Pflaumer, Peter

**Working Paper**

# Population forecasting with the Box-Jenkins approach

Diskussionsbeiträge - Serie II, No. 129

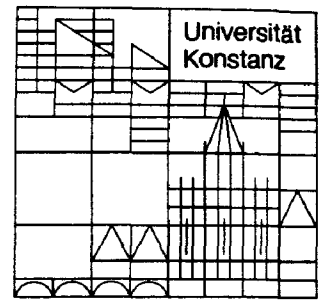
**Provided in Cooperation with:**
Department of Economics, University of Konstanz

*Suggested Citation:* Pflaumer, Peter (1991) : Population forecasting with the Box-Jenkins approach, Diskussionsbeiträge - Serie II, No. 129, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:
https://hdl.handle.net/10419/101596

**Sonderforschungsbereich 178**
**„Internationalisierung der Wirtschaft“**

Diskussionsbeiträge




Juristische
Fakultät

Fakultät für Wirtschafts-
wissenschaften und Statistik

---

Peter Pflaumer

# Population Forecasting with the Box-Jenkins Approach

---

Postfach 5560
D-7750 Konstanz

Serie II — Nr. 129
Februar 1991

# POPULATION FORECASTING WITH THE BOX-JENKINS APPROACH

**Peter Pflaumer**

Serie II - Nr. 129

Februar 1991

# POPULATION FORECASTING WITH THE BOX-JENKINS APPROACH

Peter Pflaumer

**Abstract:**

The use of the Box-Jenkins approach for forecasting the population of the United States to the year 2080 is discussed. The forecasts are based on data for 1900-1980. It is shown that no major difference exists between the Box-Jenkins approach and parabolic trend curves when making long-range predictions. An investigation of forecasting accuracy indicates that the Box-Jenkins method produces population forecasts that are at least as reliable as those done with more traditional demographic methods.

# INTRODUCTION

The Box-Jenkins approach has gained great popularity since the publication of their book in 1970. Applications can be found in many scientific fields, such as astronomy, economics or manangement science. Recently the Box-Jenkins approach has been considered for use in demography. For example, the reader might refer to articles of Saboia (1974), Lee (1974), Land/Cantor (1983) or El-Attar (1988). The utility of the Box-Jenkins model for forecasting U.S. state populations has been investigated by Voss/Palit (1981).
This paper presents the results of an investigation designed to analyse the utility of the Box-Jenkins technique for forecasting U.S. population totals. The following section is devoted to a brief outline of the theory, while the remainder deals with several forecasts and comments on the results.

# BOX-JENKINS FORECASTING PROCEDURE

This paper does not intend fully to describe the Box-Jenkins technique; only a brief outline of the procedure is given. The method is based on fitting an autoregressive moving average process (ARMA(p,q)) to a set of equally spaced observations $P_t, P_{t-1}, \ldots$ and then predicting the next values which will occur. The process has the form

$$P_t = \mu + \phi_1 P_{t-1} + \phi_2 P_{t-2} + \ldots + \phi_p P_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q},$$

where $\epsilon_t$ is a random variable with mean zero and variance $\sigma^2$. If the time series is not stationary in the mean, which will usually apply for demographic series, a difference transformation of order d may achieve stationarity. The underlying model, which has to be transformed, is called an autoregressive integrated moving average process (ARIMA(p,d,q)). The $d$ is the order of differencing necessary for stationarity.

The main steps in setting up a Box-Jenkins model are as follows:

1. Identification
2. Estimation
3. Diagnostic checking
4. Consideration of alternative models if necessary

The first step involves determining $p$, $d$ and $q$ – the order of the ARIMA-model. This procedure is usually done by means of autocorrelation functions. In the second step the parameters of the model are estimated by (nonlinear) least squares, while in the third step the residuals from the fitted model are examined to see if the chosen model is adequate. If the model is appropriate, forecasts are produced, if not, alternative models are considered (step 4).

# APPLICATIONS

This section will demonstrate the application of the Box-Jenkins technique using annual population figures $P_t$ of the U.S. from 1900 to 1980. The series had to be differenced until it appeared to be stationary. The examination of the correlograms of various differenced series indicated the necessity of differencing twice, since the correlogram of the first differenced series did not die out fast enough.

Figure 1: U.S. population (first differences) from 1900 to 1980

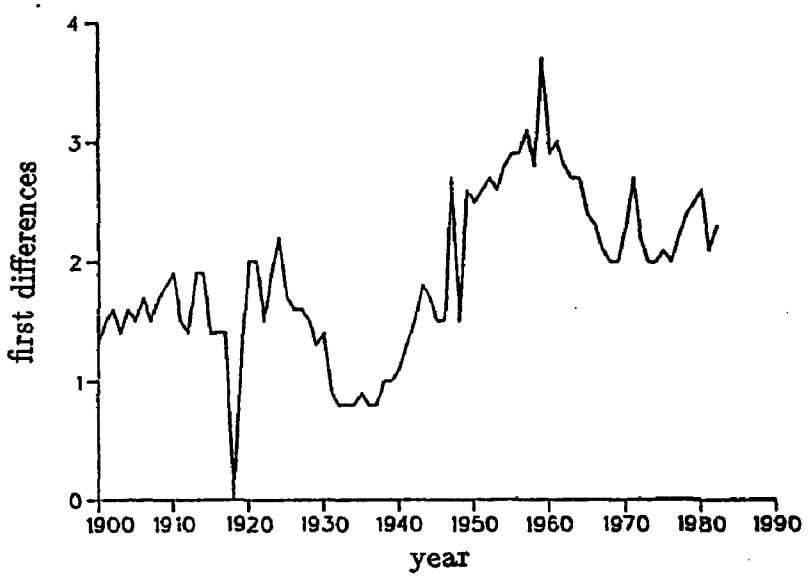



Figure 2: U.S. population (second differences) from 1900 to 1980

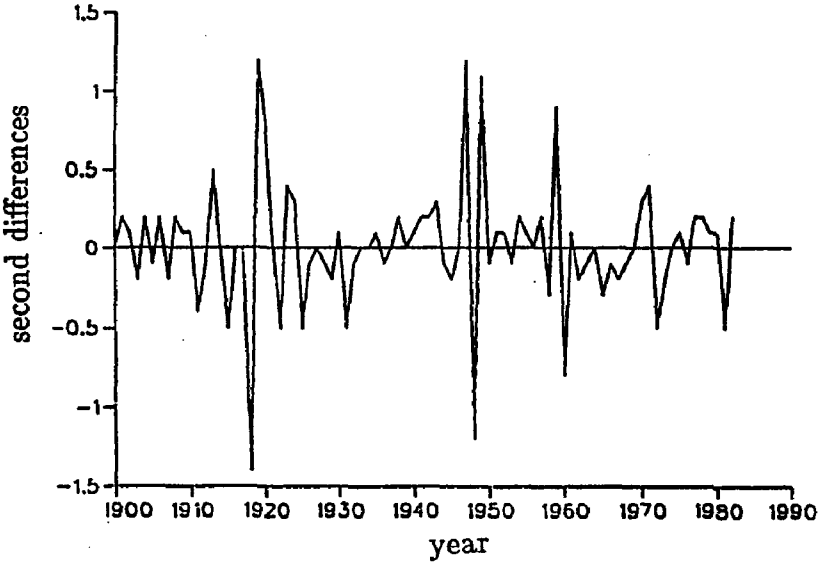


Figure 3 shows the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the second differences $\Delta^2 P_t$. The PACF is characterized by two negative spikes at lag 1 and 2 with values $r_1 = -0.22$ and $r_2 = -0.24$. The ACF is characterized by a damped sine wave. The behaviour of both correlograms is consistent with an ARIMA (2,2,0)-process for the population figures $P_t$. Having chosen an ARIMA-model to fit to the data, the next step is to estimate the parameters from the data. Applying least squares yields the following fitted equation :

$$\begin{array}{rcccccc} \Delta^2 P_t & = & 0.02 & - & 0.27\Delta^2 P_{t-1} & - & 0.24\Delta^2 P_{t-2} \\ & & & & (-2.41) & & (-2.11) \end{array}$$

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &= 0.35 \\ Q_6 &= 2.45 \\ Q_{24} &= 17.45 \end{aligned}$$

The estimation procedure provides not only point estimates for the parameters, but asymptotic standard deviations as well, so that classical hypothesis tests may be used. The t-ratios in parantheses indicate that the estimators are significant. The Box-Pierce statistics $Q_6$ and $Q_{24}$ give evidence that the residual is white noise. We therefore think that the selected ARIMA (2, 2, 0) -model provides a fairly adequate representation of the growth of the U.S. population between 1900 and 1980.

Figure 3: Autocorrelation and partial autocorrelation functions of $\Delta^2 P_t$

| Lag | Correlation | 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 |
|---|---|---|
| 0 | 1.00 | !********** |
| 1 | -0.22 | .****! . |
| 2 | -0.18 | .****! . |
| 3 | 0.02 | ! . |
| 4 | 0.17 | !*** . |
| 5 | -0.11 | **! . |
| 6 | -0.08 | **! . |
| 7 | 0.21 | !****. |
| 8 | -0.09 | **! . |
| 9 | -0.08 | **! . |
| 10 | 0.03 | !* . |
| 11 | -0.00 | ! . |
| 12 | 0.03 | !* . |
| 13 | -0.11 | **! . |
| 14 | 0.04 | !* . |
| 15 | 0.01 | ! . |
| 16 | -0.05 | *! . |
| 17 | -0.04 | *! . |
| 18 | -0.05 | *! . |
| 19 | 0.04 | !* . |
| 20 | -0.05 | *! . |
| 21 | -0.02 | ! . |
| 22 | -0.04 | *! . |
| 23 | 0.04 | !* . |
| 24 | 0.09 | !** . |

| Lag | Correlation | 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 |
|---|---|---|
| 1 | -0.22 | .****! . |
| 2 | -0.24 | *****! . |
| 3 | -0.09 | **! . |
| 4 | 0.12 | !** . |
| 5 | -0.05 | *! . |
| 6 | -0.07 | *! . |
| 7 | 0.17 | !*** . |
| 8 | -0.05 | *! . |
| 9 | -0.04 | *! . |
| 10 | 0.00 | ! . |
| 11 | -0.09 | **! . |
| 12 | 0.06 | !* . |
| 13 | -0.09 | **! . |
| 14 | -0.05 | *! . |
| 15 | 0.02 | ! . |
| 16 | -0.06 | *! . |
| 17 | -0.05 | *! . |
| 18 | -0.10 | **! . |
| 19 | -0.07 | *! . |
| 20 | -0.04 | *! . |
| 21 | -0.06 | *! . |
| 22 | -0.10 | **! . |
| 23 | 0.01 | ! . |
| 24 | 0.10 | !** . |

Since $\Delta^2 P_t$ can satisfactorily be described by an ARMA (2,0)-process, it is easy to show that the predictions of the original values $P_t$ are obtained by

$$P_t = 0.02 + 1.73P_{t-1} - 0.7P_{t-2} + 0.21P_{t-3} - 0.24P_{t-4},$$

which is a fourth order difference equation, whose sum of coefficients is zero.

The results of the projection with the Box-Jenkins model are shown in Table 1. The U.S. population will continue to grow. In 2080 a total population of 551 million will be reached. Placing a confidence interval around this point forecast, it is assumed that the population will be in the range 285 to 872 million with a probability of 95 percent.

It is interesting to compare these figures with the projection results of the U.S. Bureau of the Census (see Table 2).Their middle projection leads to a population of 310.8 million in the year 2080, and the range between their lowest and highest series is from 191.1 to 531.2 million.

Table 1: Results of the Box-Jenkins Forecasts with 95%-Confidence Intervals (Numbers in Millions)

| Year | Lower Limit of a 95% C.I. | Point Forecast | Upper Limit a 95% C.I. |
|---|---|---|---|
| 1990 | 244.4 | 254.0 | 263.6 |
| 2000 | 256.4 | 281.7 | 306.9 |
| 2010 | 265.3 | 310.6 | 355.0 |
| 2030 | 277.0 | 372.6 | 468.3 |
| 2050 | 282.9 | 440.0 | 597.1 |
| 2080 | 284.5 | 551.1 | 871.6 |

Table 2: U.S.Census Bureaus's Population Projections (Numbers in Millions)

| Year | Lowest Series | Middle Series | Highest Series |
|---|---|---|---|
| 1990 | 245.8 | 249.7 | 254.1 |
| 2000 | 256.1 | 268.0 | 281.5 |
| 2010 | 261.5 | 283.2 | 310.0 |
| 2030 | 257.4 | 304.8 | 369.8 |
| 2050 | 232.2 | 309.5 | 427.9 |
| 2080 | 191.1 | 310.8 | 531.2 |

Source: U.S. Bureau of the Census (1984)

Regarding the first and second differences in Figure 1 and Figure 2, it seems appropriate to describe the growth of the U.S. population by a parabolic trend curve in the form

$$P_t = a + bt + ct^2,$$

since the first derivative is a linear and the second derivative of $P_t$ is a constant function of time.

We will now show that the chosen Box-Jenkins model is equivalent to a parabolic trend model, when making long-term population forecasts. Considering that the above Box-Jenkins model of $P_t$ is a fourth order (stochastic) difference equation, it is possible to reduce the solution of this model to the solution of an algebraic equation, which is called the characteristic equation of the difference equation. In that special case we obtain

$$\lambda^4 - 1.73\lambda^3 + 0.7\lambda^2 - 0.21\lambda + 0.24 = 0,$$

with its roots

$$\begin{aligned} \lambda_1 = \lambda_2 &= 1 \\ \lambda_3 &= 0.135 + 0.4709i \\ \lambda_4 &= 0.135 - 0.4709i. \end{aligned}$$

This equation has repeated real roots and conjugate complex numbers. Because of the existence of repeated roots the general solution is

$$P_t = C_1(0.135 + 0.4709i)^t + C_2(0.135 - 0.4709i)^t + (C_3 + tC_4)1^t,$$

where $C_1$, $C_2$, $C_3$ and $C_4$ are arbitary constants. In order to find the particular solution, we set $P = zt^2$,which leads to

$$zt^2 - 1.73z(t-1)^2 + 0.7z(t-2)^2 + 0.21z(t-3)^2 + 0.24z(t-4)^2 = 0.02$$

$$\text{or} \quad z = 0.0066225.$$

Given $P_0 = 220.2$, $P_1 = 222.6$, $P_3 = 251.1$, and $P_4 = 227.7$, the definite solution can be written as

$$\begin{aligned} P_t = \; & (0.114 + \tfrac{0.024}{i})(0.135 + 0.4709i)^t + (0.114 - \tfrac{0.024}{i}) \\ & \cdot(0.135 - 0.4709i)^t + 220 + 2.56t + 0.0066225t^2 \\ & \text{for} \quad t = 0, 1, 2, \ldots \quad . \end{aligned}$$

Transforming into its trigonometric form leads to

$$P_t = 0.49^t(0.228\cos 1.29t + 0.048\sin 1.29t) + 220 + 2.56t + 0.0066225t^2$$

$$\text{for} \quad t = 0, 1, 2, \ldots \quad .$$

The resulting time path of $P_t$ is a damped oscillation around its particular solution, which is a parabolic trend curve. Already after a few years the oscillations can be forecasted solely by its particular solution.

# FORECASTING ACCURACY

The question here is whether the Box-Jenkins approach is capable of predicting future populations accurately. The accuracy was tested by comparing past hypothetical population forecasts of the U.S. with subsequent outcomes. This was achieved by fitting an

ARIMA-model to the first $n$ values of the population series and calculating the yearly growth rate error $e_t$ as follows

$$e_t = (\hat{r}_t - r_t),$$

where $r$ is the actual annual and $\hat{r}$ is the forecasted annual growth rate $t$ periods ahead. The time horizon of ex post forecasting ranged from 1 year to 80 years. Table 3 summarizes the important properties of the models fitted for different periods. The fitting of each model included looking at the graph of each series, its autocorrelation and partial autocorrelation functions, identifying an appropriate ARMA-process, estimating its parameters and doing diagnostic checking on the residual autocorrelations. The behaviour of all autocorrelation and partial autocorrelation functions seemed to be consistent with ARIMA (2,2,0)-models.

Table 3: ARIMA(2,2,0)-processes of U.S.population of different fitting periods

| | Parameters | | | | |
|---|---|---|---|---|---|
| fitting period | AR1 | AR2 | $\hat{\sigma}_\epsilon$ | $Q_6$ | $Q_{24}$ |
| 1790-1900 | -0.60 | -0.32 | 0.066 | 7.17 | 32.88 |
| | (-6.5) | (-3.5) | | | |
| 1790-1910 | -0.55 | -0.27 | 0.075 | 3.01 | 22.30 |
| | (-6.1) | (-3.0) | | | |
| 1790-1920 | -0.62 | -0.81 | 0.179 | 4.27 | 13.88 |
| | (-7.0) | (-7.59) | | | |
| 1790-1930 | -0.34 | -0.41 | 0.202 | 10.69 | 23.52 |
| | (-4.3) | (-5.2) | | | |
| 1790-1940 | -0.32 | -0.37 | 0.203 | 5.37 | 24.59 |
| | (-4.11) | (-4.82) | | | |
| 1790-1950 | -0.27 | -0.31 | 0.224 | 3.17 | 17.35 |
| | (-3.6) | (-4.1) | | | |
| 1790-1960 | -0.36 | -0.29 | 0.234 | 2.45 | 8.68 |
| | (-4.8) | (-0.37) | | | |
| 1790-1970 | -0.33 | -0.23 | 0.234 | 3.05 | 13.08 |
| | (-4.5) | (-3.1) | | | |
| 1790-1980 | -0.34 | -0.23 | 0.238 | 2.45 | 17.45 |
| | (-4.7) | (-3.1) | | | |

The numbers in parantheses denote $t$-ratios.

Table 4: Errors $e_t$ of short-term population forecasts with the Box-Jenkins model

| Years ahead | jump-off year 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 |
|---|---|---|---|---|---|---|---|---|
| 1 | -0.26 | 0.32 | -1.39 | 0.56 | -0.15 | -0.06 | 0.00 | 0.00 |
| 2 | -0.32 | 0.42 | -1.01 | 0.52 | -0.26 | -0.06 | 0.11 | 0.10 |
| 3 | -0.21 | 0.27 | -0.66 | 0.50 | -0.34 | -0.04 | 0.14 | 0.16 |
| 4 | -0.24 | 0.20 | -0.78 | 0.51 | -0.36 | -0.08 | 0.16 | 0.19 |
| 5 | -0.24 | 0.26 | -0.79 | 0.48 | -0.35 | -0.10 | 0.19 | 0.20 |

We first look at some short-term forecasts. Table 4 presents the $e_t$ for projections made in jump-off years 1900 through 1970. If the error $e_t$ is negative, the projections have been biased downward, that is, they have been undershooting the mark. A close look reveals a strong pattern in the data. All projections starting with 1900, 1920, 1940, and 1950 were low, and all other projections were high. Similar results were obtained by Stoto (1983) in investigating the accuracy of population projections made by the U.S.Census Bureau. The forecasting using the cohort-component method as well as the forecast based on the Box-Jenkins approach did not anticipate the baby boom in the 1950's nor the baby bust later on.

Stoto (1983) calculated the error $e_t$ of projections made by Pearl and Reed, Dublin, and the Scripps Institute (see Table 5). Their projections for the U.S. in 1970 ranged from 145 to 172 million. The errors $e_t$ ranged between -0.42 and -1.02. Comparable results from ARIMA time series projections for the U.S. in 1970, covering a longer historical span, are presented in Table 6. Those projections ranged from 166.8 to 229.5 million and the errors $e_t$ from -0.69 to 0.35. These results clearly indicate that the Box-Jenkins approach would have been better suited than other methods if it had been applied between 1900 and 1960.

Table 5: Population Projections for U.S.1970
(actual population - 204.9 million)

| Name | Year | Projection (in millions) | Base | $e_t$ |
|---|---|---|---|---|
| Pearl-Reed I | 1910 | 167.9 | 92.4 | -0.33 |
| Pearl-Reed II | 1930 | 160.4 | 123.0 | -0.61 |
| Dublin | 1931 | 151.0 | 124.1 | -0.78 |
| Scripps | 1928 | 171.5 | 120.5 | -0.42 |
| Scripps | 1931 | 144.6 | 124.1 | -0.89 |
| Scripps | 1933 | 146.0 | 125.7 | -0.92 |
| Scripps | 1935 | 155.0 | 127.4 | -0.80 |
| Scripps | 1943 | 160.5 | 136.7 | -0.90 |
| Scripps | 1947 | 162.0 | 144.1 | -1.02 |

Source: Stoto (1983)

Table 6: Population Projections for the U.S.
in 1970 with the Box-Jenkins approach
(actual pop.-204. million)

| Base Year | Projection (in millions) | $e_t$ |
|---|---|---|
| 1900 | 194.0 | -0.08 |
| 1910 | 229.5 | 0.19 |
| 1920 | 172.1 | -0.35 |
| 1930 | 187.0 | -0.23 |
| 1940 | 166.8 | -0.69 |
| 1950 | 205.3 | 0.01 |
| 1960 | 212.3 | 0.35 |

Therefore we can support the findings of Voss et al. (1981), who concluded that the strongest defense of applying Box-Jenkins methods in population prediction lies in the ex post evaluation, which shows that ARIMA-models produce population forecasts which are at least as reliable as more traditional demographic models.

# CONCLUSION

In this study two important results were mentioned. First, the Box-Jenkins approach is equivalent to a simple trend model, when making long-term population forecasts in the United States. Second, the Box-Jenkins approach had not performed worse than more complex demograhic models in projecting future population. Should therefore the Box-Jenkins approach or rather simple trend models be preferred to the component-cohort method?
It is not possible to give an answer to such a question. The application of a forecasting model depends on the specific situation and specific needs. If age-structured population forecasts are required, for example, time series models are obviously not suitable. In addition, forecasters will not reach the same conclusion about a model, so they will use different models for their forecasts. We cannot decide *a priori*, which forecasting model will perform better in future, but the accuracy of the Box-Jenkins approach or of other simple time series methods can be used as a standard for applying more complex demographic or economic forecasting techniques.

As long as they do not beat simple time series models in accuracy, they have to be improved and modified. Comparing the accuracy of time series models and component-cohort models, it is evident that further research concerning the assumptions of the demographic models is inevitable. One possibility is the construction of statistical confidence intervals for population projections. It is assumed that fertility and mortality rates are

random variables. These assumptions imply that the population size is a random variable, too. Its distribution and its resulting confidence intervals can be deduced either by theoretical methods (see Sykes (1969), Cohen (1986)) or by means of Monte Carlo simulation (see Pflaumer (1986,1988)).Generally, these confidence intervals are smaller than the confidence intervals derived by Box-Jenkins models (see e.g. Pflaumer (1988)).

# REFERENCES


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

Cohen, J.E. (1975): *Population Forecasts and Confidence Intervals for Sweden: A comparison of Model-Based and Empirical Approaches. Demography 23: 105-126.*

El-Attar, S. (1988):*Population Forecasting: An Application of Box-Jenkins Technique. American Statistical Association, Proceedings of the Social Statistics Section: 305-310.*

Jöckel, K.-H./Pflaumer, P. (1984):*Calculating the Variance of Errors in Population Forecasting. Statistics and Probability Letters 2: 211-213.*

Land, K.C. (1986):*Methods for National Population Forecasts: A Review. Journal of the American Statistical Association 81: 888-901.*

Land, K.C./Cantor, D.(1983): *ARIMA Models of Seasonal Variation in U.S. Birth and Death Rates. Demography 20: 541-568.*

Lee, R.D. (1974):*Forecasting Births in Post-Transition Populations:Stochastic Renewal with Serially Correlated Fertility. Journal of the American Statistical Association 69: 607-617.*

Pflaumer, P. (1986):*Forecasting the German Population with Monte Carlo Methods. Economics Letters 21: 385-390.*

Pflaumer, P. (1988):*Confidence Intervals for Population Projections Based on Monte Carlo Methods. International Journal of Forecasting 4: 135-142.*

Pflaumer, P. (1988):*Methoden der Bevölkerungsvorausschätzung unter besonderer Berücksichtigung der Unsicherheit*, Berlin.

Pflaumer, P. (1988):*The Accuracy of U.N.Population Projections. American Statistical Association, Proceedings of the Social Statistics Section: 299-304.*

Saboia, J.L.M. (1974):*Modeling and Forecasting Population Time Series. Demography 11: 483-492.*

Sykes, Z.M. (1969):*Some Stochastic Versions of the Matrix Model for Population Dynamics. Journal of the American Statistical Association 44: 111-130.*

Stoto, M. (1983):*The Accuracy of Population Projections. Journal of the American Statistical Association 78: 13-20.*

United Nations (1973):*The Determinants and Consequences of Population Trends. Chapter XV: Demographic Predictions. VolI.*, New York.

U.S. Bureau of the Census (1975):*Historical Statistics of the United States. Colonial Times to 1970. Bicentennial Edition*, Washington, D.C..

U.S. Bureau of the Census (1984):*Current Population Reports. Series P-25. No. 952. Projections of the Population of the United States: 1982 to 2080*, Washington, D.C..

Voss, P.R./Palit, C.D. (1981):*Forecasting State Population using ARIMA Time Series Techniques. Technical Series 70-6. University of Wisconsin*, Madison.