.72
70–74	0.01	0.64	0.75	0.16	0.70	0.79
75–79	0.27	0.74	0.81	—	0.47	0.63
80–84	0.45	0.80	0.86	0.19	0.71	0.79
85–89	0.56	0.85	0.89	0.33	0.76	0.83
90–94	0.69	0.89	0.93	0.85	0.95	0.97

males—e.g. reduction in fat and cholesterol intake—is a good reason for predicting an accelerating decline in heart disease mortality in older age groups). However, there does not seem to exist a similar basis for assuming that the current increase in the incidence of heart disease in the very young (ages 0–4) would become a decline. Still, the weight given to the latter judgement far exceeds the weight given to the former.

The overall conclusion is that for those ages for which the fit of the quadratic model is very good (small residual variance), the actual prediction variance at the target year $\text{Var}[y(n+m) - \bar{y}(n+m)]$ is often roughly the same as for ages for which the fit has been bad. The only way this can be accomplished is to exercise much stronger judgement for the latter ages. This does not appear reasonable because, as noted above, the outside information might actually be stronger for the former ages. If past data and expert judgement were combined in a statistically rational manner, the estimated uncertainty (as reflected in the residual error from a fit of a time-series regression model) should be better reflected in $\text{Var}(\bar{y}(n+m))$. Other things being equal, this would result in wider prediction intervals at the target year for those ages for which the past fit has been bad than for those ages for which the past fit has been good.

FACTORS INFLUENCING THE CHOICE OF THE STRENGTH OF EXPERT OPINION

To gain insight into the way one might choose κ in practice, let us consider its interpretation in more detail. Table I shows that the estimated strength of expert judgement at the target year

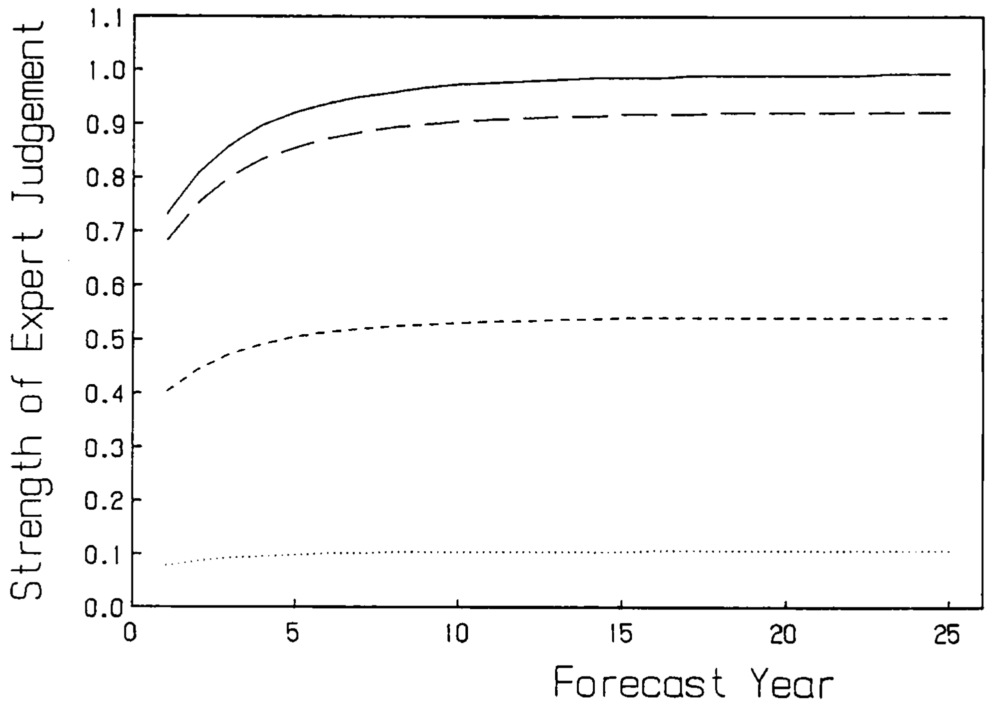

Figure 1. Strength of expert judgement as a function of the forecast year, for $\varkappa = 1$ (solid line), $\varkappa = 10$ (long dashed), $\varkappa = 100$ (short dashed), and $\varkappa = 1000$ (dotted)

is, for many ages, close to one under all three models. Alternatively, for these ages the forecast at the target year is essentially dictated by the (subjective) expert judgement. These results have been extended to the intermediate forecast years using $\zeta(n+p)$'s with $K = (1, n+p, (n+p)^2)^T$ for $p = 1, \ldots, m$.

Figure 1 is a graph of the strength of expert opinion for $\varkappa = 1, 10, 100, 1000$ during the 25 forecast years, calculated using equations (6) and (7). The curves are equal, up to a multiplicative factor $1/(\varkappa + \phi)$. The multiplier determines the starting level of the curve and the height of the final plateau. Irrespective of these values, the final plateau is rapidly reached by the middle of the forecast period. In our empirical material the range of $\varkappa$ values for male heart disease was from $\varkappa = 0$ in young ages to $\varkappa = 3153$ at age 55–59. Therefore the range was even wider than that considered in Figure 1.

Since y^* and $\varkappa$ are intended to represent information that is extraneous to the past of the time series in question, there is no simple 'reasonable range' of values for them. One way to think about the problem is to assume that mixed estimation is used to average two formal statistical forecasts. Therefore suppose one would be able to quantify all extraneous information we think we possess that is not already reflected in the past of the time series itself, and base a predictive model for the target year on such data. How large would be the prediction error of such a forecast? For example, if we would not be willing to assume that the forecast based exclusively on the (independent) extraneous information would be more accurate than that based on the past of the time series itself, we should not have $\varkappa$ smaller than ϕ, or the strength should not be over 0.5, a value frequently exceeded in Table I. In this author's view, both the empirical comparisons of trend extrapolations versus judgemental forecasts mentioned in the introduction, and the lack of validated substantive theories of demographic change, suggest that the strength of expert judgement should not exceed this bound in practice.

AN ALTERNATIVE ANALYSIS

So far, we have assumed that the second-degree model describes both the observation period (1968–85) and the forecast period up to the target year (1986–2010). In particular, we have let the past data influence all coefficients of the polynomial model. Although parsimonious from the statistical modeling point of view, this does not represent exactly the current practice in official forecasting. In fact, the US Office of the Actuary uses a linear model $f(t) = \beta_0 + \beta_1 t$ for the logarithm of the observed mortality rates, $t = 1, \ldots, n$. In other words, a constant rate of decline is assumed. A model that is well approximated by a second-degree polynomial is assumed during the forecast period, say, $f(t) = \beta_0 + \beta_1 t + \beta_2 (t-n)^2$ for $t = n+1, \ldots, n+m$. In this formulation past data have no effect on β_2. It is completely determined by the expert judgement y^*.

We note that the definition for the strength of expert opinion, as given above, is not directly applicable here, because the $n \times 3$ matrix X is of rank 2, therefore ϕ, for example, is undefined. Even more importantly, no purely data-based estimator exists for $f(t) = \beta_0 + \beta_1 t + \beta_2 (t-n)^2$ for $t = n+1, \ldots, n+m$. However, letting $y^* = \tilde{y}(n+m) = \hat{\beta}_0 + \hat{\beta}_1 (n+m) + \beta_2 m^2$, we can solve $\tilde{\beta}_2 = [y^* - \hat{\beta}_0 - \hat{\beta}_1 (n+m)]/m^2$. This shows that y^* completely dictates the point forecast at the target year, so the strength of expert opinion at the target year can be defined as 1. Furthermore, in this case, $\mathrm{Var}(y^*) = \varkappa\sigma^2 = \mathrm{Var}(\tilde{y}(n+m))$. For the earlier forecast years $t = n+1, \ldots, n+m$, we have $\tilde{y}(t) = y^*(t-n)^2/m^2 + \hat{\beta}_0[1 - (t-n)^2/m^2] + \hat{\beta}_1[t - (n+m)(t-n)^2/m^2]$, where the first term depends on y^* and the last two depend on past data. Note that $\tilde{y}(t)$ reduces to y^* at $t = n+m$. It follows that we might define the strength of expert judgement at $t = n+1, \ldots, n+m$, as $\mathrm{Var}(y^*(t-n)^2/m^2)/\mathrm{Var}(\tilde{y}(t))$, because this is the fraction of variance explained by the subjective input. A computational formula is obtained by noting that $\mathrm{Var}(y^*(t-n)^2/m^2) = \varkappa\sigma^2(t-n)^4/m^4$, and $\mathrm{Var}(\tilde{y}(t)) = \varkappa\sigma^2(t-n)^4/m^4 + \sigma^2 L^{\mathrm{T}}(U^{\mathrm{T}}\Sigma_0^{-1}U)^{-1}L$, where $L = [1 - (t-n)^2/m^2,\ t - (n+m)(t-n)^2/m^2]^{\mathrm{T}}$ and U is an $n \times 2$ matrix with the (i, j) element equal to i^{j-1}.

Under this model the coherence of expert judgement cannot be studied. The dependence of the strength of expert opinion on forecast year can, nevertheless, be evaluated. The strength of expert opinion starts from zero at $t = n$. There is an initial plateau, then a rapid rise, and a plateau at the end again, so that the strength is 1 at $t = n+m$. The position of the rapid rise depends on $\varkappa$. Since the variance 'explained' by expert judgement is directly proportional to $\varkappa$, the larger the value of $\varkappa$, the stronger the influence of expert judgement. Note the contrast with the role of $\varkappa$ in the previous model. This paradoxical property is one more reason for considering alternatives to the current forecasting practice.

DISCUSSION

In this paper we developed a formal statistical model, within which it is possible to define and estimate the strength of expert opinion in forecasting. The model was applied to US mortality data. The method is based on the so-called mixed estimation approach in linear regression, so it is applicable more generally, whenever the regression models themselves are. A modification of the model that mimics closely the subjective forecasting practices of official forecasts was also considered.

We saw that for many age groups expert judgement received a weight that appears either too high or too low relative to other ages. This was related to the residual variance of the time series in an undesirable way. We argued also that *too much weight* appears to be put on expert opinion overall. Clear instances of this were the cases in which prediction error was assumed to be less than the empirically observed lack of fit from a polynomial regression. In addition, there were many cases in which the strength of expert opinion was close to one, even though we do not appear to have enough independent information which would not already have been reflected in the past of the time series to warrant such a high weight.

The formal models we have developed suggest an alternative way of incorporating expert judgement into forecasting. It is possible that there are substantive demographic reasons for feeling more confident about the future trends of mortality for some ages than for others. Whether or not this is the case, we can let our prior information dictate the appropriate values for ζ, as long as we keep in mind that there are practical limits to what we know independently of the past of the time series. These will imply values for $\varkappa$. Together with the estimated value of σ^2, they will determine $\mathrm{Var}(\tilde{y}(n+m))$ and the widths of all prediction intervals. We can still center the intervals as desired by using y^*. The resulting prediction intervals would reflect accurately both the strength of expert judgement and the empirically estimated level of uncertainty in past data.

APPENDIX

Consider the prediction of $y(n+p)$. Define $K = [f_0(n+p), \ldots, f_k(n+p)]^{\mathrm{T}}$, so $y(n+p) = K^{\mathrm{T}}\beta + \varepsilon(n+p)$. Let $\mathbf{Y} = (Y_0^{\mathrm{T}}, y^*)^{\mathrm{T}}$, $\mathbf{X} = (X^{\mathrm{T}}, A)^{\mathrm{T}}$, $\mathbf{\Sigma}_{2,p}^{\mathrm{T}} = (\Sigma_{2,p}^{\mathrm{T}}, 0)^{\mathrm{T}}$, and $\mathbf{\Sigma} = \mathrm{diag}(\Sigma_0, \varkappa)$. It follows that the mixed forecast of $y(n+p)$ is

$$\begin{aligned}\tilde{y}(n+m) &= K^{\mathrm{T}}\tilde{\beta} + \mathbf{\Sigma}_{2,p}^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \\ &= K^{\mathrm{T}}\tilde{\beta} + \Sigma_{2,p}^{\mathrm{T}}\Sigma_0^{-1}(Y - X\tilde{\beta})\end{aligned}$$

where $\tilde{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Y}$. Recall that $\phi = A^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}A$, $\hat{\beta} = (X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}$ $X^{\mathrm{T}}\Sigma_0^{-1}Y_0$, and define $\beta^* = (X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}A\varkappa^{-1}y^*$. A direct calculation similar to that in Vinod and Ullah (1981, p. 71) shows that

$$\tilde{\beta} = M(\hat{\beta} + \beta^*)$$

where

$$M = I - (X^{\mathrm{T}}\Sigma_0^{-1}X)^{-1}AA^{\mathrm{T}}/(\varkappa + \phi)$$

At the target year $p = m$ we have $K = A$. If, in addition, $\Sigma_{2,m} = 0$, then $\tilde{y}(n+m) = K^{\mathrm{T}}\tilde{\beta}$. Since $A^{\mathrm{T}}M = \varkappa A^{\mathrm{T}}/(\varkappa + \phi)$, we get equation (1). Equation (2) follows immediately.

From the formula of $\tilde{\beta}$ we see that y^* influences different components of β differently, so, in general, equation (5) cannot be reduced to the simple weighted average form of equation (1), even when $\Sigma_{2,p} = 0$.

ACKNOWLEDGEMENTS

The author would like to thank Ms Wen-Ling Peng for help in computing and Bruce Spencer for comments on an early version of the manuscript. Detailed comments by an anonymous referee helped greatly in improving the presentation. The support of this work by the DHHS Grant 89ASPE220A is gratefully acknowledged.

REFERENCES

Alho, J., 'Stochastic methods in population forecasting', *International Journal of Forecasting*, **6** (1990), 521–530.

Alho, J., 'Effect of aggregation on the estimation of trend in mortality', *Mathematical Population Studies*, in press.

Alho, J. and Spencer, B. D., 'Uncertain population of forecasting', *Journal of the American Statistical Association*, **80** (1985), 306–14.

Alho, J. and Spencer, B. D., 'Effects of targets and aggregation on the propagation of error in mortality forecasts', *Mathematical Population Studies*, **2** (1990), 209–27.

Anandalingam, G. and Lian Chen, 'Linear combinations of forecasts: a general Bayesian model', *Journal of Forecasting*, **8** (1989), 199–214.

Andrews, G. H. and Beekman, J. A., *Actuarial Projections for the Old-Age, Survivors, and Disability Insurance Program of Social Security in the United States of America*, Itasca, IL: Actuarial Education and Research Fund, 1987.

Box, G. E. P. and Jenkins, G. M., *Time-Series Analysis. Forecasting and Control*, San Francisco: Holden-Day, 1976.

Granger, C. W. J., 'Invited review: combining forecasts—twenty years later', *Journal of Forecasting*, **8** (1989), 167–73.

Lee, R. D., 'Forecasting births in post-transition populations: stochastic renewal with serially correlated fertility', *Journal of the American Statistical Association*, **69** (1974), 607–17.

Pankratz, A., 'Time series forecasts and extra-model information', *Journal of Forecasting*, **8** (1989), 75–83.

Rao, C. R., *Linear Statistical Inference and its Applications*, 2nd edn, New York: John WIley, 1973.

Thompson, W. S. and Whelpton, P. K., *Population Trends in the United States*, New York: McGraw-Hill, 1933.

Vinod, H. D. and Ullah, A., *Recent Advances in Regression Methods*, New York: Marcel Dekker, 1981.

Wade, A., *Social Security Area Population Projections: 1987*, Actuarial Study No. 99, Washington DC: Office of the Actuary, 1987.

Whelpton, P. K., 'Population of the United States, 1925 to 1975', *American Journal of Sociology*, **34** (1928), 253–70.

Whelpton, P. K., 'An empirical method of calculating future population', *Journal of the American Statistical Association*, **31** (1936), 457–73.

Whelpton, P. K., Eldridge, H. T. and Siegel, J. S., *Forecasts of the Population of the United States*, US Bureau of the Census, Washington, DC: US Government Printing Office, 1947.

Author's biography:
Juha M. Alho is Assistant Professor of Environmental Biostatistics and Statistics at the University of Illinois at Urbana-Champaign. His recent research interests include the analysis of uncertainty in demographic forecasts and the uses of logistic regression in demographic estimation and sampling. He has publications, for example, in *Biometrics*, *Biometrika*, *Demography*, and the *Journal of the American Statistical Association*.

Author's address:
Juha M. Alho, Institute for Environmental Studies and Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Drive, Urbana, IL 61801, USA.

From August 1991: Professor Juha M. Alho, Department of Statistics, University of Joensuu, P.O. Box 111, SF-80101 Joensuu, Finland.