Forecasting Age Distribution Curves

Tucker McElroy and William Bell
U.S. Census Bureau

Disclaimer This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

The construction of population projections typically involves forecasting of several demographic components of change—births, deaths, immigration, and emigration—each with considerable demographic detail (e.g., age-race-sex for single years of age and several race/origin groups). This presents a forecasting problem of high dimension. Along with producing forecasts with this level of detail it is generally desired that forecasts reflect an age pattern consistent with age patterns observed in historical data, which typically are quite pronounced. Also, these properties are sought in long-term forecasts (say, 50 to 75 years ahead) though based on historical time series that may be of relatively moderate length, often less than the desired forecast horizon. This setting presents some challenging methodological problems for forecasting.

In this paper we investigate the application of time series methods to the forecasting of data on legal immigration to the U.S. The available data consist of historical estimates of several categories of legal immigration from 1972 to 2002. We focus, for illustration, on the category of "Hispanic employment immigrants," which denotes legal immigrants and their dependents arriving from countries of predominantly Hispanic race/origin, who have immigrated for reasons of employment. The historical data provide estimates for each category of the number of immigrants (aggregated as males plus females) for single years of age from 0 to 99 and then 100+. The same level of detail is desired in the forecasts. This data is easily transformed into a suite (or time series) of 31 age distributions simply by dividing the number of persons of a given age in a given year by the total number of persons in that particular year. If these age distributions can be forecasted ahead, they can then be multiplied by forecasts of total Hispanic employment immigrants to obtain forecasts of immigrants by age. Forecasts are desired up to 50 years ahead.

As noted above, the creation of such long-term forecasts from a small data set presents many challenges. Clearly, any forecasts should themselves be distributions, i.e., for each forecast year, the forecasted values at each age should be non-negative and should sum to one over ages. In addition, one wishes to model and forecast the main features of the distribution (such as the modes) while ignoring small perturbations that appear to be just "noise" in the historical data. This amounts to selecting a parsimonious approximation of the age curves that can be readily forecasted, while remaining true to the data's structure. Finally, since the forecast horizon is so long, it is desirable to impose constraints on the forecast methodology to outlaw implausible distortions in the ultimate forecasts of the age-distributions. These considerations have led us to consider a combination of logistic transformation, principal com-

ponents analysis, time series modelling, cubic spline smoothing, and Bayesian forecast attenuation. The combination of these statistical methods represents an innovative approach to achieving our forecasting objectives.

This paper focuses on the above methodologies and their application to the Hispanic employment immigrants data. Section 2 discusses the employment immigrant data that provides the basis for our forecasts, and Section 3 discusses the details of our methods. Section 4 discusses the application of these methods to the forecasting of Hispanic employment immigrants. Section 5 gives a summary of the approach, and an appendix contains a proof of a mathematical result used in Section 3.

Readers are cautioned that the focus in this paper is on investigation of methods for dealing with the forecasting problem, with the Hispanic employment immigrants data used for illustration. The actual forecast results are, at this stage, experimental, and should not be regarded as any sort of official Census Bureau projections.

2 Historical Data on Legal Immigration

The first step in forecasting legal immigration is to develop corresponding historical data. This task was undertaken by Hollmann (2004), who developed historical estimates of legal immigrants by age for four general types of immigrants (family, employment, refugee, and special) for each of four race/ethnic groups, nominally labelled Hispanic, Black, White, and Asian (the last three groups all referring to the non-Hispanic population). The four race/ethnic groups do not actually correspond to reported races of individual immigrants (as such detailed information is not available), but rather to a classification of countries of origin into four groups according to the predominant race/ethnicity that immigrants from these countries reported in the 2000 census. We focus here on forecasting the time series of "Hispanic employment immigrants," which means immigrants from countries of predominantly Hispanic race/ethnicity *and their dependents* by virtue of employment provisions in immigration law. (The inclusion of dependents of the actual employment immigrants means that the age range of the employment immigrants as estimated by Hollmann range from 0 to 100 and above.) Essentially the same considerations apply to modelling and forecasting time series of employment immigrants for the other three race/ethnic groups, and similar considerations apply to forecasting the "family immigrants" series as well. (Family immigrants are family members of previous immigrants, admitted via the family provisions of immigration law.) The refugee and special immigrant data are somewhat different, showing more erratic behavior.

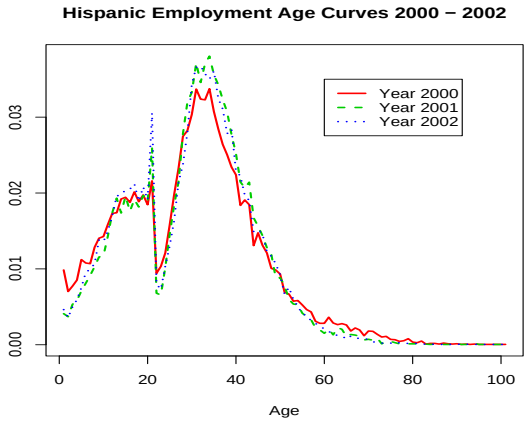
The data sources available for constructing estimates of legal immigration have some limitations that translate into errors in the historical estimates. Particularly worth noting here is the fact that a person does not become an "immigrant" until he or she has achieved legal permanent residence status in the U.S., and when this occurs the person is deemed to have immigrated as of the date of their last entry to the U.S. This "date of last entry" is often several years before the achievement of permanent residence status, implying that at any time there is a substantial pool of people already in the U.S. who will achieve immigrant status in the coming years and then be assigned as immigrants to the current year or to some past year (whenever they last entered the U.S.). To account for this pool of potential immigrants Hollmann (2004) was thus forced to impute substantial numbers of immigrants into the last years of his historical estimates. Such a process unavoidably incurs errors, and since these errors increase towards the end of the series, these errors have important implications for forecasting. In particular, although we shall forecast the series taking the data "as is," we must keep in mind that capturing every

detail of the age-pattern of the historical data, and reflecting this in the forecasts, is not necessarily justified. Some irregularities in the data, particularly in the last years, may be due to data errors.

3 Methodology

The forecasting problem at hand poses four main challenges: preserving the distribution property, dealing with the high dimensionality of the data, projecting "signal" while dispensing with noise, and attenuating the excesses of forecasts over an extended forecast horizon. One approach to handling the first three issues at once is to fit a parametric family of distributions to the data, such as a gamma curve, and then forecast the curve parameters. Rogers and Castro (1981) describe use of this approach with migration data. Bell (1997) discusses relative advantages and disadvantages of this approach, in general. The primary disadvantage is the difficulty in finding a curve that depends on a small number of parameters yet provides an adequate approximation to the data. For the Hispanic employment immigrants series, and for the employment immigrant series of the other three race/ethnic groups, the data are bimodal (see Figure 1 below) with some irregularities that can be regarded as "noise," but with others that represent real features of the data that should be preserved in forecasting. This means that finding a simple parametric curve to fit to this data would be difficult, and an adequate approximation would probably require piecing together two or more curves resulting in a moderate set of parameter series to forecast. We use other techniques to address the four problems noted. To preserve the distribution property, Section 3.1 discusses how we use a generalization of the logistic transformation proposed by Aitchison (1987). To deal with the high dimensionality of the data we use the principal components approach (PCA) proposed by Bozik and Bell (1987) – see also Bell 1997, which provides an excellent approximation to the data in terms of very few "parameters." In fact, as discussed

Figure 1: Age distribution curves for Hispanic Employment, 2000 – 2002



in Section 3.2, we use an approximation based on one principal component for the mean corrected data, a version of the approach used by Ronald Lee and his colleagues (e.g., Lee and Carter 1992). The one principal component approximation succeeds in capturing the features of the data used here that appear to be real, at least to the extent that appears important for forecasting. As noted in Section 3.3, however, the approximations to the age distributions include some irregularities that are exacerbated by long-term forecasting. To remove these irregularities we smooth the principal component vector and the mean age distribution curve using cubic splines (Section 3.4). Another problem with long-term forecasts noted in Section 3.3 is that historical trends in these data, extrapolated indefinitely into the future, yield implausible results. Section 3.5 discusses how, with a nominally Bayesian approach, we bring in "prior information" about the plausible range for future values of the series to "attenuate" the point forecasts to prevent implausible results.

3.1 Logistic Transformations

We require a method that ensures that our forecasted distributions will also be distributions, i.e., curves

that are non-negative with unit integral. A generalization of the basic logistic transformation to a multivariate context achieves this objective. Let the data be given as x_{it} for ages $i = 0, \ldots, 99, 100$ (the last age actually being 100+), and let $t = 1, \ldots, 31$ represent the years 1972 to 2002. For each year, we then obtain ratios

$$r_{it} = \frac{x_{it}}{\sum_{j=0}^{100} x_{jt}} \qquad i = 0, \ldots, 99, 100,$$

so that for each t, r_{it} defines an age distribution. Whatever forecasting methods we employ, we wish to ensure that the result is also an age distribution. The generalized logistic transformation (Aitchison 1987), when inverted, will constrain the forecasts to be positive and sum to one across ages. For each t, let

$$\gamma_{it} = \log\left(\frac{r_{it}}{r_{100,t}}\right) \qquad i = 0, \ldots, 99,$$

which defines a 100 by 31 data matrix. This transformation is reversed via

$$\begin{aligned} r_{it} &= \frac{e^{\gamma_{it}}}{(1 + \sum_{j=0}^{99} e^{\gamma_{it}})} \quad i = 0, \ldots, 99 \\ r_{100t} &= \frac{1}{(1 + \sum_{j=0}^{99} e^{\gamma_{it}})}. \end{aligned}$$

The transformed data γ_{it} can take on any real number value, so there is no constraint on the forecasts of the γ_{it}, but the inverse transformation guarantees that we obtain an age distribution for each time t. One proviso is that $r_{it} > 0$ is necessary for the transformation. There are actually many zeros in the historical Hispanic employment immigrant data x_{it}. Since the counts x_{it} tend to be fairly large when nonzero, however, we modified the data by adding one to each x_{it}.

3.2 Principal Components Analysis

The multivariate series γ_{it} has high dimension (100) but is relatively short (31 years). To simplify matters we reduce the dimension of the forecasting problem by using the principal components analysis (PCA) approach. The general approach was proposed in Bozik and Bell (1987) and is discussed further by Bell (1997). Lee and Carter (1992) used a version of PCA based on a one principle component approximation to mean corrected data of log U.S. mortality rates. Here we also use a one principle component approximation to the Hispanic employment immigrant data with mean correction, where the data are the logistically transformed ratios γ_{it}. Irregularities in the migration data (note Figure 1), coupled with our knowledge that errors in the historical data grow towards the end of the series, suggest use of a low-dimensional PCA approximation. Also, the need to forecast age distributions for each of 16 groups (four immigrant categories for each of four race/ethnic groups), along with the total number of immigrants for each of these groups (which we don't consider here), mandates that we attempt to minimize the number of principle components used. Bell (1997) notes that mean correcting the data, or alternatively subtracting from the data each year the values from the last year of data, is helpful when using a low-dimensional PCA approximation.

Let $\gamma_t = (\gamma_{0t}, \ldots, \gamma_{99t})'$ be the column vector of the transformed r_{it} for year t. We subtract a baseline curve defined as a single summary measure of all the curves; this could be the mean curve (the average over time) or the last curve γ_{31}, for example. We will model the deviations from the chosen baseline, forecast the deviations ahead in time, and add back the baseline curve. Here we use the mean curve $\bar{\gamma} = \sum_{t=1}^{31} \gamma_t/31$ as the baseline. To apply the PCA approach to the centered curves $\gamma_t - \bar{\gamma}$, we compute the sum-of-squares and cross products matrix

$$S = \sum_{t=1}^{31} (\gamma_t - \bar{\gamma})(\gamma_t - \bar{\gamma})'$$

and determine its eigenvalue-eigenvector decomposition

$$S = \Lambda D \Lambda',$$

where D is diagonal consisting of the eigenvalues of S (by convention, ordered from greatest to smallest) and Λ has columns consisting of the corresponding

orthonormal eigenvectors of S. Such a decomposition always exists, because S is non-negative. For any $J \leq 30$, the submatrix Λ_J consisting of the first J eigenvectors of Λ is of dimension 100 by J. The J-dimensional principal component approximation is obtained by regressing, for each year, the data vector $\gamma_t - \bar{\gamma}$ on Λ_J. The resulting regression coefficients are

$$\beta_t^J = (\Lambda_J' \Lambda_J)^{-1} \Lambda_J' \gamma_t. \quad (1)$$

Ordinarily in principle components $\Lambda_J' \Lambda_J$ equals the identity matrix, so that we get the principle components transformation $\beta_t^J = \Lambda_J' \gamma_t$, (Mardia, Kent, and Bibby 1979) but we will later smooth the columns of Λ_J destroying their orthonormality, so we retain the general expression as given above. The corresponding approximation of γ_t using J principal components is then

$$\hat{\gamma}_t^J = \bar{\gamma} + \Lambda_J \beta_t^J. \quad (2)$$

When $J = 0$, we set $\hat{\gamma}_t^0 = \bar{\gamma}$ by convention. Higher values of J improve the approximation of γ_t, but at the cost of a higher dimensional problem in forecasting β_t^J. Here we wish to keep J very low, but we still wish to accurately capture the main features of the curves.

In order to forecast the curves using PCA, we generate forecasts of the principal component series β_t^J. Given forecasts $\hat{\beta}_{31+h}^J$ for $h = 1, 2, \ldots, 50$ (see below for time series forecasting methods), we substitute into (2) to obtain forecasted curves

$$\hat{\gamma}_{31+h}^J = \bar{\gamma} + \Lambda_J \hat{\beta}_{31+h}^J.$$

Here we focus on the case that $J = 1$. A victory of this method was the excellent level of approximation obtained even for $J = 1$. As described in Mardia, Kent, and Bibby (1979), the ratio of the first J eigenvalues of S to the trace of S measures the proportion of total variation in the data that is explained by the first J principal components. Here this proportion was .92 for $J = 1$, and adding a second principal component only increases this proportion to .95. We thus set $J = 1$ so β_t^1 is a univariate time series that can be easily modelled and forecast.

3.3 Forecasting Principal Components

The one principal component series β_t^1 can be modelled and forecasted using univariate time series techniques. For our application, a simple ARIMA (Box and Jenkins 1976) model was deemed sufficient considering that the data limitations (length of series and errors in the historical data) did not warrant a very refined treatment. For the Hispanic employment immigrants data, we found that a decent fit was given by the ARIMA$(1, 1, 0)$ model with a trend (slope) constant:

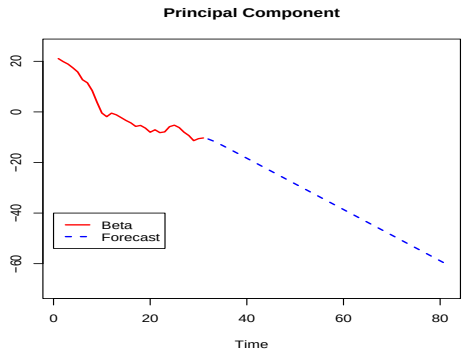
$$(1 - \alpha B)\left[(1 - B)\beta_t^1 - \mu\right] = \epsilon_t \quad (3)$$

where ϵ_t is a white noise sequence of variance σ^2, μ is a slope constant estimated via generalized least squares regression given α, the autoregressive parameter. Using the modelling capabilities of the program X-12-ARIMA (U.S. Census Bureau 2002), we obtained $\hat{\mu} = -1.01$, $\hat{\alpha} = .46$, and $\sigma^2 = 1.71$. The interpretation of the downward trend of β_t^1 is not obvious, since its product with Λ_J forms the approximation to γ_t. Since the model for β_t^1 is nonstationary, the forecasts will decrease in an unbounded fashion, possibly resulting in large deviations from the original age curves. In practice, we found this to be the case, especially when viewed 50 years out. Below in Figure 2 is a graph of β_t^1 together with 50 forecasts. The linear forecast pattern of β_t^1 creates implausible age distribution curves in the latter years of the forecast period. Empirically, we observed the following phenomena:

- Modes of the curve become increasingly high and narrow, to an unfeasible degree.
- Modes migrate to the older age groups, creating untenable results.
- Small noise perturbations in the original curves become large spikes in the long-term future.

These are all caused by the long-term behavior of the forecasts of β_t^1, which accentuate small pertur-

Figure 2: Plot and forecasts of principal component series for Hispanic Employment



bations in the principal components approximation to the age distributions into much larger perturbations in the forecasted age distributions 50 years out. On the other hand, the historical pattern of the series β_t^1 shows a steady downward trend that clearly should be reflected in the short-term forecasts. The problem is thus that the model (3) provides a reasonable description of the historical behavior of the series, and plausible short-term forecasts, but implausible long-term forecasts. Furthermore, *any* model or forecasting procedure that extrapolates the historical downward trend of β_t^1 indefinitely into the future will produce implausible long-term forecasts.

The first two problems noted above can be resolved by attenuating the forecasts of β_t^1. While simply truncating the forecast at a pre-specified limit would address the problems, this would yield somewhat strange results around the time point of truncation. Instead, we achieve a gradual attenuation of the forecasts by imposing a prior probability density function in the forecast period, as described Section 3.5 below. The amplification of the irregularities in the age distributions that appear to be "noise" can be addressed also through smoothing of the mean curve $\bar{\gamma}$ and Λ_1. When $J = 1$, the principal component approximation for the age distribution curve γ_t is $\bar{\gamma} + \Lambda_1 \beta_t^1$. Since β_t^1 is a scalar (univariate) time series, if both $\bar{\gamma}$ and Λ_1 are smooth over age then all the forecasted age distributions will be smooth. Actually, we can selectively smooth $\bar{\gamma}$ and Λ_1 to maintain any non-smooth features of the age distributions that appear to be real while smoothing away those that appear to be "noise." In Section 3.4 below, we investigate the use of cubic spline smoothers for this purpose.

The other product of the ARIMA forecasts is a standard error at each future time point. These standard errors increase as a function of the forecast horizon h, and are used in the Bayesian methods discussed below. As discussed in Bozik and Bell (1987), it is possible to translate such standard errors on forecasts of β_t^1 (actually, we need to use the full variance-covariance matrix of the forecast errors) into standard errors for the forecasted age distribution curves, but we do not pursue this calculation in this paper.

3.4 Cubic Spline Smoothers

Smoothing via cubic splines is discussed in Hastie and Tibshirani (1990). The basic concept is to fit a cubic polynomial to every pair of consecutive data points on the desired curve. This only provides two constraints for 4 unknowns; the remaining constraints are obtained from smoothness conditions between adjacent cubics. For a smoother curve, one can leave out certain data points, or adjust the smoothing parameters which govern the goodness of fit in the polynomial fitting.

In our implementation, we used the *smooth.spline* function in the *R* programming language (R Development Core Team 2004), with a smoothing parameter chosen such that small noise perturbations were eliminated, while preserving the major features of the curve. For our applications, we spline smoothed both the mean curve $\bar{\gamma}$ and the Λ_1 curve in the PCA decomposition. It is necessary for both of these curves to be smooth in order to ensure that forecasts are also smooth. However, it is important to avoid oversmoothing, and thus some care in the selection of

smoothing parameters is required. For $\bar{\gamma}$, we used an automatic choice of the smoothing parameter, whereas for Λ_1 we used the value of 0.5. We made our choice for Λ_1 based on aesthetic considerations.

One complication associated with using the spline smoother is that certain features believed to be intrinsic to the curve could get identified as noise, due to their structure. For example, the sharp drop in the Hispanic employment immigrants age distribution at age 22 is not an anomaly due to noise, but is a real feature of the data. Specifically, this drop reflects the fact that the "employment immigrants" below age 21 are primarily dependents of the actually employed immigrant, while few young adults over age 21 qualify as dependents. Hence, a sharp drop-off in dependents of employment immigrants between ages 20 and 22 results in the observed large drop in the aggregate (actually employed immigrants plus dependents) age distribution. On the other hand, the value at age 21 reflects a combination of dependents and actually employed immigrants, with a boost relative to the slightly lower ages due to age misstatement (an incentive to report oneself as 21 to qualify). Thus, we may wish to smooth out somewhat the peak at 21 as noise, but we wish to retain the drop at age 22 in the forecasts. Cubic splines applied to the full age distribution will automatically smooth out the drop at age 22. In order to force the preservation of this feature, we spline smooth the curve in two separate applications – up to age 21 and then 22 and up. Figure 3 below demonstrates the spline smoother on the transform of the mean age curve $\bar{\gamma}$. Note that the spline smoothing takes place on the logistic transformed data γ_t; when the transform is reversed, as depicted in Figure 3, the resulting curve is still smooth, but satisfies the properties of an age distribution, as discussed in Section 3.1 above.

3.5 Bayesian Attenuation

Bayesian methods have been used in time series modelling and forecasting – see, e.g., Thompson and Miller (1986). Below we formulate a fairly general approach to forecasting with *a priori* beliefs about the future. Let x denote an n-dimensional vector of observed data, and let y denote an unknown future value (this can be generalized to a multivariate scenario, but here y is scalar for simplicity). Let Model 1 portray the scenario that x and y are jointly normal:

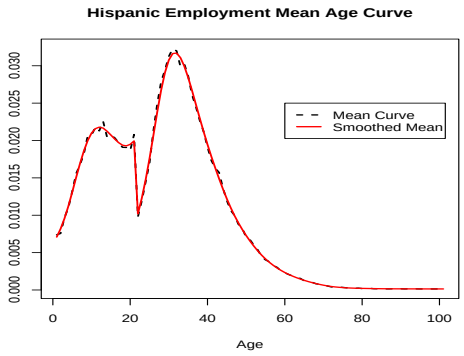
$$\begin{bmatrix} y \\ x \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma), \tag{4}$$

where $\mu = [\mu_y, \mu_x]'$ and Σ is the covariance matrix. Then we can easily write down the joint multivariate normal density $p_1(y, x)$, as well as the conditional density $p_1(y|x)$. Model 2 specifies bounds for y to be the numbers L and U (with $L < U$); this could be extended to form bounds L_i and U_i for each x_i, but we won't pursue this here. So the joint density $p_2(y, x)$ is essentially just a truncated version of $p_1(y, x)$, and is given by:

$$\begin{aligned} p_2(y, x) &= p_1(y, x) 1_{\{L<y<U\}}/c \\ c &= \Pr{}_1[L < y < U] \end{aligned} \tag{5}$$

where $\Pr_1(\bullet)$ denotes the probability of the event computed as if Model 1 were true, and $1_{\{L<y<U\}}$ is the indicator of the event that y is between L and U. Roughly speaking, we wish to combine time series forecasts with prior assumptions about what values are allowed. This is done by determining $p_1(y|x)$

Figure 3: Mean age curve and smoothed mean age curve for Hispanic Employment



purely from our time series model, and combining this with predetermined limits L and U for the future value y . The forecast conditional on (i) the data, and (ii) our *a priori* assumptions about the future values, has density $p_2(y|x)$, the conditional density in Model 2. Proposition 1, whose proof is in the appendix, gives a formula for this density.

Proposition 1 *The conditional density in Model* 2 *is given by*

$$p_2(y|x) = \frac{p_1(y|x)1_{\{L<y<U\}}}{\Pr_1[L < y < U|x]}.$$

The minimum mean square error prediction of y *from* x *in Model* 2 *is given by*

$$E_2(y|x) = \int y\, p_2(y|x)\, dy = \frac{\int_L^U y\, p_1(y|x)\, dy}{\Pr{}_1[L < y < U|x]}.$$

Remark 1 *The quantity* $\Pr_1[L < y < U|x] = \int_L^U p_1(y|x)\, dy$ *forms the appropriate normalization for* $p_2(y|x)$ *and is easily computed, since* $p_1(y|x)$ *is known.*

We apply Proposition 1 to the forecasts of the principal component series β_t^1. Here $(\beta_1^1, \beta_2^1, \ldots, \beta_{31}^1)$ plays the role of the data vector x, and any particular future value β_{31+h}^1 is our y. Our ARIMA model will provide $\hat{y} = E_1(y|x)$ as well as the variance V of $\hat{y}$, which does not depend on x. Thus

$$p_1(y|x) = \frac{1}{\sqrt{V}}\ phi\left(\frac{y-\hat{y}}{\sqrt{V}}\right)$$

where ϕ denotes the standard normal density, and hence

$$p_2(y|x) = \frac{\phi\left(\frac{y-\hat{y}}{\sqrt{V}}\right)1_{\{L<y<U\}}}{\sqrt{V}\left(\Phi(\frac{U-\hat{y}}{\sqrt{V}}) - \Phi(\frac{L-\hat{y}}{\sqrt{V}})\right)}$$

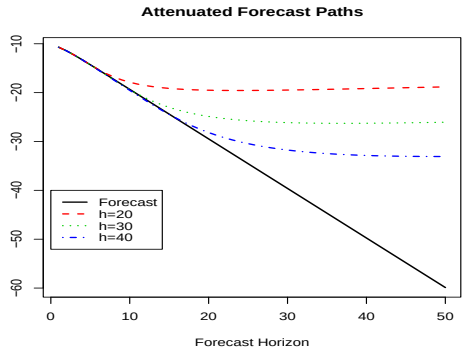
where Φ is the cumulative distribution function of the standard normal random variable. Finally, by integrating against y we obtain the optimal estimate under Model 2:

$$E_2(y|x) = \hat{y} - \sqrt{V}\,\frac{\phi(\frac{U-\hat{y}}{\sqrt{V}}) - \phi(\frac{L-\hat{y}}{\sqrt{V}})}{\Phi(\frac{U-\hat{y}}{\sqrt{V}}) - \Phi(\frac{L-\hat{y}}{\sqrt{V}})}. \qquad (6)$$

When a particular forecast $\hat{y}$ from the unconstrained Model 1 is extremely high or low, the Bayesian attenuation modifies this prediction towards the midpoint $(L+U)/2$.

The next practical question is "how should one choose L and U so as to best attenuate the forecasts?" The values of L and U have no obvious interpretation, and therefore some amount of trial and error is required. Figure 4 below plots the principal component forecasts $\hat{\beta}_{31+h}^1$ for various choices of L. U was set to the threshold of zero (greater than the maximum of all the forecasts), since no attenuation in this direction is necessary. Some "data-driven" choices for L are the various forecasts themselves, e.g., $L = \hat{\beta}_{31+h}^1$ for forecast horizons $h = 20, 30, 40$.

Figure 4: Forecast paths of Hispanic Employment with Bayesian attenuation. The original ARIMA forecast is plotted, together with attenuated forecasts with lower limit L is equal to the ARIMA forecast at horizon $h = 20, 30, 40$.



4 Application to Hispanic Employment Immigration

Here we summarize the procedures used on the Hispanic employment immigrants age distribution data.

1. Transform the data: add 1 to all data x_{it}, and normalize to form ratios. Apply the generalized logistic transformation, obtaining γ_t.

2. Compute the mean curve $\bar{\gamma}$. This is then spline-smoothed, taking any jags or special data features into account. We used age 21 as a breakpoint.

3. Compute the sum of squares and cross products matrix of $\gamma_t - \bar{\gamma}$, using the smoothed mean curve instead of the usual $\bar{\gamma}$. Find the eigenvectors to obtain the Λ matrix.

4. Select J and spline-smooth each column of Λ_J. We used $J = 1$, and customized the smoothing parameter to the value 0.5 in the R program.

5. Compute β_t^J via the formula (1).

6. Formulate a time series model for the principal component time series β_t^J. In our application $J = 1$, so we need only a univariate time series model for β_t^1. An ARIMA (1,1,0) model with trend constant was selected.

7. Forecast the principle component series, obtaining point forecasts and standard errors at each forecast horizon.

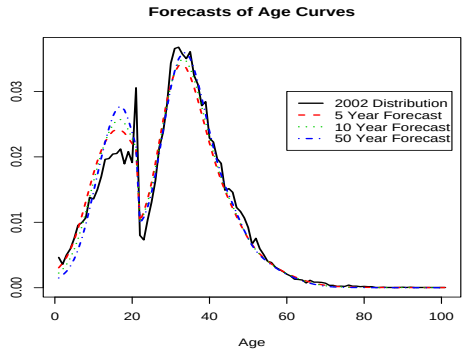
8. Modify the forecasts using Bayesian attenuation. One must decide upon an appropriate prior distribution for the forecast. We used the 30-year forecast value of the principal component as the lower limit in our Bayesian prior, and 0 as the upper limit.

9. Take the modified forecasted principal components, and apply (2). This gives the spline-smoothed, Bayesian-attenuated forecasts of the logistic-transformed data.

10. Undo the logistic transform to obtain the forecasted age distribution curves.

This is essentially the implementation used in our R program. Steps 1 through 5 are done in R, while steps 6 and 7 use the ARIMA modelling and forecasting capabilities of X-12-ARIMA. Then the output is read back into a second R program, which completes steps 8 through 10. The resulting forecasted age distribution curves for Hispanic employment immigrants are displayed in Figure 5. The forecasts have

Figure 5: Forecasted age distributions for Hispanic Employment, at 5, 10, and 50 year horizons, compared to the final year of data.



the desired properties discussed in the introduction: they are actual distributions, they are not overly simplistic (i.e., they don't smooth over relevant features of the data), and they are locally smooth. The use of PCA ensures that the curves give a fairly accurate representation of the data, with the accuracy controllable through the number of principal components used. The spline smoothing takes care of noise in a local fashion. Finally, the Bayesian attenuation ensures that forecasts at horizon $h = 50$ are not implausible. The difference between the age distribution in the last year of data and that from the first year of forecasts is somewhat large due to our use of $\bar{\gamma}$ as the baseline curve. If we instead used γ_{31} as the baseline curve, the initial forecasts would conform more closely to the particular behavior of the last year of data. Since the error in the data is highest in the last year, tying the forecasts more closely to the pattern in the last year is not necessarily desirable.

5 Summary

The PCA approach has previously been applied to age-specific fertility and mortality curves. The ap-

plication here to the age distribution of Hispanic employment immigrants posed some different challenges. For U.S. fertility and mortality rates for major race groups (e.g., white or nonwhite) the data could be regarded as quite accurate, providing some rationale for using sufficient principal components to provide a very accurate approximation. The immigrant data are of lesser accuracy, particularly in the last years of the data, so there was less reason to use more than one principal component to obtain a very accurate approximation.

In addition, the nature of the appropriate forecast functions differ across the applications. For fertility relatively flat forecasts were appropriate, as fertility rates have not shown extended downward or upward trends (since the post-war "baby boom" and subsequent "baby bust.") (Log) mortality rates, on the other hand, have consistently moved downward over time, so that forecast functions with downward linear trends continuing into the distant future appear reasonable. For the immigrant age distributions the historical data produce a first principle component series showing a steady downward linear trend, but forecasting this trend to continue indefinitely eventually produces implausible age distributions. Hence, the forecasts of the principle component series were attenuated using a Bayesian approach. Also, the mean age distribution curve and principle component vector were spline smoothed because use of the unsmoothed data led to accentuation of irregularities in these vectors, producing implausible forecast age distributions in the long term.

Finally, since the data we wished to forecast here are age distributions it was necessary to produce forecasts that are themselves age distributions (nonnegative values that sum to one). We achieved this objective by applying the generalized logistic transformation to the data, forecasting in the transformed scale, and inverse transforming the results to yield forecasted age distributions.

Acknowledgements The authors thank Fred Hollmann and Ward Kingkade of the Census Bureau for providing the immigration data, and communicating the age forecasting problem.

6 Appendix: Proof of Proposition 1

The desired conditional density is $p_2(y|x) = p_2(y,x)/p_2(x)$, and

$$\begin{aligned} p_2(x) &= \int p_2(y,x)\,dy \\ &= \int p_1(y,x)\,1_{\{L<y<U\}}\,dy/c \\ &= \int_L^U p_1(x)\,p_1(y|x)\,dy/c \\ &= p_1(x)\,\Pr{}_1(L<y<U|x)/c, \end{aligned}$$

from which it follows that

$$\begin{aligned} p_2(y|x) &= \frac{p_1(y,x)1_{\{L<y<U\}}/c}{p_1(x)\,\Pr{}_1(L<y<U|x)/c} \\ &= \frac{p_1(y|x)1_{\{L<y<U\}}}{\Pr{}_1(L<y<U|x)}. \end{aligned}$$

The formula for $E_2(y|x)$ follows at once. □

The methods of Bayesian attenuation can easily be generalized. Let $f(x)$ be a probability density function, and let $p_2(y,x) = p_1(y,x)f(y)/c$, where $c = \int p_1(y)f(y)\,dy = E_1[f(y)]$. Then (6) becomes

$$E_2(y|x) = \hat{y} + \sqrt{V}\frac{\int z\phi(z)f(\hat{y}+\sqrt{V}z)\,dz}{\int \phi(z)f(\hat{y}+\sqrt{V}z)\,dz}.$$

If f is symmetric about its mean $\mu = \int yf(y)\,dy$, then $E_2(y|x) = \hat{y}$ when $\hat{y} = \mu$.

References

[1] Aitchison, J. (1987), *The Statistical Analysis of Compositional Data*, London: Chapman and Hall.

[2] Bell, William R. (1997) "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates," *Journal of Official Statistics*, **13**, 279-303.

[3] Bozik, J. and Bell, W. (1987), "Forecasting Age Specific Fertility," SRD Research Report No. RR-87/19, U.S. Census Bureau, available at http://www.census.gov/srd/papers/pdf/rr87-19.pdf.

[4] Box, G.E.P. and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

[5] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall.

[6] Hollmann, Frederick W. (2004), "Toward a Legal Immigrant Data Series for Population Projections," paper presented at the 2004 meeting of the Southern Demographic Association, Hilton Head Island, South Carolina.

[7] Lee, Ronald D. and Carter, Lawrence R. (1992), "Modeling and Forecasting the Time Series of U.S. Mortality," *Journal of the American Statistical Association*, **87**, 659-671.

[8] Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, San Diego: Academic Press.

[9] R Development Core Team (2004), *R: A language and environment for statistical computing*, Vienna, Austria. http://www.R-project.org

[10] Rogers, A. and Castro, L. (1981), *Model Migration Schedules*, RR-81-30, Laxenburg, Austria: International Institute for Applied Systems Analysis.

[11] Thompson, P. A. and Miller, R. B. (1986), "Sampling the Future: A Bayesian Approach to Forecasting From Univariate Time Series Models," *Journal of Business and Economic Statistics*, **4**, 427–436.

[12] U.S. Census Bureau (2002), *X-12-ARIMA Reference Manual*, Version 0.2.10, U.S. Census Bureau, available at http://www.census.gov/srd/www/x12a/x12down_pc.html.