Popul Res Policy Rev (2007) 26:347–369
DOI 10.1007/s11113-007-9034-9

Precision, bias, and uncertainty for state population forecasts: an exploratory analysis of time series models

Jeff Tayman · Stanley K. Smith · Jeffrey Lin

Received: 4 January 2006 / Accepted: 2 July 2006 / Published online: 23 June 2007

Abstract Many researchers have used time series models to construct population forecasts and prediction intervals at the national level, but few have evaluated the accuracy of their forecasts or the out-of-sample validity of their prediction intervals. Fewer still have developed models for subnational areas. In this study, we develop and evaluate six ARIMA time series models for states in the United States. Using annual population estimates from 1900 to 2000 and a variety of launch years, base periods, and forecast horizons, we construct population forecasts for four states chosen to reflect a range of population size and growth rate characteristics. We compare these forecasts with population counts for the corresponding years and find precision, bias, and the width of prediction intervals to vary by state, launch year, model specification, base period, and forecast horizon. Furthermore, we find that prediction intervals based on some ARIMA models provide relatively accurate forecasts of the distribution of future population counts but prediction intervals based on other models do not. We conclude that there is some basis for optimism regarding the possibility that ARIMA models might be able to produce realistic prediction intervals to accompany population forecasts, but a great deal of work remains to be done before we can draw any firm conclusions.

Keywords ARIMA · Forecast accuracy · Forecast uncertainty · Population forecasts · Prediction intervals

J. Tayman (✉)
San Diego Association of Governments (Retired), 2142 Diamond Street, San Diego, CA 92109, USA
e-mail: jtayman@san.rr.com

S. K. Smith
University of Florida, Gainesville, FL, USA

J. Lin
University of California, San Diego, CA, USA

Introduction

A substantial amount of research over the last several decades has dealt with the measurement and evaluation of uncertainty in population forecasts. Much of this research has focused on the development and application of time series models. Some researchers have developed models of total population growth (e.g., Alho and Spencer 1997; Pflaumer 1992), whereas others have focused on individual components of growth—typically mortality and fertility (e.g., Keilman et al. 2002; Lee 1974, 1992; Lee and Tuljapurkar 1994; McNown and Rogers 1989), but occasionally migration as well (e.g., De Beer 1993). These studies have focused primarily on issues such as sources of uncertainty in population forecasts, the development of models that provide specific measures of uncertainty, and how point forecasts and prediction intervals vary from one model to another. Few have evaluated the accuracy of the resulting forecasts or analyzed the out-of-sample validity of prediction intervals.

Most of the research on time series forecasting models has focused on the national level. Population forecasts, however, are widely used for planning and analytical purposes at the state and local levels (Smith et al. 2001). Although a number of studies have evaluated the precision and bias of state and local forecasts (e.g., Murdock et al. 1984; Rayer 2004; Smith and Sincich 1992; Tayman 1996; White 1954), only a few have attempted to evaluate forecast uncertainty (e.g., Smith and Sincich 1988; Swanson and Beck 1994; Tayman et al. 1998). Given the widespread use of subnational population forecasts for decision-making purposes, the growing importance of time series models in national population forecasting, and the increasing emphasis in the literature on measures of uncertainty, we believe an analysis of time series forecasting models for subnational areas is in order.

Few studies have developed and evaluated time series models for subnational areas. In the most comprehensive study, Voss et al. (1981) tested several ARIMA models for states and chose a single model for their detailed analyses. Using a number of different launch years and forecast horizons, they used this model to construct population forecasts for the 48 contiguous states in the United States. They evaluated forecast accuracy by comparing the resulting forecasts with census counts and census-based population estimates. They also compared the accuracy of ARIMA forecasts with the accuracy of several other forecasting models and found it to be roughly the same. The authors briefly discussed the use of ARIMA models for constructing prediction intervals and examined their efficacy as a measure of forecast uncertainty, but did not pursue that line of research.

In this article, we select six ARIMA models for states in the United States. Using these models and a series of annual population estimates from 1900 to 2000, we construct population forecasts for four states chosen to reflect a range of population size and growth rate characteristics; both of these factors are known to affect forecast accuracy (Smith et al. 2001). The forecasts are based on a variety of combinations of model, launch year, base period, and forecast horizon; these factors are also known to affect forecast accuracy (Smith et al. 2001). We compare the forecasts with census counts for the corresponding years and attempt to answer the following questions:

(1) What is the impact of differences in model specification, length of base period, and length of forecast horizon on precision, bias, and the width of prediction intervals?
(2) How consistent are the results from one state to another?
(3) How consistent are the results from one launch year to another?
(4) What proportion of future populations fall within the prediction intervals?
(5) What do these results tell us about the usefulness of time series models for forecasting subnational populations and for assessing the uncertainty of those forecasts?

We focus on ARIMA models and do not investigate other time series models or statistical methods that borrow strength across time and space or rely on non-Gaussian error distributions (e.g., Granger and Newbold 1986). Such alternative approaches are potentially useful but lie beyond the scope of the present study.

We have two basic objectives. The first is to summarize the out-of-sample error characteristics of commonly used ARIMA forecasting models. The second is to evaluate the out-of-sample performance of the prediction intervals produced by those models. By "out-of-sample," we mean that data from a historical base period (e.g., 1900–1950) are used to produce forecasts for subsequent years (e.g., 1960, 1970, and 1980). By using out-of-sample point forecasts and prediction intervals, we can simulate actual forecasting situations in which information beyond the historical base period is unknown. Although this is an exploratory analysis based on a limited number of states, we believe these simulations provide useful information regarding the forecasting performance of commonly used time series models and the validity of using such models as predictors of population forecast uncertainty.

Data and terminology

Using a set of annual population estimates from 1900 to 2000 (U.S. Census Bureau 1956, 1965, 1971, 1984, 1993, 2002), we started by analyzing population change by decade for all states in the United States with the exception of Alaska and Hawaii, which did not have data back to 1900. For our empirical analysis, we chose four states that exhibited widely varying size and growth rate characteristics: Wyoming, Maine, Florida, and Ohio. These states reflect a broad range of population growth patterns and provide a diverse data set for conducting an exploratory analysis at the state level.

Wyoming is the smallest state in the United States, with a population of just under half a million in 2000. As shown in Table 1, its growth rates fluctuated considerably from one decade to the next, including one decade with negative growth. Maine is also a small state (1.3 million in 2000), but exhibited moderate and relatively stable growth rates between 1900 and 2000. Florida is the 4th largest state, with almost 16 million residents in 2000. It has grown rapidly but unevenly since 1900, with growth rates ranging between 27% and 78% per decade. Ohio is the 7th largest state, with a population of 11.4 million in 2000. Ohio has grown much less rapidly than Florida, but its growth rates have fluctuated considerably from one decade to the next.

Table 1 Percentage population change in the 20th century by decade, sampled states

State	1900–1910	1910–1920	1920–1930	1930–1940	1940–1950	1950–1960	1960–1970	1970–1980	1980–1990	1990–2000
Florida	42.6	27.2	52.9	30.2	46.7	78.1	37.9	42.6	32.4	23.2
Maine	7.2	3.5	3.8	6.1	8.0	6.3	2.9	12.4	9.3	3.7
Ohio	15.0	21.2	14.9	4.0	15.2	22.0	10.2	0.7	0.6	4.6
Wyoming	58.1	34.0	14.7	10.6	16.0	14.1	1.5	41.1	−4.2	8.8

These four states followed markedly different population growth patterns during the 20th century. Florida's population grew by 13.2 million between 1950 and 2000, compared to only 2.3 million between 1900 and 1950. Maine also added more residents during the second half of the century than the first half: 360,000 compared to 222,000. Growth in Wyoming was about the same in both time periods, as the state added 197,000 residents between 1900 and 1950 and 204,000 between 1950 and 2000. Ohio was the only state in our sample that added fewer residents in the second half of the century than the first, growing by 3.8 million between 1900 and 1950 and 3.4 million between 1950 and 2000.

We use the following terminology to describe population forecasts:

(1) Base year: the year of the earliest population size used to make a forecast.
(2) Launch year: the year of the latest population size used to make a forecast.
(3) Target year: the year for which population size is forecasted.
(4) Base period: the interval between the base year and launch year.
(5) Forecast horizon: the interval between the launch year and target year.

For example, if data from 1900 through 1950 were used to forecast population size in 1980, then 1900 would be the base year, 1950 would be the launch year, 1980 would be the target year, 1900–1950 would be the base period, and 1950–1980 would be the forecast horizon.

ARIMA modeling

A number of different time series models can be used for forecasting purposes. In this study, we use ARIMA models based on past population values and the dynamic and stochastic properties of error terms. Like other extrapolation methods, these models do not require knowledge of underlying structural relationships; rather, they are based on the assumption that past values provide sufficient information for forecasting future values. The two main advantages of univariate ARIMA models are: (1) they require historical data only for the population of the area being forecasted, and (2) their underlying mathematical and statistical properties provide a basis for developing probabilistic intervals to accompany point forecasts (Box and Jenkins 1976; Nelson 1973). ARIMA models are commonly used for forecasting purposes, but the methods used in developing and applying those models are more complex than is true for most extrapolation methods.

ARIMA model specifications

The general ARIMA model is expressed as ARIMA(p,d,q) where p is the order of the autoregressive term, d is the degree of differencing, and q is the order of the moving average term. Models based on time intervals of less than one year may also require seasonal components that are not relevant in the present study. Our aim in this study is to investigate the behavior of commonly used ARIMA specifications that reflect different assumptions about future population trajectories. To this end, we analyzed six ARIMA models.

Models 1 and 2 contain a constant term and incorporate first-order differences and first-order terms for p and q. Model 1 contains a first-order autoregressive term but no moving average term; it is identified as ARIMA(1,1,0). This model has been used for population forecasts of Sweden (Cohen 1986; Saboia 1974) and states in the United States (Smith and Sincich 1992); some analysts have found it to outperform more complex time series formulations (e.g., Voss et al. 1981). Model 2 contains a first-order moving average term but no autoregressive term; it is identified as ARIMA(0,1,1) and is equivalent to a simple exponential smoothing model. Models of this type have been used by Alho (1990) to forecast U.S. mortality. De Beer (1993) used a similar model, but it did not require differencing because the net migration time series for Netherlands he analyzed was already stationary. Forecasts from ARIMA models with a first difference and constant term will follow a linear trend, with the constant term equal to the slope of the trend.

Models 3 and 4 contain a constant term and incorporate second-order differences. Model 3 contains a second-order autoregressive term but no moving average term; it is identified as ARIMA(2,2,0) and was used by Pflaumer (1992) to forecast U.S. population. Model 4 contains a first-order moving average term but no autoregressive term; it is identified as ARIMA(0,2,1). Models of this type have been used by Cohen (1986) and Saboia (1974). Some analysts have suggested that second differences may be optimal for modeling human populations and have provided evidence that models containing higher order differences may outperform models using only first differences (e.g., McNown and Rogers 1989; Saboia 1974). Forecasts from models such as Models 3 and 4, with second-order differences and a constant term, follow a quadratic trend and their prediction intervals diverge more quickly and are wider than intervals based on models containing first-order differences (Makridakis et al. 1998).

These four models have often been used to forecast population change or the components of population growth. We also investigated two other ARIMA models. Model 5 contains second-order differences and moving average parameters but does not have a constant term; it is identified as ARIMA(0,2,2) and is equivalent to Holt's linear trend exponential smoothing method. This model extrapolates the trend in the historical data series with more weight given to the last two observations (Makridakis et al. 1998). The population trajectory for Model 5 tends to be more accelerated than a model with first differences and a constant term, but less accelerated than a model with second differences and a constant term (Nelson 1973).

Our last ARIMA model (Model 6) has the same specifications as Model 2 but uses the natural logarithm of population; it is identified as ARIMA ln(0,1,1). One characteristic of prediction intervals based on natural log transformations is that although they are not symmetric around the population forecast, they are symmetric around the forecast of the transformed population (Nelson 1973). Pflaumer (1992) examined forecasts based on an ARIMA(1,1,0) model using the natural logarithm of the U.S. population. We evaluated both ARIMA ln(1,1,0) and ARIMA ln(0,1,1) specifications using the autocorrelation and partial autocorrelation functions, the augmented Dickey–Fuller test, the Bayesian Information Criteria, and the Portmanteau test. (We discuss these and other model identification techniques in the following section.) We chose the ln(0,1,1) model because it does not appear to be misspecified for any state and the model selection criteria favored it over the ln(1,1,0) model.

As noted below, we fit each model for each state using 15 different combinations of base period and launch year. For individual states and the combination of forecasts across states, we can judge the performance of these six models in terms of precision, bias, and the performance of prediction intervals. Although particular models may be misspecified for particular states, we believe that combining forecasts across states will provide interesting and useful findings.

To address the issue of potential model misspecification, we created a 7th model (Hybrid) using the five ARIMA specifications applied to the untransformed population (i.e., Models 1–5). These five models represent a plausible range of population trajectories for each state. To build the hybrid model, we applied model identification techniques to each state/launch year combination and selected the best of the five ARIMA models. The next section discusses the procedures used to identify ARIMA specifications and choose the specific models that comprise the hybrid model.

ARIMA model identification

ARIMA model identification refers to the process for determining the best values of p, d, and q, which typically range from 0 to 2. The d value must be determined first because a stationary time series is required to properly identify the values of p and q (e.g., Brockwell and Davis 2002; Granger 1989). A nonstationary time series can generally be converted into a stationary one by taking first or second differences. The long-term dynamics of ARIMA models are generally controlled more by the differencing term (d) than by the values of the autoregressive and moving average terms (p and q), which have their greatest impact on short-term dynamics (Chatfield 2000).

Fifty observations is often suggested as the minimum required for identifying ARIMA models (e.g., Granger and Newbold 1986; McCleary and Hay 1980; Meyler et al. 1998; Saboia 1974). We therefore used 50-year base periods in our attempts to identify the best models. In our analyses of forecast errors, however, we used base periods of five different lengths (10, 20, 30, 40, and 50 years) in order to test for the effects of differences in length of base period on forecast accuracy and the performance of the prediction intervals. This is discussed more fully in the following section.

The Box–Jenkins (1976) approach to ARIMA model identification relies on assessing the patterns of the autocorrelation function (ACF) and partial autocorrelation function (PACF) and their standard errors. This quasi-formal approach to model identification is subjective and highly dependent on the skill and interpretation of the analyst, especially in the case of mixed ARMA models (Granger and Newbold 1986; Meyler et al. 1998). To deal with this problem, several less subjective methods have been developed to help identify the best ARIMA model. These methods are based on statistical tests for stationarity (e.g., Dickey et al. 1986) and statistics such as the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) for selecting the best values of p and q while avoiding models with too many parameters (Brockwell and Davis 2002).

We used the augmented Dickey–Fuller unit root test to identify the degree of differencing required to obtain a stationary series. This test performs a regression of the differenced series $Y'_t = Y_t - Y_{t-1}$ against a lagged term (Y_t) and lags in the differenced series, which are usually set to three (Makridakis et al. 1998):

$$Y'_t = \phi Y_{t-1} + b_1 Y'_{t-1} + b_2 Y'_{t-2} + b_3 Y'_{t-3}.$$

If the series Y_t is stationary, φ will be negative and significant, causing a rejection of the null hypothesis that there is a unit root. We evaluated the augmented Dickey–Fuller test for six series: the original series, a logarithmic transformation of the original series, and first and second differences of both the original and transformed series.

We also analyzed the AIC and BIC statistics for the six ARIMA models. Improved fit of the ARIMA model lowers AIC and BIC values, while additional terms that do not increase the likelihood more than the penalty amount increase them (Makridakis et al. 1998). Thus, the smallest values for the AIC and BIC statistics are desirable when selecting an ARIMA model. The BIC measure is preferable to the AIC measure since it less likely to lead to an over-parameterized model (Brockwell and Davis 2002; Meyler et al. 1998). The results of our analyses were similar using both the AIC and BIC statistics; we present only the latter in this study.

The top half of Table 2 reports MacKinnon approximate *p*-values for the augmented Dickey–Fuller test; *p*-values less than .05 indicate a stationary series. The lowest-differenced instances that result in a stationary series are highlighted, once for the unlogged series and once for the logged series. We focus on the least amount of differencing necessary to achieve a stationary series; this helps us avoid selecting an over-differenced model that can lead to inflated sample variances and prediction intervals (Makridakis et al. 1998; Meyler et al. 1998). We note, however, that prediction intervals from time series models have often been found to underestimate forecast uncertainty even for models that have been appropriately specified (Chatfield 2000).

The original series was nonstationary in every state/launch year combination, as evidenced by *p*-values that exceeded .86. Across launch years, Florida required second differencing in order to achieve stationarity in the unlogged series. For the logarithmic series, stationarity was achieved with a single-order difference. In Maine and Ohio, the augmented Dickey–Fuller test indicated a stationary series after a single-order difference, for both the original and logarithmic series. Results

Table 2 Objective criteria for identifying ARIMA specifications[a]

	Dickey–Fuller test probabilities*					
	Florida			Maine		
	1950	1960	1970	1950	1960	1970
Original Series	.999	1.000	.999	.923	.990	.967
1st Difference	.088	.761	.605	.005	.002	.003
2nd Difference	.000	.000	.000	.000	.000	.000
ln (Original series)	.962	.997	.950	.895	.982	.950
ln (1st Difference)	.010	.033	.017	.004	.001	.002
ln (2nd Difference)	.000	.000	.000	.000	.000	.000
	Ohio			Wyoming		
	1950	1960	1970	1950	1960	1970
Original Series	.923	.990	.967	.866	.948	.873
1st Difference	.005	.002	.003	.332	.017	.022
2nd Difference	.000	.000	.000	.001	.000	.000
ln (Original Series)	.895	.982	.950	.263	.553	.716
ln (1st Difference)	.004	.001	.002	.288	.012	.016
ln (2nd Difference)	.000	.000	.000	.000	.000	.000
	Bayesian Information Criteria (BIC)					
	Florida			Maine		
	1950	1960	1970	1950	1960	1970
Model 1: (1,1,0)	544	574	575	360	373	373
Model 2: (0,1,1)	546	594	595	365	376	376
Model 3: (2,2,0)	546	576	577	367	388	390
Model 4: (0,2,1)	529	571	572	363	384	386
Model 5: (0,2,2)[b]	536	572	573	362	379	376
Model 6: ln(0,1,1)	−212	−213	−227	−306	−310	−313
	Ohio			Wyoming		
	1950	1960	1970	1950	1960	1970
Model 1: (1,1,0)	602	622	622	293	308	310
Model 2: (0,1,1)	603	625	625	294	309	310
Model 3: (2,2,0)	604	636	636	295	327	329
Model 4: (0,2,1)	594	628	628	291	314	317
Model 5: (0,2,2) [b]	596	626	626	290	313	312
Model 6: ln(0,1,1)	−280	−283	−286	−229	−245	−262

[a] Based on a sample of 50 observations

[b] Model does not include a constant term

*$p < .05$ rejects the hypothesis of a unit root (nonstationary time series)

were less consistent across launch years in Wyoming. For both the logged and unlogged series, data for launch year 1950 indicated that a second-order differenced series was needed to achieve stationarity, whereas a single-order difference was

sufficient for launch years 1960 and 1970. The first-order differenced logarithmic series is the most consistent for the states in our sample, being stationary in 11 out of 12 state/launch year combinations.

The bottom half of Table 2 reports the BIC statistics, with the numbers for the best model (i.e., the one with the lowest non-negative value) highlighted for each state/launch year combination. In general, the BIC results were consistent with the findings of the augmented Dickey–Fuller test. Because of the logarithmic transformation, criteria for the logarithmic series are not comparable to criteria for the original series. In general, the ARIMA ln(0,1,1) model performed well across all states and launch years, with the exception of Wyoming for the 1950 launch year. In this case, a second-order difference was required to make the logarithmic series stationary, but we still favor a logarithmic model with a first- rather than second-order difference because the latter yielded prediction intervals that appeared unrealistically wide, especially for longer forecast horizons.

Turning to the results for the unlogged series, BIC values in Florida were lowest for the ARIMA(0,2,1) model. In Maine, the BIC for the ARIMA(1,1,0) model was substantially lower than that for other models across launch years. In Wyoming, the BIC suggested that an ARIMA(0,2,2) model without a constant was best for the 1950 launch year, while an ARIMA(1,1,0) was best for later launch years. In Ohio, the BIC suggested that ARIMA(0,2,1) was the best model for the 1950 launch year, while the ARIMA(1,1,0) model was the best for the other launch years. However, as noted above, there are valid reasons for preferring the lowest degree of differencing required for stationarity. Since the Dickey–Fuller test indicated stationarity in the single-differenced series, we prefer the simpler ARIMA(1,1,0) model for Ohio even for the 1950 launch year because it has the smallest BIC of the two single-order difference models tested.

Based on the stationarity test, the BIC, and our analysis of the ACF and PACF, we constructed a hybrid model by choosing the best individual model from Models 1–5 for each state and launch year. We also performed the Portmanteau test (Granger and Newbold 1986) on the residuals and found that the random residuals "white noise" requirement was satisfied for the identified models. This model identification process indicated that relatively few specifications were needed for the 12 state/launch year combinations. For the logarithmic series, the ARIMA ln(0,1,1) model fit 11 of 12 state and launch years. For the untransformed series, the hybrid model consisted of an:

- ARIMA(1,1,0) model for all launch years in Ohio and Maine and the 1960 and 1970 launch years in Wyoming;
- ARIMA(0,2,1) model for all launch years in Florida; and
- ARIMA(0,2,2) model without a constant for the 1950 launch year in Wyoming.

Forecasts and prediction intervals

We applied each of the six individual models to each state using three launch years (1950, 1960, and 1970), base periods of five lengths (10, 20, 30, 40, and 50 years),

and forecast horizons of three lengths (10, 20, and 30 years). This allowed us to analyze a wide range of values for each of variable and a large number of combinations of launch year, base period, and forecast horizon. With the addition of the hybrid model, this approach gave us a total of 315 point forecasts and associated prediction intervals for each state ($7 \times 3 \times 5 \times 3$).

We compared each point forecast to the population count for the relevant target year. We refer to the resulting percentage differences as *forecast errors*, although they may have been caused partly by errors in the population counts themselves. Following Lee and Tuljapurkar (1994), we express the size of the prediction interval as a *half-width* by dividing one-half of the difference between the upper and lower ends of the interval by the point forecast and multiplying the result by 100. For symmetrical intervals, the half-width reflects the percentage distance between the point forecast and the lower and upper bounds of the prediction interval. We calculated half-widths for both 95% and 68% prediction intervals, but report only the latter. In this study, then, a half-width of 15% means that about two-thirds of future populations are expected to fall within plus or minus 15% of the point forecast.

Forecast errors were measured in two ways. The mean absolute percentage error (MAPE) is the average when the direction of error is ignored; it is a measure of precision, or how close the forecasts were to out-of-sample population counts regardless of whether they were too high or too low. The mean algebraic percentage error (MALPE) is the average when the direction of error is accounted for; it is a measure of bias, or the tendency of forecasts to be too high or too low. Both measures have been used frequently in evaluations of population forecast accuracy (e.g., Ahlburg 1992; Keilman 1999; Pflaumer 1992; Rayer 2004; Smith and Sincich 1992).

To simplify the analysis, we started by aggregating the point forecasts and half-widths from each of the four states and three launch years. We calculated the average errors and half-widths of these 12 forecasts for each combination of model, base period, and forecast horizon. Finally, we evaluated the results separately for each state and launch year.

Results averaged over all states and launch years

The results averaged over all states and launch years are shown in Tables 3–5. Several patterns stand out. We first discuss those related to the six individual ARIMA models and then those related to the hybrid model.

The results for Models 1 and 2 were very similar. For every combination of base period and forecast horizon, the MAPEs, MALPEs, and half-widths were almost the same for Model 2 as for Model 1. It appears that the inclusion or omission of the first-order autoregressive term or the moving average term had little impact on the resulting forecasts; similar results were reported by Voss et al. (1981).

MAPEs and MALPEs for Models 3 and 4 were similar to each other, but half-widths were not. For every combination of base period and forecast horizon, half-widths were much larger for Model 3 than Model 4. Apparently, differences in the specification of these two nonlinear-growth models had little impact on precision and bias but had a substantial impact on the measurement of uncertainty.

Table 3 All states and launch years: mean absolute percentage error (MAPE) by model, length of base period, and length of forecast horizon

	Horizon length	Base period length 10	20	30	40	50
Model 1: (1,1,0)	10	9.2	9.9	9.9	9.9	10.1
	20	13.7	14.8	14.8	15.1	15.8
	30	15.9	17.4	17.7	18.2	18.6
Model 2: (0,1,1)	10	9.5	9.9	10.1	10.3	10.7
	20	13.7	14.8	14.9	15.5	16.4
	30	16.4	18.0	18.4	18.9	19.4
Model 3: (2,2,0)	10	19.0	13.3	12.7	12.4	11.9
	20	36.0	20.8	19.5	16.8	15.5
	30	56.0	30.7	28.2	22.7	20.2
Model 4: (0,2,1)	10	18.7	14.6	13.7	11.9	11.5
	20	41.3	26.7	20.3	17.2	16.8
	30	67.9	41.8	27.9	21.3	19.2
Model 5: (0,2,2)[a]	10	10.9	12.7	12.7	12.7	11.6
	20	16.1	19.4	19.6	19.8	18.0
	30	16.5	20.4	20.2	20.2	17.9
Model 6: ln(0,1,1)	10	10.7	11.0	8.9	8.7	9.3
	20	16.3	17.1	13.0	12.5	13.2
	30	24.7	25.0	18.0	16.8	19.0
Model 7: Hybrid	10	11.9	11.1	10.7	9.8	9.7
	20	18.6	16.1	15.5	13.9	13.8
	30	19.8	16.0	14.6	12.8	12.5

[a] Model does not include a constant term

MAPEs and half-widths were generally smallest for Models 1 and 2 and largest for Models 3 and 4, with those for Models 5 and 6 falling somewhere in between. These results were found for almost every combination of base period and forecast horizon. The linear-growth models thus tended to produce more precise forecasts and narrower prediction intervals than the nonlinear-growth models. For MAPEs, differences were often very large for 10-year base periods but became steadily smaller as the base period increased; at 50 years, differences were fairly small. For half-widths, differences were often quite large for all base periods, especially for longer forecast horizons.

The impact of the length of the base period on forecast precision varied by model. For Models 1, 2, and 5, MAPEs generally *increased* with increases in the base period, whereas for Models 3, 4, and 6 they generally *declined*. These results were found for all three forecast horizons. In this sample, then, 10 years of base data were sufficient to obtain maximum precision for Models 1, 2, and 5 but for Models 3, 4, and 6 precision increased continuously as more base data were added (although

Table 4 All states and launch years: mean algebraic percentage error (MALPE) by model, length of base period, and length of forecast horizon

	Horizon length	Base period length				
		10	20	30	40	50
Model 1: (1,1,0)	10	−3.6	−4.3	−5.4	−6.1	−6.3
	20	−4.8	−6.0	−7.9	−9.1	−9.5
	30	−6.4	−7.7	−10.2	−11.8	−12.4
Model 2: (0,1,1)	10	−3.6	−4.7	−6.0	−6.9	−7.2
	20	−4.9	−6.4	−8.5	−9.8	−10.4
	30	−6.5	−8.2	−10.8	−12.5	−13.1
Model 3: (2,2,0)	10	7.3	5.2	4.2	3.6	2.8
	20	22.2	15.9	12.5	10.5	8.5
	30	36.6	26.5	19.9	15.8	12.3
Model 4: (0,2,1)	10	7.7	7.3	4.0	1.8	1.0
	20	26.5	21.1	12.5	7.4	5.3
	30	46.7	35.4	20.7	12.0	8.4
Model 5: (0,2,2)[a]	10	−0.5	0.9	0.1	0.3	−0.9
	20	0.7	3.7	2.4	2.9	0.7
	30	0.0	4.3	2.5	3.2	0.2
Model 6: ln(0,1,1)	10	0.6	0.6	−0.4	−0.7	0.1
	20	6.4	6.2	4.1	3.5	5.3
	30	14.8	14.3	10.4	8.9	11.5
Model 7: Hybrid	10	−1.4	0.8	−0.8	−1.3	−1.9
	20	−0.5	4.1	1.1	0.1	−1.0
	30	−1.0	5.3	1.1	−0.4	−1.8

[a] Model does not include a constant term

the increases became steadily smaller as the base period became longer). It is noteworthy that Models 1 and 2 are linear-growth models and, although Model 5 is nonlinear, it is more nearly linear than Models 3, 4, and 6. It appears that models with first-order differences can be estimated fairly accurately using a relatively short data series, whereas models with second-order differences require a considerably longer series.

Increasing the length of the base period reduced the size of the half-width for all six models and for every length of forecast horizon, indicating that additional base data reduced the uncertainty associated with population forecasts (or, at least, additional data reduced this measure of uncertainty). The reductions were particularly great for Models 3 and 4, especially for longer horizons. It should be noted, however, that narrower prediction intervals are not necessarily better; what matters is how well they measure forecast uncertainty (Chatfield 2000). We will investigate this issue later in the article.

Table 5 All states and launch years: average half-width of the 68% prediction interval by model, length of base period, and length of forecast horizon

	Horizon length	Base period length 10	20	30	40	50
Model 1: (1,1,0)	10	7.7	7.2	7.0	6.4	5.9
	20	12.2	10.9	10.4	9.4	8.7
	30	15.7	13.7	12.9	11.6	10.6
Model 2: (0,1,1)	10	7.0	6.2	5.8	5.1	4.7
	20	10.8	9.2	8.3	7.3	6.6
	30	13.8	11.6	10.2	8.9	8.0
Model 3: (2,2,0)	10	28.4	22.9	21.2	18.0	16.0
	20	75.1	56.1	51.7	43.5	38.4
	30	135.9	92.0	86.0	71.8	63.2
Model 4: (0,2,1)	10	20.8	15.2	12.8	11.0	10.0
	20	52.2	34.2	29.0	24.8	22.3
	30	96.1	53.3	46.2	39.7	35.7
Model 5: (0,2,2)[a]	10	15.2	13.1	11.2	9.5	7.6
	20	32.9	27.4	23.1	19.3	14.9
	30	51.5	41.4	35.0	29.1	22.1
Model 6: ln(0,1,1)	10	9.7	8.9	8.5	7.8	7.5
	20	17.2	14.7	13.6	12.2	11.7
	30	25.3	20.4	18.5	16.3	15.4
Model 7: Hybrid	10	13.1	9.5	8.0	7.1	6.3
	20	29.1	17.0	13.8	12.1	10.7
	30	54.3	23.5	18.7	16.3	14.3

[a] Model does not include a constant term

Models 1 and 2 had a negative bias whereas Models 3 and 4 had a positive bias. This result was found for every combination of base period and forecast horizon. That is, Models 1 and 2 tended to produce forecasts that were too low and Models 3 and 4 tended to produce forecasts that were too high. Furthermore, the absolute value of the MALPEs for these four models became larger as the horizon increased—both when MALPEs were positive and when they were negative—indicating that the magnitude of the bias increased with the length of the forecast horizon. Models 5 and 6, on the other hand, displayed relatively little bias. The only exception was the 30-year horizon for Model 6, which displayed a fairly substantial upward bias.

The impact of the length of the base period on MALPEs varied by model. For Models 1 and 2, increasing the base period exacerbated the downward bias of the forecasts. For Models 3 and 4, increasing the base period reduced the upward bias. For Models 5 and 6, increasing the base period had no consistent impact on bias. As we note below, these results were not found for all individual states and launch years.

Both MAPEs and half-widths increased steadily with the length of the forecast horizon. This result was found for every model and every base period. This is not surprising, of course: Longer horizons create greater uncertainty and larger errors because they provide more opportunities for growth to deviate from predicted trends. Similar results have been reported in many other studies (e.g., Cohen 1986; Rayer 2004; Smith and Sincich 1992; Voss et al. 1981).

The hybrid model (Model 7) performed very well on tests of precision and bias, especially when the base period was at least 20 years long. MAPEs were generally similar to those found in the individual model with the smallest MAPE; sometimes, they were smaller than for any individual model. Half-widths were generally larger than those found for Models 1 and 2 but smaller than those found for Models 3 and 4. MALPEs were very small, indicating a low degree of bias; in most instances, MALPEs from the hybrid model were smaller (in absolute value) than MALPEs from any individual model. Furthermore, the hybrid model showed no consistent direction of bias, as MALPEs were sometimes positive and sometimes negative. MAPEs and half-widths generally increased with the length of the forecast horizon and declined with the length of the base period, but MALPEs displayed no consistent relationship with either variable.

Results by state

Several patterns are apparent when errors and half-widths from the four states and three launch years are averaged together. Do the same patterns appear when averages of the three launch years are calculated separately for each state? For the most part they do, but not in every instance. Detailed results of this analysis are available from the authors upon request; we summarize them here.

For every state and all combinations of forecast horizon and base period, MAPEs for Model 1 were very similar to those for Model 2. For these two models, Maine had substantially smaller MAPEs than any other state and Florida had the largest. In terms of precision, then, the two linear-growth models performed best in a state with a slow, steady growth rate and worst in a state with a high, volatile growth rate.

MAPEs for Models 3–6 differed considerably from each other. Model 3 had notably smaller MAPEs than the other nonlinear models in Florida and Model 6 had notably smaller MAPEs in Maine. Models 3 and 4 had particularly large MAPEs in Maine and Wyoming. For all six models in every state, MAPEs increased almost monotonically with the length of the forecast horizon.

For Models 1 and 2, increasing the length of the base period reduced MAPEs slightly in Maine, Wyoming, and Ohio. In Florida, however, increasing the length of the base period consistently *raised* MAPEs for both models. For Models 3 and 4, increasing the length of the base period generally reduced MAPEs in all four states, especially for longer forecast horizons. For Models 5 and 6, increases in the base period sometimes raised MAPEs and other times reduced them. For linear models, then, increasing the base period beyond 10 years did not improve precision appreciably (and sometimes made it substantially worse), whereas for nonlinear models it generally improved precision.

For every state, model, and forecast horizon, MALPEs for Models 1 and 2 were very similar to each other. For Models 3–6, however, MALPEs differed considerably. The direction of bias varied by model and by state. In most instances, MALPEs for Models 1 and 2 had negative signs for Maine, Wyoming, and Florida, and positive signs for Ohio. This most likely reflects the fact that the first three states had larger population increases in the second half of the century than the first half, whereas Ohio had larger increases in the first half. Models 3–5 had predominantly positive signs in every state except Florida and Model 6 had predominantly positive signs in Florida and Ohio and predominantly negative signs in Maine and Wyoming.

Bias generally increased with the length of the forecast horizon. If MALPEs were positive for 10-year horizons, they generally became larger positive numbers as the horizon increased. If they were negative for 10-year horizons, they generally became larger negative numbers as the horizon increased. This occurred for almost all combinations of state, model, and base period. The only exception was when MALPEs for 10-year horizons were close to zero; in these instances, the impact of increases in length of horizon was inconsistent across states, models, and base periods.

The impact of changes in length of base period on bias varied by state and model. For Florida, increasing the base period increased the negative bias for Models 1, 2, 3, and 5 (sometimes substantially), but reduced bias for Models 4 and 6. For Wyoming, it substantially reduced the upward bias for Models 3 and 4 but had little effect for the other models. For Ohio, it generally reduced upward bias, but the effects were very small for some models and horizons. For Maine, it sometimes exacerbated the upward or downward biases, but generally had little impact. There does not appear to be any consistent relationship between the length of the base period and the tendency for forecasts to be too high or too low.

For all states, horizons, and base periods, half-widths for Models 1 and 2 were similar to each other, but the degree of similarity was not quite as high as it was for MAPEs. In most instances, half-widths for Models 1 and 2 were smaller (sometimes substantially so) than half-widths for Models 3–6. Model 3 generally had the largest half-widths of all six models. For all states, models, and base periods, half-widths increased with the length of the horizon; in many instances, the increases were quite large.

For most combinations of state, model, and horizon, increasing the base period reduced the half-width. The reductions were especially large for Models 3 and 4, and were generally greater for longer horizons than for shorter horizons. The only exception was Model 1 in Florida, where increasing the length of the base period slightly raised the size of the half-width.

The hybrid model performed reasonably well in every state. MAPEs were generally moderate, falling somewhere between the smallest and the largest of the six individual models; they were usually closer to the smallest than the largest and in some instances were smaller than for any individual model. Half-widths for the hybrid model were also fairly moderate, increasing with the length of forecast horizon in every state and falling with increases in the base period in every state except Ohio. MALPEs were uniformly negative in Florida and uniformly positive in the other three states (except for the 10-year base period in Maine); the extent of the

downward bias in Florida was considerably greater than the extent of the upward bias in the other states.

Results by launch year

How similar are the results from one launch year to another? To answer this question, we calculated the average errors and half-widths of the state forecasts for each of the three launch years, by model, length of base period, and length of forecast horizon. Again, detailed results are available from the authors upon request; we summarize them here.

For every launch year and every combination of model, base period, and forecast horizon, MAPEs for Model 1 were very similar to those for Model 2. For 1950, Models 1, 2, and 6 generally had smaller MAPEs than Models 3–5. For short base periods and long horizons, the differences were sometimes very large. For 1960, Models 1 and 2 generally had the smallest MAPEs for 10-, 20-, and 30-year base periods, but for 40- and 50-year base periods MAPEs were similar for all six individual models. For 1970, MAPEs were roughly similar for all six models for short base periods, but for long base periods MAPEs were generally smaller for Models 3 and 4 than for the other models. In most instances, MAPEs increased with the forecast horizon for all six individual models in all three launch years.

For Models 1 and 2, MAPEs for launch years 1950 and 1970 were similar to each other but were somewhat larger than MAPEs for 1960. Models 5 and 6 also had generally larger MAPEs for 1950 and 1970 than for 1960, but did not display as much consistency from one launch year to another as Models 1 and 2. For Models 3 and 4, MAPEs were generally largest for 1950 and smallest for 1970. Differences in MAPEs from one launch year to another were considerably larger for Models 3 and 4 than for the other models. Overall, the linear-growth models displayed somewhat more consistency from one launch year to another than did the nonlinear-growth models.

MALPEs varied considerably from one launch year to another. For Models 1 and 2, MALPEs were negative for every combination of base period and forecast horizon for launch years 1950 and 1970, but were positive for a number of combinations for 1960. For Models 3 and 4, MALPEs were positive for most base-horizon combinations for 1950 and 1960, but were negative for a number of combinations for 1970. Models 5 and 6 displayed decidedly mixed results. As has been noted before, bias appears to vary substantially (and unpredictably) from one launch year to another (e.g., Rayer 2004; Smith and Sincich 1992). Lengthening the base period had no consistent impact on the results, sometimes raising MALPEs and other times reducing them.

In all three launch years, half-widths were similar for Models 1 and 2 for most base period/forecast horizon combinations. In every combination, they were smaller (usually much smaller) than the corresponding half-widths for Models 3 and 4. Half-widths for Models 5 and 6 generally fell somewhere between those for Models 1 and 2 and those for Models 3 and 4. There was no consistent relationship between the launch year and the size of the half-width; sometimes they were largest for 1950, sometimes for 1960, and sometimes for 1970. For every combination of model, base

period, and launch year, half-widths increased monotonically with the length of the forecast horizon. In most combinations of model, forecast horizon, and launch year, half-widths declined as the base period increased.

In most instances, then, the results regarding the effects of differences in base period, forecast horizon, and model on forecast errors and half-widths were about the same for each individual launch year as they were when the launch years were aggregated. The values of the errors and half-widths themselves, however, often differed considerably from one launch year to another.

We also evaluated the results for the hybrid model (Model 7) for each launch year. MAPEs from the hybrid model typically were similar to those found for the most precise individual model and sometimes were smaller than for any of the individual models. The hybrid model itself generally had the largest MAPEs for 1950 and the smallest for 1960. For most combinations of base period and forecast horizon, the hybrid model exhibited a downward bias in 1950 and 1970 and an upward bias in 1960; however, the magnitude of the bias was generally fairly small. Half-widths were typically larger than those found for Models 1 and 2 but smaller than those found for Models 3 and 4.

A test of prediction intervals

Many different time series models could be constructed using different base periods, launch years, and sets of assumptions; each implies a different set of prediction intervals for each forecast horizon. How well do the models analyzed in this study perform in terms of predicting the uncertainty of future population growth?

One way to address this question is to calculate the number of population counts falling inside the prediction intervals associated with each set of forecasts (e.g., Cohen 1986; Keyfitz 1977; Smith and Sincich 1988; Swanson and Beck 1994; Voss et al. 1981). Table 6 shows the calculations for each combination of model, base period, and forecast horizon for forecasts aggregated across all states and launch years. Each cell is based on 12 forecasts (four states and three launch years). If the 68% prediction intervals provide valid measures of uncertainty, they will encompass approximately eight of the 12 out-of-sample population counts. Cells in which between seven and nine counts fell within the prediction interval are highlighted in the table.

According to this criterion, the prediction intervals for Models 1–3 did not provide valid measures of uncertainty. The intervals associated with Models 1 and 2 were too narrow. In no set of forecasts did more than six of the 12 population counts fall inside the predicted interval; in some sets, only two or three fell inside. In this study, then, the two linear-growth models consistently underestimated uncertainty. Similar results were reported by Voss et al. (1981).

Differences in the length of the forecast horizon did not have much effect on the number of counts falling inside the prediction intervals for Models 1 and 2, but increasing the length of the base period generally led to a smaller number falling inside. The latter result is particularly noteworthy, as it suggests that fewer observations in the base period *improved* the validity of this measure of uncertainty for linear-growth models.

Table 6 All states and launch years: number of population counts falling within the 68% prediction interval by model, length of base period, and length of forecast horizon

	Horizon length	Base period length				
		10	20	30	40	50
Model 1: (1,1,0)	10	6	5	6	5	4
	20	5	3	5	4	2
	30	5	4	5	5	4
Model 2: (0,1,1)	10	5	5	4	4	4
	20	6	3	3	2	2
	30	5	4	5	3	3
Model 3: (2,2,0)	10	7	10	10	10	9
	20	10	11	11	11	11
	30	11	12	12	12	12
Model 4: (0,2,1)	10	6	6	5	7	7
	20	7	9	8	9	8
	30	6	8	9	10	11
Model 5: (0,2,2)[a]	10	5	6	6	6	5
	20	7	6	7	6	5
	30	7	7	9	9	8
Model 6: ln(0,1,1)	10	6	4	7	7	5
	20	8	6	8	8	6
	30	7	7	8	8	7
Model 7: Hybrid	10	7	5	5	6	5
	20	6	4	6	5	4
	30	7	7	8	8	7

[a] Model does not include a constant term

The prediction intervals associated with Model 3 were too wide. In all but two sets of forecasts, 10, 11, or all 12 counts fell inside the predicted intervals. This model consistently overestimated uncertainty, especially for longer forecast horizons.

Models 4–7 performed considerably better than Models 1–3. For 10-year horizons, between four and seven forecasts fell inside the prediction intervals; for 20-year horizons, between five and nine; and for 30-year horizons, between six and eleven. Model 4 performed particularly well for the 20-year horizon and Models 5–7 performed particularly well for the 30-year horizon. Although this sample is too small to support general conclusions, these results suggest that some time series models may produce prediction intervals that provide fairly realistic measures of forecast uncertainty.

For the four nonlinear-growth models, the length of the base period did not have a consistent impact on the number of counts falling inside the prediction intervals. For Models 3 and 4, the number increased slightly with the length of the base

period, but for Models 5 and 6 it followed no clear pattern, sometimes rising and sometimes falling. The reduction in uncertainty associated with increases in the length of the base period noted in Table 5 for nonlinear models was thus supported by the empirical evidence: the intervals became smaller, but the number of counts falling inside them remained about constant.

Increasing the length of the forecast horizon had a small but generally positive effect on the number of counts falling inside the prediction intervals for the four nonlinear-growth models. This suggests that these models may produce prediction intervals that grow more rapidly than the true degree of uncertainty as the forecast horizon becomes longer. More research is needed before we can draw firm conclusions on this point.

Summary and conclusions

Time series models have several drawbacks compared to most extrapolation methods used for population forecasting. They are considerably more complex and difficult to apply than simpler methods. Specifying them correctly requires a high level of expertise and a substantial time commitment; this can be burdensome when models must be specified for a large number of geographic areas (e.g., states or counties). Furthermore, they have not been found to produce more accurate forecasts than simpler methods (Smith et al. 2001). However, they offer one major advantage compared to simpler methods: they provide prediction intervals to accompany their point forecasts. This characteristic has led to a substantial amount of research on time series models over the last several decades. Very little of this research, however, has considered subnational population forecasts or attempted to evaluate out-of-sample forecast accuracy or the empirical validity of prediction intervals.

In this study, we developed six individual ARIMA time series models reflecting a variety of population growth trajectories, applied them using data from four states in the United States, and evaluated the accuracy of the resulting population forecasts. To address the model misspecification that may occur when observations are combined across states, we also developed a hybrid model (Model 7) based on the best individual model (using untransformed data) for each state and launch year. The best individual model was identified by analyzing autocorrelation and partial autocorrelation functions and applying Dickey–Fuller, Bayesian Information Criteria, and Portmanteau statistical tests.

Using data for a variety of launch years, base periods, and forecast horizons between 1900 and 2000, we constructed 315-point forecasts and sets of prediction intervals for each state. We evaluated these forecasts by comparing them to population counts for the corresponding target years. Although the evidence was not always clear-cut, a number of distinct patterns emerged.

The two linear-growth models (Models 1 and 2) produced forecasts that differed very little from each other, leading to MAPEs, MALPEs, and half-widths that were much the same for both models. It appears that the presence or absence of first-order autoregressive and moving average terms does not have much impact on forecasts

from linear-growth models. Similar results have been reported before (Alho 1990; Voss et al. 1981).

For linear-growth models, 10 years of base data were generally sufficient to achieve—or at least come close to—maximum forecast precision (i.e., the smallest MAPEs). This result was found for every state and launch year, even for 20- and 30-year forecast horizons. Similar results for simple extrapolation techniques have been reported before (e.g., Rayer 2004; Smith and Sincich 1990). Although longer base periods may be desirable for other purposes, they do not appear to be necessary for maximizing the precision of forecasts based on linear-growth models.

Models 3 and 4 are nonlinear-growth models. For these models, 10 years of base data generally were *not* sufficient to achieve—or even come close to—maximum forecast precision. Rather, MAPEs declined as base periods increased, albeit at a generally diminishing rate. Typically, the declines were greater for longer horizons than shorter horizons. Although Models 1 and 2 produced more precise forecasts than Models 3 and 4 when the base periods were relatively short, their superiority diminished as the base period increased.

Models 5 and 6 are also nonlinear-growth models, but are not as explosive as Models 3 and 4. Consequently, the impact of increases in the base period for these two models was more nearly similar to that observed for Models 1 and 2 than to that observed for Models 3 and 4. For Model 5, MAPEs increased slightly with increases in the base period, whereas for Model 6 they declined slightly. As was true for Models 1 and 2, longer base periods did not have much impact on forecast precision for these two models.

On average, the linear-growth models had a negative bias and the nonlinear-growth models had a positive bias. However, this result was not found for every state and launch year. Given this finding and the fact that bias has been found to vary considerably from one launch year to another (e.g., Rayer 2004; Smith and Sincich 1988, 1992), we do not believe there is enough evidence to draw any general conclusions regarding the bias inherent in different types of time series forecasting models. We note that changes in the length of the base period had no consistent impact on bias for either type of model; this result has also been reported before (Rayer 2004; Smith and Sincich 1990).

The hybrid model performed very well on tests of precision and bias. Its MAPEs were generally similar to—and sometimes smaller than—the smallest MAPE found for any individual model. In most instances, its MALPEs were smaller (in absolute value) than those found for individual models. Furthermore, the hybrid model showed no consistent direction of bias. These results were found in most analyses of individual states and launch years as well as when the data were aggregated across all states and launch years.

Models 1 and 2 produced the smallest half-widths and Model 3 produced the largest. This result was found for virtually every combination of state, launch year, base period, and forecast horizon. Small half-widths are not necessarily a positive characteristic of population forecasts, however; what matters is how well the prediction intervals derived from those half-widths measure forecast uncertainty. As shown in Table 6, the prediction intervals produced by Models 1 and 2 were considerably too narrow (i.e., too few counts fell inside) while those produced by

Model 3 were considerably too wide (i.e., too many fell inside). Models 4–6 and the hybrid model were more successful in producing realistic prediction intervals than Models 1–3; for these models, the number of counts falling inside the intervals was often close to the predicted number. These results suggest that some specifications of time series models may provide fairly realistic measures of forecast uncertainty but others do not.

In our view, the major advantage of time series models compared to simpler extrapolation methods is their ability to produce prediction intervals to accompany point forecasts. If those prediction intervals cannot provide realistic measures of uncertainty, much of the potential value of time series models is lost. Given the results of the present study, we believe there is some basis for optimism regarding the possibility that time series models might be able to produce realistic prediction intervals. Although some models performed poorly in this regard, others (particularly Models 6 and 7) performed very well.

Many different time series forecasting models can be specified, each providing a different set of point forecasts and prediction intervals (e.g., Cohen 1986; Keilman et al. 2002; Lee 1974; Sanderson 1995). These models are subject to errors in the base data, errors in specifying the model, errors in estimating the model's parameters, and structural changes that invalidate the model's statistical relationships over time (Lee 1992). It is therefore not surprising that the models examined in this study produced widely differing values for MAPEs, MALPEs, and half-widths and provided mixed results regarding the number of population counts falling within the predicted intervals.

Clearly, the development of empirically valid prediction intervals based on time series models requires a significant investment of time and effort. The generally good performance of the hybrid model illustrates the importance of using best practices to identify the optimal model for a particular forecast. Applying an arbitrarily specified or a single unanalyzed model to a large number of states or local areas is not likely to produce useful results.

Many questions remain to be answered before we can draw any firm conclusions, however. Would results similar to those reported here be found for other states and time periods? Would different model specifications or evaluation criteria lead to different results? Can the use of location-specific characteristics improve the model selection process? Can ways other than the hybrid model described here be found for combining forecasts? What about forecasts made at the local level? How would their error characteristics compare to those shown here for states?

Research on other approaches to measuring uncertainty is also needed, such as basing prediction intervals on the distribution of errors in previous forecasts (e.g., Keyfitz 1981; Smith and Sincich 1988; Stoto 1983; Tayman et al. 2005), the application of expert judgment (e.g., Lutz et al. 1999), or some combination of approaches (e.g., Keilman et al. 2002). We believe further research on forecast accuracy and the measurement of uncertainty will be both intellectually interesting and practically useful, giving data users a more realistic understanding of the potential accuracy of population forecasts and helping decision makers plan more effectively for an uncertain future.

Acknowledgments The authors thank the two referees for their thoughtful comments and suggestions that greatly improved this paper.

References

Ahlburg, D. (1992). Error measures and the choice of a forecast method. *International Journal of Forecasting, 8*, 99–100.

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

Alho, J., & Spencer, B. (1997). The practical specification of the expected error of population forecasts. *Journal of Official Statistics, 13*, 203–225.

Box, G., & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden Day.

Brockwell, P., & Davis, R. (2002). *Introduction to time series and forecasting* (2nd ed.). New York, NY: Springer-Verlag.

Chatfield, C. (2000). *Time-series forecasting*. Boca Raton, FL: Chapman & Hall/CRC.

Cohen, J. (1986). Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches. *Demography, 23*, 105–126.

De Beer, J. (1993). Forecast intervals of net migration: The case of the Netherlands. *Journal of Forecasting, 12*, 585–599.

Dickey, D., Bell, W., & Miller, R. (1986). Unit roots in time series models: Tests and implications. *American Statistician, 74*, 427–431.

Granger, C. (1989). *Forecasting in business and economics* (2nd ed.). San Diego, CA: Academic Press.

Granger, C., & Newbold, P. (1986). *Forecasting economic time series* (2nd ed.). San Diego, CA: Academic Press.

Keilman, N. (1999). How accurate are the United Nations world population projections? In W. Lutz, J. Vaupel, & D. Ahlburg (Eds.), *Frontiers of population forecasting* (pp. 15–41). New York, NY: Population Council. Supplement to Vol. 24 of *Population and Development Review*.

Keilman, N., Pham, D., & Hetland, A. (2002). Why population forecasts should be probabilistic – illustrated by the case of Norway. *Demographic Research, 6*, 409–453.

Keyfitz, N. (1977). *Applied mathematical demography*. New York, NY: John Wiley.

Keyfitz, N. (1981). The limits of population forecasting. *Population and development review, 7*, 579–593.

Lee, R. (1974). Forecasting births in post-transition populations: Stochastic renewal with serially correlated fertility. *Journal of the American Statistical Association, 69*, 607–617.

Lee, R. (1992). Stochastic demographic forecasting. *International Journal of Forecasting, 8*, 315–327.

Lee, R., & Tuljapurkar, S. (1994). Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association, 89*, 1175–1189.

Lutz, W., Sanderson, W., & Scherbov, S. (1999). Expert-based probabilistic population projections. In W. Lutz, J. Vaupel, & D. Ahlburg (Eds.), *Frontiers of Population Forecasting* (pp. 139–155). New York, NY: Population Council. Supplement to Vol. 24 of *Population and Development Review*.

Makridakis, S., Wheelwright, S., & Hyndman, R. (1998). *Forecasting methods and applications* (3rd ed.). New York, NY: John Wiley.

McCleary, R., & Hay, R. (1980). *Applied time series for the social sciences*. Beverly Hills, CA: Sage Publications.

McNown, R., & Rogers, A. (1989). Forecasting mortality: A parameterized time series approach. *Demography, 26*, 645–660.

Meyler, A., Kenny, G., & Quinn, T. (1998). *Forecasting Irish inflation using ARIMA models*. Technical Paper 3/RT/98. Economic Analysis, Research, and Publication Department. Central Bank of Ireland.

Murdock, S., Leistritz, F., Hamm, R., Hwang, S., & Parpia, B. (1984). An assessment of the accuracy of a regional economic-demographic projection model. *Demography, 21*, 383–404.

Nelson, C. (1973). Applied time series analysis for managerial forecasting. San Francisco, CA: Holden Day.

Pflaumer, P. (1992). Forecasting U.S. population totals with the Box–Jenkins approach. *International Journal of Forecasting, 8*, 329–338.

Rayer, S. (2004). *Assessing the accuracy and bias of trend extrapolation methods for population projections: The long view*. Paper presented at the annual meeting of the Southern Demographic Association, Hilton Head SC.

Saboia, J. (1974). Modeling and forecasting populations by time series: The Swedish case. *Demography, 11*, 483–492.

Sanderson, W. (1995). Predictability, complexity, and catastrophe in a collapsible model of population, development, and environmental interactions. *Mathematical Population Studies, 5*, 259–279.

Smith, S., & Sincich, T. (1988). Stability over time in the distribution of population forecast errors. *Demography, 25*, 461–474.

Smith, S., & Sincich, T. (1990). The relationship between the length of the base period and population forecast errors. *Journal of the American Statistical Association, 85*, 367–375.

Smith, S., & Sincich, T. (1992). Evaluating the forecast accuracy and bias of alternative population projections for states. *International Journal of Forecasting, 8*, 495–508.

Smith, S., Tayman, J., & Swanson, D. (2001). *State and local population projections: Methodology and analysis*. New York, NY: Kluwer Academic/Plenum Publishers.

Stoto, M. (1983). The accuracy of population projections. *Journal of the American Statistical Association, 78*, 13–20.

Swanson, D., & Beck, D. (1994). A new short-term county population projection method. *Journal of Economic and Social Measurement, 20*, 25–50.

Tayman, J. (1996). The accuracy of small-area population forecasts based on a spatial interaction land-use modeling system. *Journal of the American Planning Association, 62*, 85–98.

Tayman, J., Rayer, S., & Smith, S. (2005). *Prediction intervals for county population forecasts*. Paper presented at the annual meeting of the Southern Demographic Association, Oxford MS.

Tayman, J., Schafer, E., & Carter, L. (1998). The role of population size in the determination and prediction of population forecast errors: An evaluation using confidence intervals for subcounty areas. *Population Research and Policy Review, 17*, 1–20.

U.S. Census Bureau. (1956). *Estimates of the population of states: 1900 to 1949*. Current Population Reports, Series P-25, No. 139 (Released on the Internet February 1996.)

U.S. Census Bureau. (1965). *Revised estimates of the population of states and components of population change 1950 to 1960*. Current Population Reports, Series P-25, No. 304 (Released on the Internet April 1995.)

U.S. Census Bureau. (1971). *Preliminary intercensal estimates of states and components of population change 1960 to 1970*. Current Population Reports, Series P-25, No. 460 (Released on the Internet August 1996.)

U.S. Census Bureau. (1984). *Intercensal estimates of states and components of population change 1970 to 1980.* Current Population Reports, Series P-25, No. 957 (Released on the Internet February 1995.)

U.S. Census Bureau. (1993). *Intercensal estimates of states and components of population change 1970 to 1980.* Current Population Reports, Series P-25, No. 1106 (Released on the Internet August 1996.)

U.S. Census Bureau. (2002). Table CO-EST2001-12-0 – Series of intercensal state population estimates: April 1, 1990 to April 1, 2000. (Released on the Internet April 2002.)

Voss, P., Palit, C., Kale, B., & Krebs, H. (1981). *Forecasting state populations using ARIMA time series techniques*. Report prepared by the Wisconsin Department of Administration and the University of Wisconsin-Madison.

White, H. (1954). Empirical study of the accuracy of selected methods of projecting state populations. *Journal of the American Statistical Association, 49*, 480–498.