EVALUATING THE PERFORMANCE OF THE LEE-CARTER METHOD FOR FORECASTING MORTALITY*

RONALD LEE AND TIMOTHY MILLER

Lee and Carter (LC) published a new statistical method for forecasting mortality in 1992. This paper examines its actual and hypothetical forecast errors, and compares them with Social Security forecast errors. Hypothetical historical projections suggest that LC tended to underproject gains, but by less than did Social Security. True e_0 *was within the ex ante 95% probability interval 97% of the time overall, but intervals were too broad up to 40 years and too narrow after 50 years. Projections to 1998 made after 1945 always contain errors of less than two years. Hypothetical projections for France, Sweden, Japan, and Canada would have done well. Changing age patterns of mortality decline over the century pose problems for the method.*

Important policy decisions are made today on the basis of forecasts of the elderly population far into the future. Public pension policies are the prime example: fundamental changes have been proposed for the U.S. Social Security system in part because of a projected financial crisis 37 years from now, caused by population aging. Old-age dependency ratios are the key variable in these forecasts; they depend on the number of elderly people in the numerator and the number of working-age people in the denominator. The denominator depends heavily on future trends in fertility and perhaps in migration, but these are notoriously difficult to forecast. The elderly in the numerator have already been born, at least for forecasts up to a 65-year horizon, and so the numerator is on firmer ground. Yet Keilman (1997) found systematic underprediction of the elderly population in a number of industrial nations. He reported that after 15 years, forecast errors of –15% are not uncommon for females age 85+ (Keilman 1997: 272). A recent National Academy of Sciences study reported that U.N. projections made between 1965 and 1990 contained average errors of about –10% for the elderly populations of Europe and North America after 15 years (net of baseline error; National Research Council 2000:46; average errors in Third World countries were smaller but also negative).

Although immigration must have contributed to these errors, the main culprit is the systematic underprediction of mortality decline and gain in life expectancy. The National Academy study also reported, "For industrial countries, increases in life expectancy have been under-projected" (p. 132); Keilman (1998:38) reached a similar conclusion. We suggest that these problems continue in the recent and current forecasts made in industrial nations, including those by the U.S. Social Security Administration.

Mortality forecasts typically are based on the forecaster's subjective judgments, sometimes buttressed by expert opinion, and these judgments have tended to underestimate the pace of subsequent mortality decline in recent decades. In another approach, which is not without its own problems, the role of judgment is reduced by using extrapolation to forecast mortality.

Recently Lee and Carter (1992; henceforth LC) developed a new extrapolative method for modeling and forecasting mortality based on the analysis of long-term trends, and used it to make probabilistic forecasts of U.S. mortality to 2065. Since that time, the method has attracted a certain amount of attention and acceptance as well as some criticism. The most recent Census Bureau population forecasts (Hollmann, Mulder, and Kallan 2000) use the LC forecast as a benchmark for their long-run forecast of U.S. life expectancy. The two most recent Social Security technical advisory panels have recommended that the trustees adopt the method, or forecasts consistent with it. The method also has been used to forecast mortality in a number of other countries (most recently for the G-7 nations; see Tuljapurkar, Li, and Boe 2000).

Our main purpose in this paper is to make a careful, detailed assessment of the performance of the Lee-Carter method for forecasting mortality. There are only limited possibilities for the kind of ex post analysis of forecast performance conducted by Keilman and the NAS panel and reported above, but we will examine performance over the nine years that are available since the jump-off in 1989. Yet because the LC method involves less subjective judgment than those that have been used in the past, we can construct hypothetical forecasts with jump-off years earlier in the twentieth century, pretending we had only the data available up to that point, and comparing the subsequent pseudo-forecasts with the actual outcomes. We also conduct some similar but less detailed experiments using the LC method to produce forecasts for Japan, Canada, France, and Sweden, with 1950 as the jump-off year. Finally, we examine age patterns of decline during the twentieth century and consider the possibil-

*Ronald Lee, Demography and Economics, University of California, 2232 Piedmont Ave., Berkeley, CA 94720; E-mail: rlee@demog.berkeley.edu. Timothy Miller, Demography, University of California at Berkeley. Research for this paper was funded by Grant R37-AG11761 from NIA. We thank John Wilmoth for making available mortality data for the United States, France, Sweden, and Japan, through the Berkeley Mortality Data Base. We thank Statistics Canada and François Nault for making Canadian data available. John Wilmoth and Ken Wachter provided very useful suggestions for the analysis, and two anonymous referees provided valuable comments and suggestions.

ity that the age pattern has changed over time, contrary to the assumptions of the method. We discuss the suggestions and criticisms that have been made in light of the results of these studies.

OVERVIEW OF THE LC APPROACH

The basic LC model of age-specific death rates (ASDRs, and denoted $m_{x,t}$) is

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}. \quad (1)$$

Here a_x describes the general age shape of the ASDRs, while k_t is an index of the general level of mortality. The b_x coefficients describe the tendency of mortality at age x to change when the general level of mortality (k_t) changes. When b_x is large for some x, then the death rate at age x varies greatly when the general level of mortality changes (as with $x = 0$ for infant mortality, for example); when b_x is small, the death rate at that age varies little with changes in the general level of mortality (as is often the case with mortality at older ages). The model assumes that all the ASDRs move up or down together, although not necessarily by the same amounts, because all are driven by the same period index, k_t. In principle, not all the b_x need to carry the same sign; in the case of different signs, movement in opposite directions could occur. In practice, however, all the b_x have the same sign, at least when the model is fitted over fairly long periods. The proportional rate of decline of any death rate is given by $b_x(dk / dt)$. If dk / dt is constant—that is, if k_t is declining linearly—then each ASDR will decline at its own age-specific exponential rate, proportional to b_x, and depending on the rapidity of the decline in k_t. Gomez de Leon (1990), using exploratory data analysis on the historical data for Norway, selected the same model from a larger set of possibilities.

The strategy is to estimate this model on the historical data for the population in question, obtaining values for a_x, b_x, and k_t. The values of k_t form a time series, with one value for each year of data. Standard statistical methods then can be used to model and forecast this time series. LC selected a random walk with drift as the appropriate model, which has the form

$$k_t = k_{t-1} + c + e_t. \quad (2)$$

In this specification, c is the drift term, and k is forecast to decline linearly with increments of c, while deviations from this path, e_t, are incorporated permanently in the trajectory. The variance of e_t is used to calculate the uncertainty in forecasting k over any given horizon. The drift term, c, also is estimated with uncertainty, and the standard error of its estimate can be used to form a more complete measure of the uncertainty in forecasting k.

The projected k then can be used in Eq. (1), together with the estimated a_x and b_x, to calculate forecasts of the ASDRs; any desired life table functions can be derived from these. The probability intervals on the forecasts of k then can be used in the same way to calculate intervals for the forecasts of the ASDRs and (because these are all linear functions of the same k) the forecast of e_0. In addition, however, forecast errors in the ASDRs and e_0 derive from the $\varepsilon_{x,t}$ and from uncertainty about the true values of a_x and b_x. LC showed that these latter sources of error matter less and less as the forecast horizon lengthens, and are dominated by uncertainty about k in the long run. For a forecast horizon of 10 years, 98% of the standard error of the forecast of e_0 is accounted for by uncertainty in k. For the individual age-specific rates, the other sources of uncertainty are more important initially and remain important longer; after 25 years, however, most account for less than 10% of the standard error of the forecasts (see Lee and Carter 1992: table B2).

Inspection of Eq. (1) shows that the right-hand side of the equation includes no observed variable, so ordinary regression methods cannot be used to estimate the model. LC described a simple approximate method using regression methods, but the singular value decomposition (SVD) gives an exact least squares fit. In addition, if a_x, b_x, and k_t form one set of coefficients for the model, then a_x, b_x/A, and $A \times k_t$ will be an exactly equivalent set for any constant A. Similarly, $a_x - b_x \times A$, b_x, $k_t(1 + A)$ will be an equivalent formulation for arbitrary constant A. LC stipulated a unique representation by setting a_x equal to the average of the logarithms of $m_{x,t}$ over the data period, and setting the average value of k_t equal to zero. In this case the sum of the b_x values is unity.

In reference to Eq. (1), one can easily imagine adding a second term or additional terms expressing interactions of age and time effects: for example, b_x^2 / k_t^2. A number of authors have fitted models of this sort (Bell and Monsell 1991; Tuljapurkar and Li 2000; Wilmoth, Vallin, and Caselli 1990). Of course one can achieve an increasingly better within-sample fit by adding terms of this sort. In our view, however, it is preferable to retain the simpler model, which runs less risk of projecting into the future continuing changes in age patterns that in fact may be one-time historical changes or transitory alterations.

The method has a number of appealing features. The basic model is very simple. Although its use for forecasting involves a number of steps, each is simple in itself. The method is "relational," in demographers' terminology. That is, it involves the transformation of actual existing mortality schedules for each study population. Therefore, on the one hand, it is largely nonparametric; on the other, it incorporates particular features of a given population's mortality pattern.

The method is also probabilistic in the sense that it involves statistical fitting of models. The quality of the fit of the historical data can be used to provide probability intervals for the forecasts. As a matter of empirical fact, in the applications of the method to date, which involve at least 10 national data sets, the historical trend in k has always been highly linear with time, and the random walk with drift has given a good fit. This approximate linearity is useful for forecasting: it contrasts with the typically nonlinear trajectories of life expectancy, which rises at a decelerating rate when age-specific mortality rates decline at constant exponential rates. Nonetheless, if a data series extends sufficiently far back in time, the linearity of decline clearly would cease to hold.

Finally, the method also can be used as the basis of a simple-model life table system. Indirect estimation methods can be developed to expand the mortality data available as the basis for forecasting.

A number of criticisms and suggestions have been made since the original LC article was published, and the original method has been modified and extended in various ways. Some observers think that the probability bands are implausibly narrow (e.g., Alho 1992:673). Others argue that many age-specific rates are so low that they cannot realistically be projected to decline much further. Some believe that biomedical information should inform the forecasts, perhaps through incorporation of expert opinion, as is done by the Social Security actuaries. Some call for more within-sample testing of the methods; others question whether the a_x and b_x should be treated as invariant. Bell (1997) noted that the model did not fit the jump-off data very well.

Considerable work has been done to refine and extend the method since the original LC article. Wilmoth (1993) has developed improved fitting methods based on weighted SVD. Methods for modeling and forecasting regional systems of mortality have been developed (Lee and Nault 1993), as have more accurate procedures for dealing with the jump-off year (Bell 1997). Alternatives for modeling mortality for the oldest old have been explored. Consideration has been given to the special role of leader and follower countries (Wilmoth 1998). The LC method has been applied to cause-of-death data (Wilmoth 1998) to the sexes separately, and by race (Carter 1996a; Carter and Lee 1992). Carter (1996b) has projected k_t using a state space model in which the drift term is itself a random walk.[1] Many applications have been made to countries other than the United States (e.g., Lee and Rofman 1994; Tuljapurkar et al. 2000). Lee (2000) summarized the model's development, extensions, and applications such as stochastic forecasts of Social Security system finances.

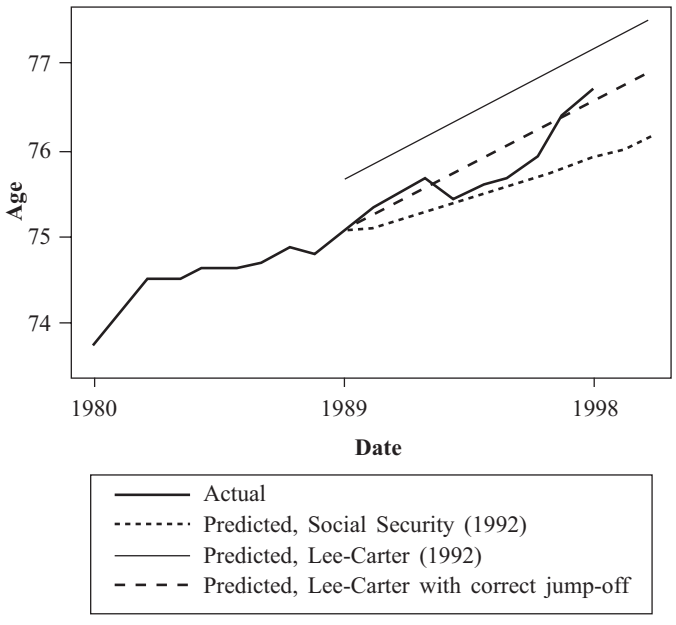
ASSESSING THE ORIGINAL FORECAST

In their original article, LC noted that the model would not fit the age-specific mortality data exactly in the jump-off year; thus the initial conditions for the forecast would not be quite right. This situation inevitably would lead to error, which would be particularly important in the early years of the forecast. They noted that it would be possible to set a_x equal to the most recently observed log age-specific rates, and thereby to fit the initial conditions exactly (with $k_t = 0$). They argued, however, that this practice might extrapolate idiosyncratic features of mortality in the jump-off year; therefore it was preferable to estimate a_x as the average values of the log death rates (Lee and Carter 1992:665–66). In retrospect, this apparently was a mistake: the 0.6-year error in e_0 at the jump-off year caused significant bias in the forecasts for the first decade, as we shall see below and as Bell (1997) pointed out. (LC estimated e_0 for 1989 at 75.66 years, whereas official data set it at 75.08.) Bell (1997) assessed the performance of four mortality forecasts: LC (as published), LC (with the jump-off year corrected), McNown and Rogers (1989), and the Social Security actuaries (SSA). He concluded that the LC forecasts were more accurate than the SSA or McNown-Rogers, but that a corrected LC forecast was better still.

Figure 1 displays the original LC mean forecast of e_0, another forecast similar but with the correct jump-off level, and the SSA projections made at the same time. The bias in the original LC projections is apparent, but it is also apparent that those projections correctly identified the trend in e_0. SSA appears to be somewhat low, ending about 0.8 year below the actual e_0. The adjusted LC projection is about 0.2 year too low in 1998 (the latest data available to us). Over this period, the actual e_0 always remains well within the 95% prediction interval for both the original and the adjusted LC.

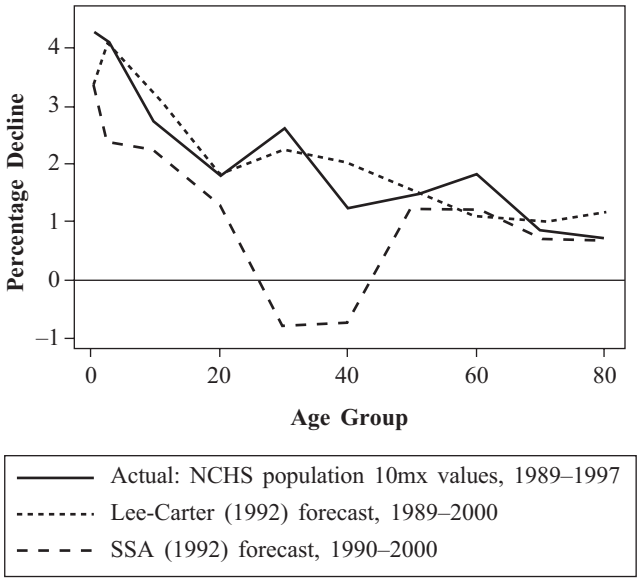
If the forecasts of e_0 performed well from 1989 to 1998, what of the forecasts of the individual age-specific rates? Once again, there are certainly errors due to the errors in initial conditions. Figure 2 focuses instead on the LC projected age-specific rate of decline of death rates from 1989 to 1997 for sexes combined, because this rate will not be affected by the errors in initial rates. It also plots the actual rates of decline and those projected by SSA. We see a striking agreement between the LC forecast and the actual rates of decline, particularly at the older ages. The SSA projections, however, incorrectly forecast slower mortality decline in the young adult years. We return to this topic later, for a different perspective on the age pattern of decline.

FIGURE 1. FORECASTS OF LIFE EXPECTANCY FROM 1989



1. Somewhat surprisingly, the results were barely distinguishable from those derived with the ordinary random walk model.

FIGURE 2. AVERAGE ANNUAL DECLINE IN AGE-SPECIFIC MORTALITY, 1989–1997: ACTUAL AND FORECASTS OF LEE-CARTER (1992) AND SSA (1992)



ANALYSIS OF LC PROJECTION ERRORS IN HYPOTHETICAL HISTORICAL PROJECTIONS

The Nature of the Tests

The original LC article included some tests of forecast performance within the historical data period, but none of these involved reestimating a_x, b_x, and k_t. Tests were restricted to forecasting k_t from different starting points in the historical period. Here we conduct a more rigorous test, in which we completely refit the model on each chosen subsample of data. Our earliest experimental forecast is based on data from 1900 through 1920. Our next forecast uses data 1900 through 1921; our next, 1900 through 1922; and so on until our last forecast, in which we use data from 1900 through 1997 to create a forecast for 1998. In this way we have 78 different forecasts for mortality one year ahead; 77 for a two-year horizon; and finally one forecast with a 78-year horizon. We reestimated the a_x and b_x for each set of data. Then we reestimated k_t for these years conditional on these a_x and b_x estimates, by choosing k_t (in the second stage) so as to match exactly the given value of e_0 in the data for the year in question.[2] This approach departs slightly from the procedure followed originally by LC: they chose k_t to match total deaths, which requires annual age-distributed population data as well.

Once we had estimated k_t for each year of the sample, we did not follow standard diagnostic methods to choose an optimal ARIMA model for each data subsample. Instead we assumed that the random walk with drift model held. We fitted it and used it to forecast k_t over the desired time range, and derived probability distributions from the subsample ARIMA errors.

LC introduced a dummy variable for the influenza epidemic of 1918. Our preference today would be to include the dummy variable (permitting a one-time positive change in k in 1918, followed by a one-time equal negative change in k in 1919), and in the forecast to incorporate a $1/T$ chance of an identical positive and negative change in k, where T is the length of the base period over which the model was fit. This has a small effect on both the mean and the variance of the forecast. We did not do this for the experimental forecasts described here.

Although these retrospective tests are something like the ex post analysis of forecasting errors, there are also significant differences. First, we developed the method with the benefit of the preceding century of mortality experience; thus we would be surprised if our method failed to accord with that history. Second, a forecaster would have to decide how far back in time to go in fitting the model to historical data. Mortality data for the United States do not start until 1900; even then, they cover only a limited number of the states. All our forecasts use data back to 1900, although the jump-off year of our first forecast is 1920. Third, we have assumed that a random walk with drift is the forecasting model always used, although the rate of drift is estimated afresh for each forecast. It would not be feasible to choose manually an optimal ARIMA model specification for each of the 78 forecast jump-off years. If we had done this, the short-run performance of the model presumably would have been better, but the long-run performance might have been worse.

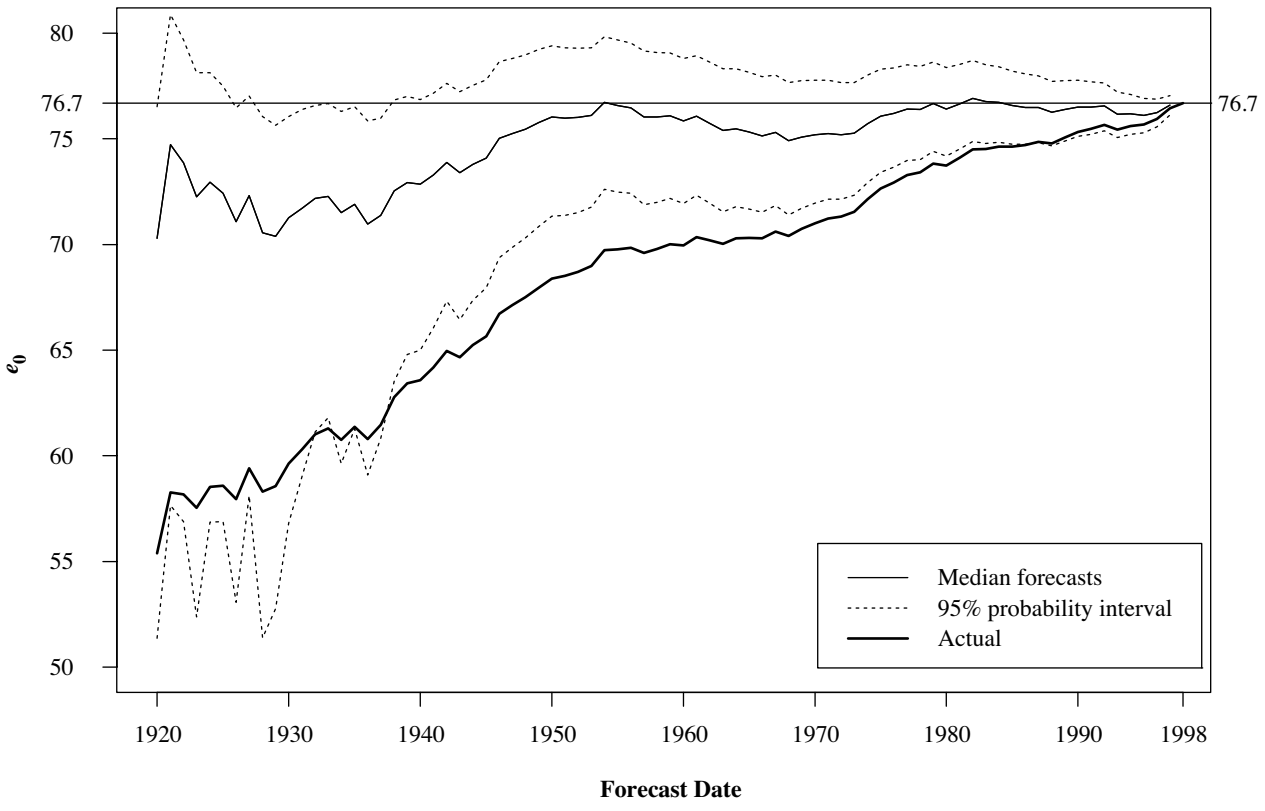
Forecasting to 1998 (e_0)

Figure 3 plots all 78 forecasts for life expectancy in 1998; each originates from a different jump-off year, and each covers a different forecast horizon. Each forecast for 1998 is plotted above its jump-off date. The 95% probability intervals are plotted as well. The horizontal line indicates the observed value of life expectancy for 1998; thus it is the true value relative to which the forecasts can be assessed.

We point out several details. First, although the hypothetical forecasts tend to be too low, generally they are fairly close to the actual value for 1998. The earlier forecasts, using data through the 1920s and 1930s, are on average five years below the true value; beginning in 1946, all forecasts are within two years of the correct value. Overall, the mean forecasts look quite good. Second, the 95% probability intervals failed to include the true value for 1998 in 12 of the 78 forecasts, or 15% of the time, compared with the intended percentage of 5%. Third, the median forecast for 1998 fell

2. In all cases, the data we use are taken from the SSA database, as maintained on the Berkeley Mortality Data Base website, http://www.demog.berkeley.edu/wilmoth/mortality. In the original article, LC used NCHS data; for the period before 1933 they estimated age-specific mortality and e_0 indirectly, using the age distribution of the total population, total deaths per year, and the a_x and b_x coefficients as estimated by SVD from the data 1933 to 1987, after the death registration area was complete.

FIGURE 3. e_0 FORECASTS FOR 1998 BY FORECAST DATE



below the actual value for 1998 in 74 of the 78 forecasts, or 95% of the time; that figure suggests downward bias.

Errors by Forecast Horizon (e_0)

It is also useful to assess forecast errors (defined as forecast value less actual value) by horizon. We have done this for horizons of 1, 5, 10, 20, 40, and 60 years. For a one-year horizon, we have 78 different jump-off dates; for the 60-year horizon, we have only 19. For each forecast, we find the percentile in the probability distribution of that forecast where the observed value falls. For example, if the actual value corresponds to the median of the forecast distribution, we assign it 50; if it corresponds to the lower 7% of the distribution, we assign it 7; and so on. We then plot the frequency distribution of these percentile scores. If the probability distribution associated with each forecast in fact describes the probability distribution of errors, this frequency distribution should be uniform between 0 and 100. If the actual distribution of percentiles tends to be concentrated in the middle, around 50, this indicates that the distribution of the errors is clustered more tightly than our forecast leads us to expect. If there are fewer percentiles in the middle of the distribution and more toward 0 and 100, our forecast understates the width of the error distribution. If most of the true values fall below the 50th percentile, then most of the time we have overpredicted; if they fall above the 50th percentile, we have tended to systematically underpredict the true value.

Figure 4 is the histogram of the percentiles for five-, 20, and 40-year horizons. The actual forecast errors match the predictive distribution quite well at the five-year horizon. By 20 years, however, a tilt toward positive errors is unmistakable, and this tilt intensifies at the 40-year horizon. The vertical scales are compressed increasingly as the errors become more concentrated.

Table 1 presents various measures of forecast performance, including the mean squared error (MSE), the mean absolute percent error (MAPE), the average error (bias), the percentage of positive errors, and the proportion of actual values that fall within the 95% probability interval of the forecast. The table reports performance by forecast horizons and offers a summary covering all forecast horizons.

The average errors are negative, an indication that the method tended to underpredict gains in life expectancy in the United States, particularly when launched from earlier dates. The "percentage underprojected" column confirms this point. For example, 91% of errors for 31- to 40-year projection horizons were negative (predicted e_0 less than actual), as were 100% of errors beyond a 50-year horizon. The 95% confi-

FIGURE 4. PERCENTILE ERROR DISTRIBUTION BY FORECAST LENGTH

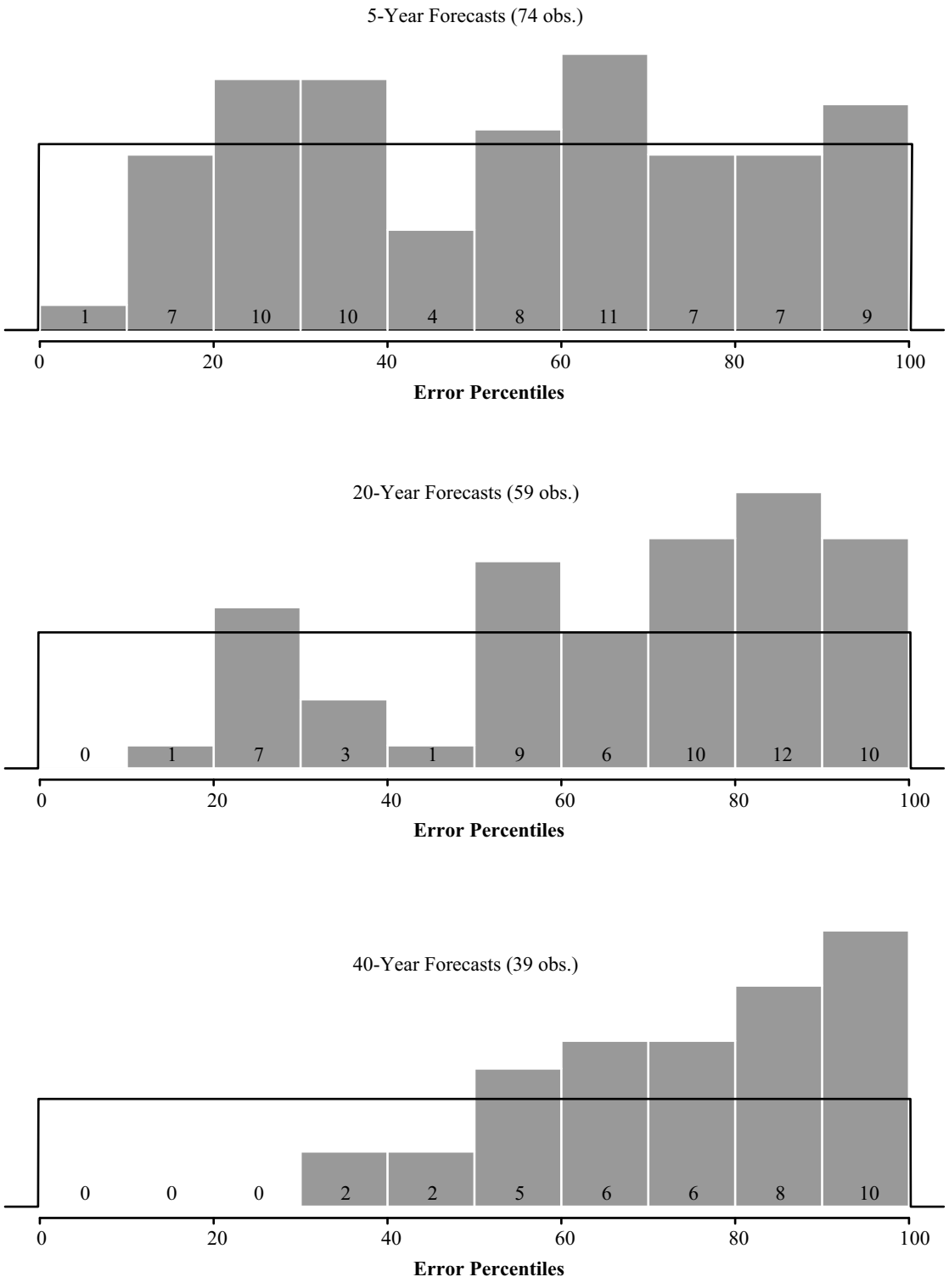


TABLE 1. MEASURES OF FORECAST PERFORMANCE BY FORECAST HORIZON

Forecast Horizon (N)	Average Error	MAD	RMSE	MAPE	% Under-Projected	% Within 95% Interval
1–5 (380)	–0.11	0.45	0.60	0.16	54	99
6–10 (355)	–0.32	0.82	1.03	0.47	56	100
11–20 (635)	–0.73	1.23	1.60	1.15	67	97
21–30 (535)	–1.37	1.47	1.99	2.03	84	100
31–40 (435)	–1.68	1.73	2.14	2.45	91	100
41–50 (335)	–2.23	2.25	2.75	3.41	96	95
51–60 (235)	–3.54	3.54	3.75	5.07	100	89
61–78 (171)	–4.38	4.38	4.53	5.39	100	80
All (3,081)	–1.49	1.76	2.34	2.45	78	97

dence bounds contain the actual e_0 value 97% of the time for all horizons combined. Yet they appear to be too broad for intervals up to a 40-year horizon and too narrow for those beyond a 50-year horizon.

Error Correlations by Age and Horizon

As noted briefly above, Eq. (1) contains an error term, $\varepsilon_{x,t}$, because the expression does not provide a perfect representation of variation in age-specific rates over time. In formulating the probability intervals for the life expectancy forecasts, we ignored this error term and incorporated only errors arising from the innovation in k_t and from errors in estimating the drift term. If we were interested only in e_0 and if the $\varepsilon_{x,t}$ were uncorrelated across age, this assumption might be relatively harmless because some 20 different values of $\varepsilon_{x,t}$ enter into the calculation of any life expectancy; these will tend to cancel each other out, leaving a small net effect. If the errors are correlated, however, such that those for older ages tend to move together and those for younger ages tend to move together, they might exert an important influence even on life expectancy. The estimation of the a_x and b_x coefficients also contains errors, which are not taken into account in our probability intervals for the e_0 forecasts.

We find that forecast errors tend to show a strong positive correlation at younger ages and to be less strong at older ages; the errors at young ages are correlated only weakly with those at older ages. At longer horizons, correlations become more positive because of the dominance of errors in k.

ASSESSING LC ON HISTORICAL TIME SERIES FROM OTHER COUNTRIES

We also conducted a simple within-sample test for Sweden, Japan, France, and Canada. For Canada, France, and Sweden we constructed a forecast to 1995 from a jump-off date of 1950. For France and Sweden, we used data from 1900 to 1950. For Canada, data are available only from 1922 to 1950. For Japan, suitable data are available from 1950, so we took the later jump-off year of 1973. For France, we used dummy variables to capture the profound effects of both World War I and World War II, but we did not allow for a possible recurrence in the future, which would have greatly increased the variance of the forecast. Such decisions reflect the analyst's judgment.

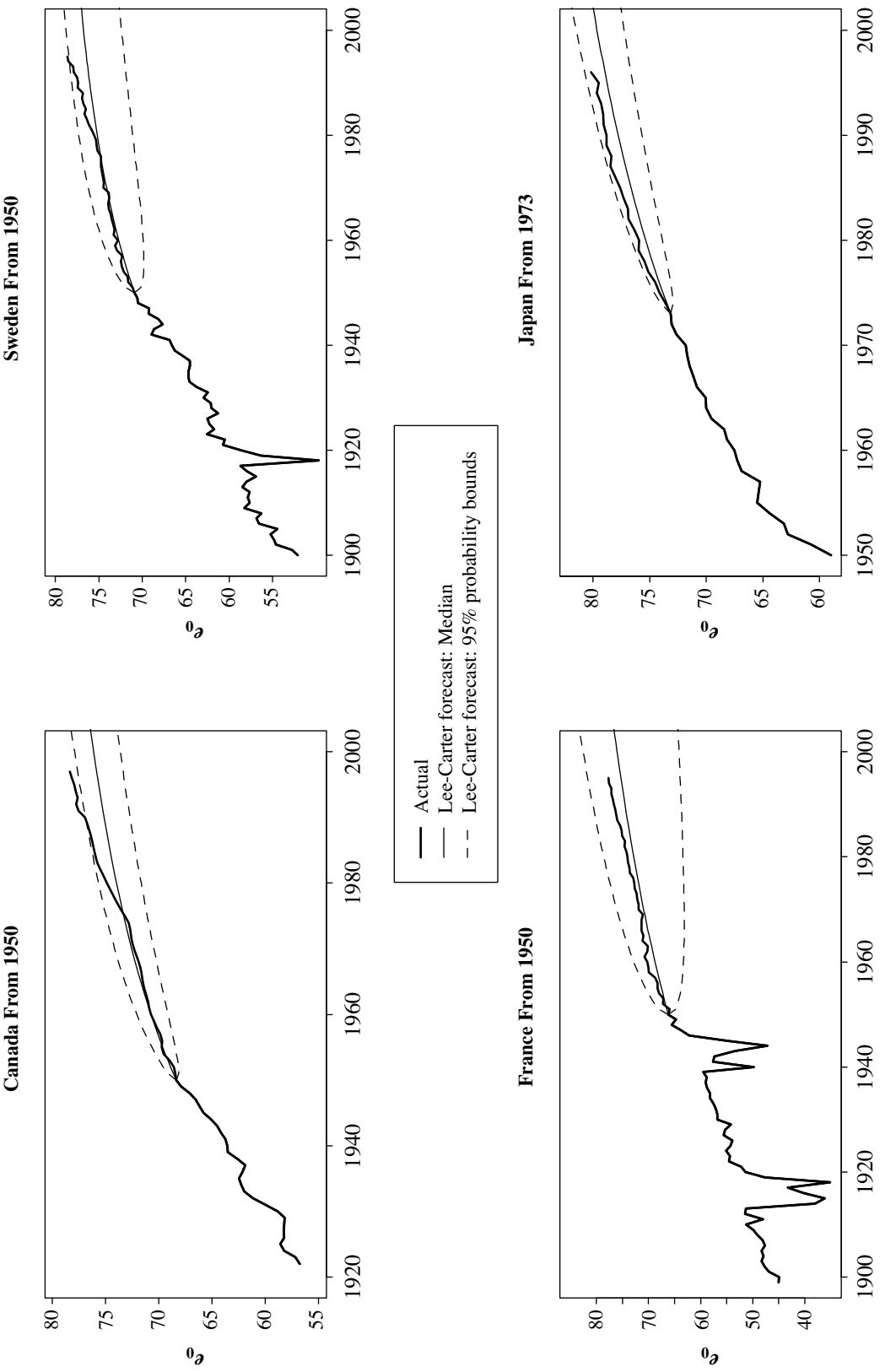
The results are shown in Figure 5. If the method had been used to forecast 1995 e_0 for Sweden, starting in 1950, it would have been on target until 1980, and two years too low in 1995. Results for France and Canada are very similar. For Japan, forecasts from 1973 to 1996 are below the actual value; they are one year too low by 1996. In all the forecasts combined, the 95% probability bounds contain the actual e_0 values for 152 of 162 forecast values or 94%, very close to the 95% coverage predicted by the models. In every country, however, we find a systematic tendency to underpredict future gains in life expectancy, just as in the United States. We return to this topic later.

THE CHANGING AGE-SHAPE OF MORTALITY

A number of observers have suggested that the b_x coefficients might vary over time; LC did not explore this possibility. Kannisto et al. (1994) found that the rate of mortality decline had been accelerating over recent decades for ages 80 to 100. Horiuchi and Wilmoth (1995) showed that in a number of countries, mortality at older ages now declines more rapidly than at lower ages, in a reversal of the historical pattern. This research suggests the importance of the possibility that the age pattern of mortality decline may change over time, and may not be described accurately by a fixed set of b_x coefficients. The a_x coefficients will always be changing over different historical periods because they are the average log death rates. The level of these averages will change as mortality declines; the shape will change because the b_x coefficients tell us that at different ages, mortality declines at different rates. This poses no problem because the changing shape and level of the a_x are implicit in the b_x, and no additional treatment is necessary.

Our earlier examination of LC's postpublication performance showed that LC predicted correctly the age pattern of mortality decline as well as the increase in e_0 over the past nine years. This suggests that the fixed b_x assumption has worked well. We learn otherwise, however, on closer exami-

FIGURE 5. LC FORECASTS OF LIFE EXPECTANCY



nation of the age pattern of decline in the United States. The top panel of Figure 6 depicts the average rate of decline for the sexes' combined mortality by age in the United States for 1900 to 1950 and for 1950 to 1995. It suggests that an important change has occurred: mortality now is declining at roughly the same rate across all ages above 15, whereas for the first half of the century it declined far more rapidly at the younger ages. The lower panels of Figure 6, which show changes in the historical age pattern of mortality decline in Sweden, France, Canada, and Japan, indicate similarly striking alterations, with a flattening of the age profile of decline.

Is this a long-term change, rooted in the changing cause structure of mortality or in the resistance of mortality at different ages to biomedical progress? Or is it due to what we hope will be transitory influences on young adult mortality in industrial nations, such as AIDS and accidents? We are not sure, but it is more prudent to assume that these are long-term changes and to incorporate them somehow into our forecasts. A simple and satisfactory solution, adopted by Tuljapurkar et al. (2000), is to base the forecast on data since 1950, and to assume fixed b_x over that range but not over the whole century. Only about 6% of life table deaths now occur in the age range affected by the changing age pattern (say, from 10 to 50); thus the change in the age pattern of decline has relatively weak effects on the forecast of e_0. It seems likely that the systematic tendency of the LC method to underpredict gains in e_0 at long horizons is due in some way to this changing age pattern of decline, but it is not clear exactly how.

COMPARISON OF OFFICIAL FORECASTS FROM SSA AND OTHERS WITH LC FORECASTS

Forecasting to 1998

We have examined the historical record of SSA projections, including two earlier projections that were used by SSA but prepared by other agencies. Figure 7 examines forecasts of e_0 for 1998. The dots represent the high-middle-low forecasts of the SSA; the lines represent the median with LC's 95% probability interval. The figure shows that the official middle projections have been systematically too low: by 12 years in 1930, by about seven years in the 1940s, and then by two to four years until the projections made in 1980. In that year, SSA forecasts jumped too high for a few years and then fell too low again. The SSA estimates reacted strongly to the slow mortality gains of the 1960s, and then to the rapid gains of the 1970s. By contrast, the LC method responds only modestly to these fluctuations because they exert only a modest effect on the average trend over the century. The LC method also tends to be somewhat low in early years, but performs substantially better than SSA. Its median forecast would have been closer to the true value in 1998 for most forecast horizons. It picks up the correct track for 1998 considerably earlier.

Figure 7 also displays the high-low range of SSA projections along with LC's 95% probability interval. The true value of e_0 for 1998 lies beyond the high bound for most of the SSA forecasts until 1970. Alho (1990) found similar results in his comparison of trend extrapolation with SSA forecasts of standardized mortality rates.

Errors by Horizon: Comparison With LC

In assessing errors by forecast horizon, we have restricted our sample to post-1950 government forecasts. We have only three early government forecasts (pre-1950), which provided e_0 forecasts for only a few selected years in the future. As a result, the analysis of errors by length of horizon is complicated for these groups. For comparison with LC, we use both the full sample (1920–1997) and a restricted sample that matches the period of the SSA forecasts (1950–1997). For LC, we have hypothetical forecasts for every year. For SSA, the forecasts were issued irregularly until 1980; after that year they became available annually.

In our calculations we have weighted each SSA forecast by the reciprocal of the number of forecasts issued within the decade. In this way, each decade contributes equally to the error estimates. Without weighting, the SSA results are dominated by forecasts made since 1980, and the longer horizon forecasts count for little.

Figure 8 compares the average bias in the SSA and LC forecasts by length of forecast horizon. Horizons are presented by single year from 1 to 7 and then are grouped (8–12, 13–17, 18–22, 23–27, 28–38, 39–46, and 39–60 years). SSA forecasts issued since 1950 compare favorably with LC forecasts issued since 1920 for horizons up to 15 years, and perform less well thereafter. Yet when we compare only the LC forecasts issued during the same period as the SSA projections (since 1950), we find that LC performs very much better at all horizons.

The General Problem of Official Forecasts

Long-run government forecasts have relied on expert opinion, which proved to be too pessimistic about the future. This pessimistic outlook might be attributed to the mood of the country at the time the forecasts were issued. Two of the earliest population forecasts were produced by the National Resource Committee (1937) during the Great Depression and by the National Resource Planning Board (1943) during World War II. Yet at those times the data were telling a different story because mortality had been declining quite rapidly over the previous decades.

A quote from the 1943 report is revealing in this regard. Thompson and Whelpton stated their objection to statistical forecasting methods such as extrapolation: "More important, the extrapolation of past trends according to such formulas might show future trends which seemed incompatible with present knowledge regarding the causes of death and the means of controlling them" (National Resource Planning Board 1943:10). This suggests an alternative explanation for the pessimistic bias of expert opinion: present knowledge informs us about current limits, but not about the future means of overcoming them. The Lee-Carter approach bases its long-run forecasts on the century-long decline in mortality, in which limits have been continuously confronted and overcome.

FIGURE 6. AVERAGE ANNUAL REDUCTION IN AGE-SPECIFIC DEATH RATES

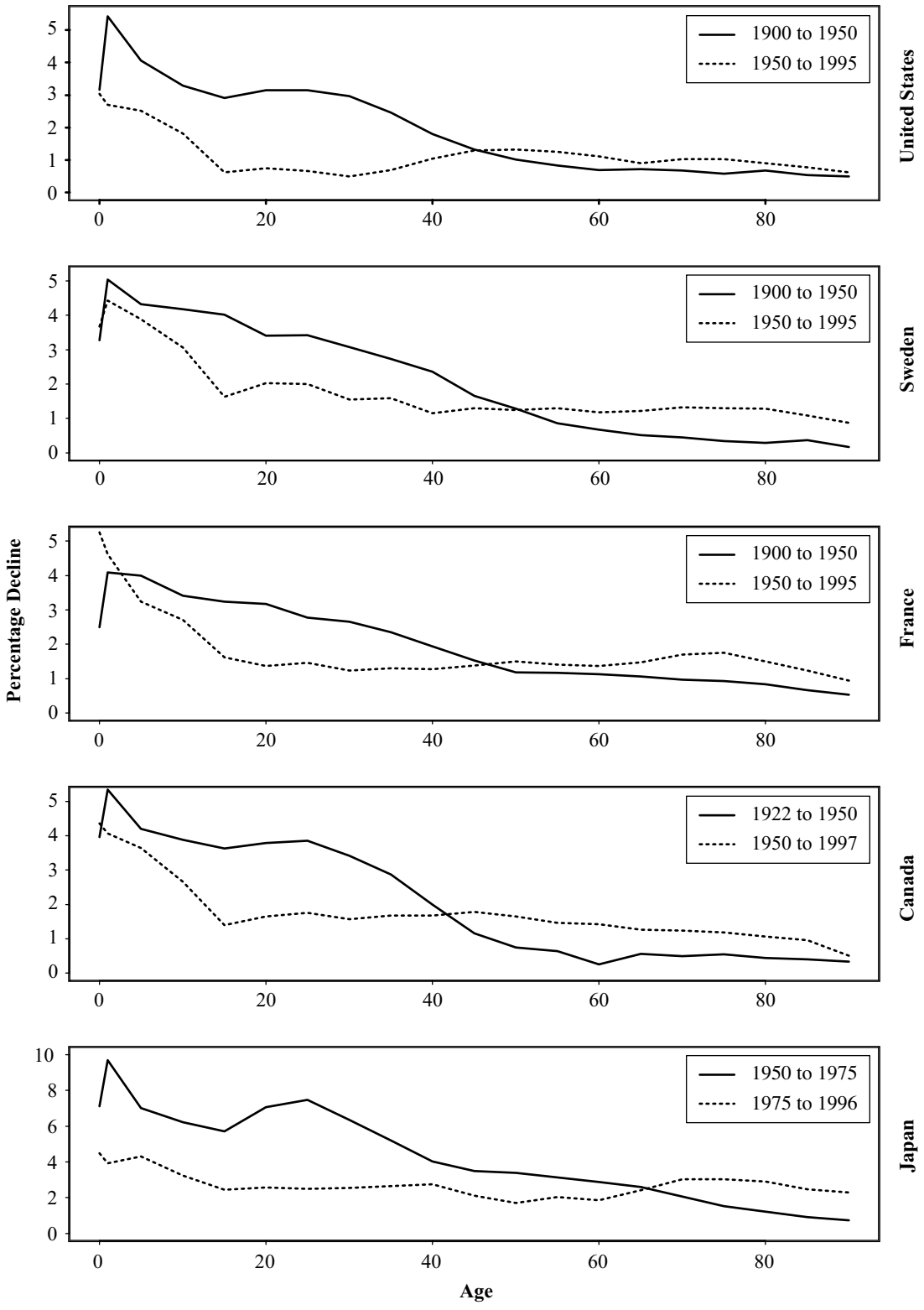


FIGURE 7. FORECASTS OF e_0 IN 1998 BY FORECAST DATE

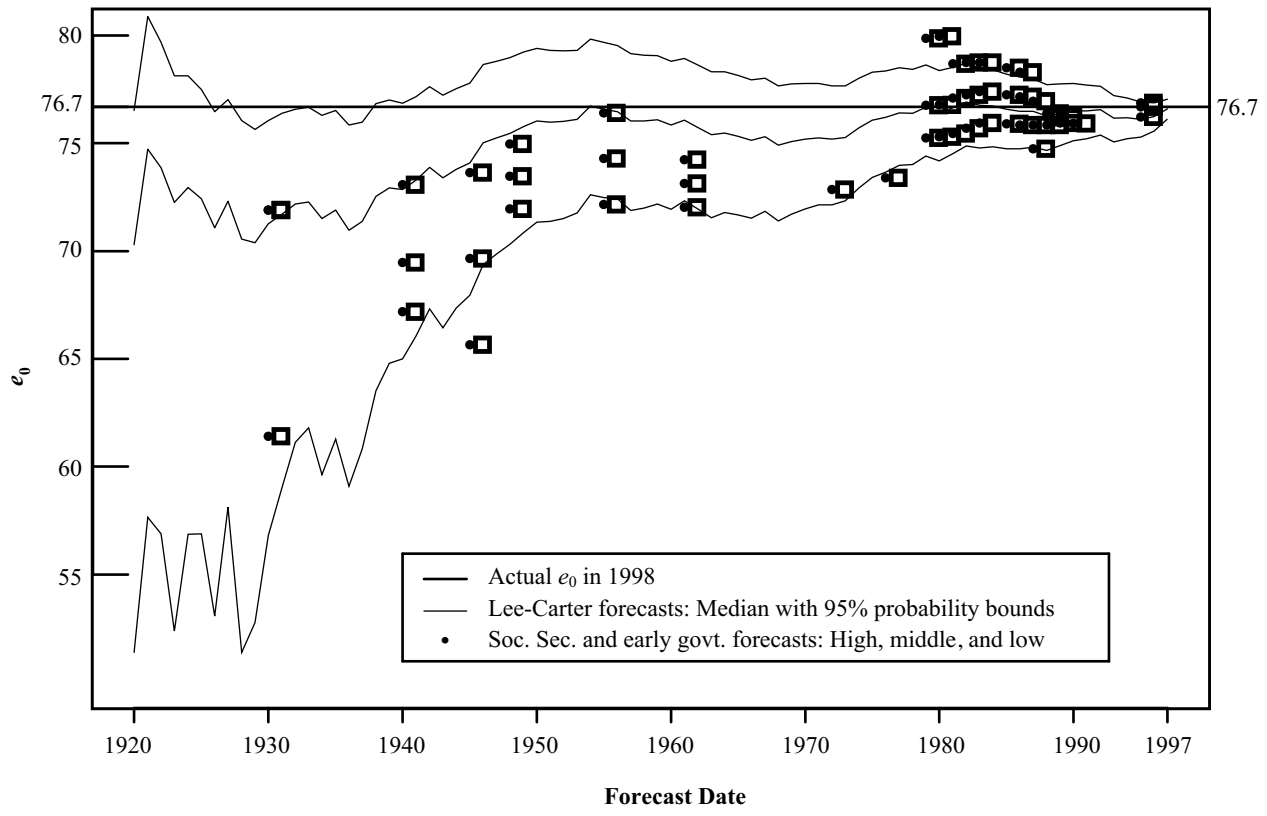
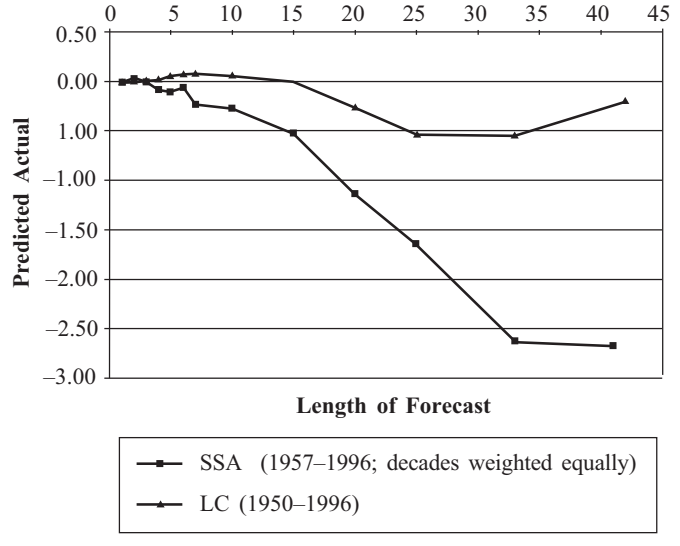


FIGURE 8. AVERAGE BIAS IN FORECASTS OF LIFE EXPECTANCY



CONCLUSIONS

We can extract several lessons from these investigations. First, the LC forecasts of life expectancy and the age pattern of mortality performed quite well for the period since publication, at least after adjusting for an error in jump-off level.

Second, hypothetical LC projections from various historical jump-off dates in the twentieth century would have performed well. For forecasts with jump-offs after 1945, LC projections are always within two years of the actual e_0 in 1998. The forecasts tend to underpredict future gains, however, especially those in the distant future. Although the 95% probability bounds contain the true value of e_0 97% of the time, the bounds appear to be too broad for horizons up to 40 years and too narrow for those beyond 50 years.

Third, Social Security projections also have systematically underpredicted gains in e_0 since 1950. The average error and mean squared error for LC forecasts since 1950 are substantially lower than those of SSA since 1950, when each decade is given equal weight.

Fourth, LC life expectancy forecasts for Canada, Sweden, and France with jump-off year 1950, and for Japan

with jump-off year 1973 would have performed very well. As in the United States, however, the forecasts would have systematically underpredicted actual gains, particularly at longer horizons.

Fifth, contrary to a basic assumption in the Lee-Carter model, the age pattern of mortality decline has shifted systematically in the United States, Sweden, France, Canada, and Japan in the second half of the twentieth century, with a flattening of the age-specific rates of decline above age 15. Although this has distorted e_0 projections only slightly, it can exert a substantial effect on projected death rates at ages from 1 to 35. For example, the median forecast for the U.S. life expectancy in 2075 based on post-1950 data is 86.2 years, about 0.5 year higher than the forecast based on post-1900 data. The age-specific rates for ages 1 to 35 are 30% to 80% lower in the forecast based on post-1900 data. The absolute errors are small, however, because the projected rates themselves are so low.

Finally, the results overall suggest that the LC method produces surprisingly accurate forecasts over rather long periods. Used for long-term forecasts within the twentieth century, the LC method generally would have tended to underpredict future gains in life expectancy. The probability intervals, despite some problems, also are surprisingly successful in containing the true outcomes.

These findings bear on some of the criticisms and suggestions addressed to the LC method, which we mentioned briefly above. Some observers suggested that the probability bounds were too narrow. We found that for forecasts of life expectancy at birth up to 50 years into the future, the probability bounds are too broad rather than too narrow; for longer forecasts, however, they are somewhat too narrow, with 80 to 90% coverage rather than the intended 95%. Some argue that many age-specific rates are so low that they cannot realistically be projected to decline much further. Although death rates at ages 10 to 50 continue to decline, the rates of decline have decelerated in relation to those at other ages. These changes in the age distribution of decline may reflect an approach to lower limits. Yet the declines at the younger and older ages continue unabated or have accelerated. Some have suggested that biomedical information should inform the forecasts—but if so, how? The Social Security actuaries have used expert opinion on mortality decline by cause of death. We find that their forecasts have been systematically too low, more so than those of the LC extrapolative approach, and that the mean squared errors of their forecasts have been greater than those of LC as well.

Some have questioned whether the relative pace of decline by age should be treated as invariant over the century. This question is justified: in the second half of the century, mortality at older ages has declined more rapidly relative to that at younger ages than in the first half of the century. Some have suggested that it would be preferable to take the most recently observed age-specific death rates as the jump-off point for the forecasts, rather than the fitted age distribution in the jump-off year. On the basis of the past 10 years' experience, this point also appears to be correct.

The Lee-Carter method takes a simple extrapolative approach. It is easy to think of reasons why its long-run forecasts should fail. Indeed, we have uncovered a number of shortcomings in the method's performance. We are impressed overall, however, not by the shortcomings we have found so far, but rather that they are not larger and more numerous. In our tests, the method performs better than we had reason to expect, both in predicting the future (and the pseudo future) and in indicating uncertainty.

Our results suggest that projections using the LC method should be taken seriously. For example, Tuljapurkar et al. (2000:792) used this method to project for a number of industrial nations that e_0 will be one to four years higher in 2050 than indicated by official projections, with larger discrepancies for Japan. It may well be that national and international agencies today are continuing their systematic underprediction of life expectancy. As industrial nations strive to confront the long-term funding problems of their public pension systems, realistic projections of mortality are particularly important. Although we cannot know what future mortality trends will be, we suggest that LC-type forecasts provide a useful baseline for planning.

REFERENCES

Alho, J.M. 1990. "Stochastic Methods in Population Forecasting." *International Journal of Forecasting* 6:521–30.

———. 1992. "Modeling and Forecasting the Time Series of U.S. Mortality." *Journal of American Statistical Association* 87:673–74.

Bell, W.R. 1997. "Comparing and Assessing Time Series Methods for Forecasting Age Specific Demographic Rates." *Journal of Official Statistics* 13:279–303.

Bell, W.R. and B.C. Monsell. 1991. "Using Principal Components in Time Series Modeling and Forecasting of Age-Specific Mortality Rates." Pp. 154–59 in *Proceedings of the Social Statistics Section, American Statistical Association*.

Carter, L. 1996a. "Long-Run Relationships in Differential U.S. Mortality Forecasts by Race and Gender: Non-Cointegrated Time Series Comparisons." Presented at the annual meetings of the Population Association of America, May 9, New Orleans.

———. 1996b. "Forecasting U.S. Mortality: A Comparison of Box-Jenkins ARIMA and Structural Time Series Models." *Sociological Quarterly* 37:127–44.

Carter, L. and R.D. Lee. 1992. "Modeling and Forecasting U.S. Mortality: Differentials in Life Expectancy by Sex." *International Journal of Forecasting* 8(3):393–412.

Gomez de Leon, J. 1990. "Empirical DEA Models to Fit and Project Time Series of Age-Specific Mortality Rates." Central Bureau of Statistics, Oslo, Norway. Unpublished manuscript.

Hollmann, F.W., T.J. Mulder, and J.E. Kallan. 2000. "Methodology and Assumptions for the Population Projections of the United States: 1999 to 2100." Working Paper 38, Population Division, U.S. Bureau of the Census.

Horiuchi, S. and J.R. Wilmoth. 1995. "The Aging of Mortality Decline." Presented at the annual meetings of the Population Association of America, April 6, San Francisco.

Kannisto, V., J. Lauristsen, A.R. Thatcher, and J.W. Vaupel. 1994.

"Reductions in Mortality at Advanced Ages: Several Decades of Evidence From 27 Countries." *Population and Development Review* 20:793–810.

Keilman, N. 1997. "Ex-Post Errors in Official Population Forecasts in Industrialized Countries." *Journal of Official Statistics (Statistics Sweden)* 13:245–77.

———. 1998. "How Accurate Are the United Nations World Population Projections?" *Population and Development Review* 24:15–41.

Lee, R.D. 2000. "The Lee-Carter Method for Forecasting Mortality, With Various Extensions and Applications." *North American Actuarial Journal* 4:80–91.

Lee, R.D. and L. Carter. 1992. "Modeling and Forecasting the Time Series of U.S. Mortality." *Journal of the American Statistical Association* 87:659–71.

Lee, R.D. and F. Nault. 1993. "Modeling and Forecasting Provincial Mortality in Canada." Presented at the World Congress of the International Union for the Scientific Study of Population, August 24–September 1, Montreal.

Lee, R.D. and R. Rofman. 1994. "Modelación y Proyección de la Mortalidad en Chile" (Modeling and Forecasting Mortality in Chile). *NOTAS* 22:182–213.

McNown, R. and A. Rogers. 1989. "Forecasting Mortality: A Parametrized Time Series Approach." *Demography* 26:645–60.

National Research Council. 2000. *Beyond Six Billion: Forecasting the World's Population,* edited by J. Bongaarts and R.A. Bulatao. Washington, DC: National Academy Press.

National Resource Planning Board. 1943. *Estimates of the Future Population of the United States, 1940–2000.* Washington, DC: U.S. Government Printing Office.

National Resources Committee. 1937. *Population Statistics: Material Prepared for a Study of Population Problems.* Washington, DC: U.S. Government Printing Office.

Tuljapurkar, S. and N. Li. 2000. "Mortality Change: The Structure of Short-Term Variability." Presented at the annual meetings of the Population Association of America, March 23–25, Los Angeles.

Tuljapurkar, S., N. Li, and C. Boe. 2000. "A Universal Pattern of Mortality Decline in the G-7 Countries." *Nature* 405:789–92.

Wilmoth, J.R. 1993. "Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change." Technical report, Department of Demography, University of California, Berkeley.

———. 1998. "Is the Pace of Japanese Mortality Decline Converging Toward International Trends?" *Population and Development Review* 24:593–600.

Wilmoth, J.R., J. Vallin, and G. Caselli. 1990. "When Does a Cohort's Mortality Differ From What We Might Expect?" *Population: English Selection* 2:93–126.