IPRPD
International Journal of Business & Management Studies
ISSN 2694-1430 (Print), 2694-1449 (Online)
Volume 02; Issue no 09: September, 2021

Analyzing U.S. Birth and Population with Structural Time Series Models

Jae J. Lee [1]

[1] *Business Analytics, School of Business, State University of New York, New Paltz, USA*

Abstract

This paper uses the structural time series models to analyze U.S. Birth data (from 1946 to 2017) and U.S. Population data (from 1946 to 2018). Main focus is to study what kind of the stochastic structures that U.S. Birth and U.S. Population can be fitted. This paper uses the local level model, fixed trend model and local linear trend model. With the state space form of these three models, the Kalman filter is used to estimate unknown parameters, to predict one step ahead data and to do filtering data. Based on the AIC and validity check, the best fitted models for U.S. Birth and Population are recommended

Keywords: U.S. Birth; U.S. Population; Structural time series model; State Space Form; Kalman filter

1. Introduction

Univariate or multivariate time series data can be modelled by the structural time series model. The structural time series model formulates a time series data directly in terms of the meaningful components such as trend, seasonal, cycle and irregular. It has an advantage to extract the unobserved meaningful components from observed data. Structural time series model has natural setting for the state space formulation that is required for the Kalman filter.
This paper is to fit three structural time series models to yearly data of U.S. Birth counts (female, male and total counts) from 1946 to 2017 and of U.S. Population counts (female, male and total counts) from 1946 to 2018. Structural time series models that this paper entertained are a random walk model, a fixed trend model and a linear trend model. To estimate parameters of entertained model, to forecast future values and to extract unobserved components from observed data, the Kalman filter (Kalman 1960 & Harvey 1989) is applied with the corresponding state space form (Durbin & Koopman 2012) for each structural model. To compare models for each data, AIC (Akaike Information Criterion) is used.

The plan of this paper is as follows. In section 2, the structural time series model is presented. It also shows how to set up the state space form for each structural model entertained. In Section 3, the Kalman filter is introduced. This section also shows what R package to use for the Kalman filter. In section 4, results of analyzing U.S. Birth data and U.S. Population data by female, male and total are presented. Finally, section 5 concludes the paper.

2. Structural Time Series Model and State Space Form

The structural time series model formulates a time series data directly in terms of meaningful components such as trend, seasonal, cycle and irregular. Since the model formulates data with meaningful components, outcomes of the entertained model are easily interpreted. In the structural time series model, each component has its own disturbance. The characteristics of each disturbance determine the characteristics of time series. Harvey and Todd (1983) compared the structural time series model with Box and Jenkins' ARIMA model. Harvey and Peters (1990) showed the number of methods to compute the maximum likelihood estimator of unknown parameters of the structural time series model.

The first structural time series model entertained in this paper is the local level (LL) model, namely,

$$\begin{cases} \boldsymbol{y_t = \mu_t + \varepsilon_t} \\ \boldsymbol{\mu_t = \mu_{t-1} + \eta_t} \end{cases} \qquad \textbf{(2.1)}$$

where y_t is the observed data, μ_t is the unobserved trend component, η_t is the disturbance (or irregular) component that shows the stochastic behavior of the trend of time series, and ε_t is the disturbance component which shows the stochastic behavior of other than the trend. It is noted that (2.1) is the state space form itself. The state space form (2.1) has 1x1 state vector, μ_t.

Second model entertained is the linear trend (LT) model, namely,

$$\begin{cases} \boldsymbol{y_t = \mu_t + \varepsilon_t} \\ \boldsymbol{\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t} \\ \boldsymbol{\beta_t = \beta_{t-1} + \zeta_t} \end{cases} \qquad \textbf{(2.2)}$$

where y_t is the observed data, μ_t is the unobserved trend component, β_t is the unobserved slope component, ζ_t is the disturbance component that shows the stochastic behavior of the slope of the trend of time series, and η_t and ε_t are defined as (2.1). The state space form of (2.2) is

$$\begin{cases} \boldsymbol{y_t = (1 \quad 0)\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix} + \varepsilon_t} \\ \boldsymbol{\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix} = \begin{pmatrix}1 & 1 \\ 0 & 1\end{pmatrix}\begin{pmatrix}\mu_{t-1} \\ \beta_{t-1}\end{pmatrix} + \begin{pmatrix}\eta_t \\ \zeta_t\end{pmatrix}} \end{cases} \qquad \textbf{(2.3)}$$

The state space form (2.3) has 2x1 state vector, $\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix}$.

A variation of the linear trend model is the fixed trend (FT) model, namely,

$$\begin{cases} \boldsymbol{y_t = \mu_t + \varepsilon_t} \\ \boldsymbol{\mu_t = \mu_{t-1} + \beta + \eta_t} \end{cases} \qquad \textbf{(2.4)}$$

where y_t is the observed data, β is the deterministic slope of the trend, μ_t, η_t, and ε_t are defined as before. The state space form of (2.4) is

$$\begin{cases} \boldsymbol{y_t = (1 \quad 0)\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix} + \varepsilon_t} \\ \boldsymbol{\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix} = \begin{pmatrix}1 & 1 \\ 0 & 1\end{pmatrix}\begin{pmatrix}\mu_{t-1} \\ \beta_{t-1}\end{pmatrix} + \begin{pmatrix}1 \\ 0\end{pmatrix}\eta_t} \end{cases} \qquad \textbf{(2.5)}$$

The state space form (2.5) has 2x1 state vector, $\begin{pmatrix}\mu_t \\ \beta_t\end{pmatrix}$.

3. Kalman Filter

Introduced by Kalman (1960) and Kalman and Bucy (1961), the Kalman filter was employed in many areas by control engineers and physical scientists. Since time series models such as ARIMA models (Box and Jenkins, 1976), structural time series models (Harvey, 1989) and ARMAX models (Hannan and Deistler, 1988) can be built in the specification of the Kalman filter, statisticians in time series analysis widely used the Kalman filter for model specification, parameter estimations, diagnostics, forecasting, filtering and smoothing. The first published paper applied in time series area was Harrision and Stevens (1971) which applied in Bayesian forecasting. Since then, the Kalman filter has been used for analyzing time series data in many areas. For example, disease control (Gove and Houston, 1996), actuary claim reserves forecasting (Chukhrova and Johannssen, 2017), rain fall forecasting (Zulfi et al., 2018), and machine learning (Nobrega and Oliveira, 2019).

The popularity of the Kalman filter is on the flexibility in model specification. The Kalman filter can be employed for both univariate and multivariate time series and for both time variant structure or time invariant structure of time series. Time series data, Y_t, for t = 1, 2, …T denote the observed values of a time series of interest. Y_t could be a univariate or multivariate. Also, let α_t denote the unobserved component vector, called the state vector. For the Kalman filter, it is assumed that the observed data, Y_t and the unobserved component vector, α_t has the linear relationship as

$$\boldsymbol{Y_t = Z_t\alpha_t + \varepsilon_t, \quad t = 1, 2, \ldots, T} \qquad \textbf{(3.1)}$$

in which Y_t is *N* x *1* (N = 1 for the univariate and N >1 for the multivariate), Z_t is *N* x *m* matrix, α_t is *m* x *1* vector, ε_t is *N* x *1* vector of disturbances. In equation (3.1), Z_t is assumed to be a known quantity which shows the relationship between Y_t and α_t, and ε_t is distributed by a multivariate normal with mean of *N* x *1* vector of zero and covariance of N x N matrix, h_t. Please note that the dimension of Y_t and that of α_t are not necessarily same. Equation (3.1) is called the observation (or measurement) equation. For a univariate Y_t, the observation equation can be written as

$$Y_t = z_t'\alpha_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T \qquad (3.2)$$

in which Y_t is a scalar, z_t' is *1* x *m* vector, α_t is *m* x 1 vector, ε_t is a scalar disturbance term whose distribution is a normal with mean of zero and a scalar variance h_t.

The other equation required in Kalman filter is called the system (or transition) equation, which shows how the state variable, α_t is varied over time. The system equation is also linear as

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t, \qquad t = 1, 2, \ldots, T \qquad (3.3)$$

where T_t is *m* x *m* matrix, R_t is *m* x *g* matrix, η_t is *g* x *1* vector of disturbances whose distribution is a multivariate normal with mean of *g* x *1* vector of zero and covariance of g x g matrix, Q_t. It is noted that matrices, Z_t, T_t, and R_t, and covariance matrices, h_t and Q_t may or may not change over time. If all matrices do not change over time, the model is called the time-invariant Kalman filter.

Equation (3.1) and (3.3) together is called the state space form of Kalman filter for a multivariate time series data and (3.2) and (3.3) together is for a univariate time series data. The state space form of Kalman filter has three assumptions:

Assumption 1) the initial state vector, α_0 has a mean of $\mathbf{a}_0$ and Covariance matrix of $\mathbf{P}_0$
Assumption 2) the disturbances ε_t and η_t are independent each other. This assumption could be relaxed.
Assumption 3) the disturbances ε_t and η_t are independent with the initial state, α_0.

With the state space form of (3.1) and (3.3), or (3.2) and (3.3) and the assumptions, the Kalman filter works as a recursive algorithm to provide the estimates of the state variable, α_t using time series data available, $\mathbf{Y}_T = (y_T, \ldots, y_1)$. Given the information of initial state variable, α_0, the Kalman filter starts off the recursive algorithm. Using data at time t-1, $\mathbf{Y}_{t-1} = (y_{t-1}, \ldots, y_1)$, the algorithm predicts the state variable, α_t. Then once data y_t is available at time t, the algorithm updates the state variable, α_t and predicts α_{t+1}. At time t+1, using data y_{t+1}, updates α_{t+1} and predicts α_{t+2}.

The recursive algorithm provides three estimates of the state variable, α_t. First, the filtered estimate is the estimate of α_t given $\mathbf{Y}_t = (y_t, \ldots, y_1)$. That is, the filtered estimate is the estimate of α_t based on the observations available at time t. Second, the forecast estimate is the estimate of α_t for t = T+1, T+2, … given $\mathbf{Y}_T$. That is, the forecast estimates are estimates of α_T, α_{T+1}, ... based on all observations available, $\mathbf{Y}_T$. The smoothed estimate is the estimate of α_t for t = 1, 2, …T given $\mathbf{Y}_T$. That is, the smoothed estimates are estimates of α_1, α_2, …, α_T based on all observations available, $\mathbf{Y}_T$. These all three estimates are minimum mean square estimators (MMSE).

The Kalman filter needs the information of initial state variable, α_0 in order for the recursive algorithm to start off. The information required are mean and covariance matrix of α_0. If the state variable, α_t is stationary, then the mean and covariance matrix of α_0 are given by the mean and covariance matrix of the unconditional distribution of α_t. If the state variable is non-stationary, then the distribution of α_0 should be given as a diffuse prior, that is, putting the covariance matrix of α_0 as kI where k is a very large scalar and I is an identity matrix. If the state variable α_t has both stationary and nonstationary elements, then the covariance matrix of α_0 is given by
$P_0 = \begin{bmatrix} kI & 0 \\ 0 & P \end{bmatrix}$ where I is an identity matrix of d x d with d being the number of nonstationary elements in the state variable α_t and P is the covariance matrix of stationary elements in the state variable α_t (Harvey, 1989).

In the state space form in (3.1) and (3.3) for a multivariate data or in (3.2) and (3.3) for a univariate data, there are some unknown parameters in matrices, Z_t, T_t, h_t and Q_t. Before running the recursive algorithm, these unknown parameters should be estimated. If the disturbances of ε_t and η_t are normally distributed, the likelihood function of the observations could be obtained from the Kalman filter via the prediction error decomposition (Harvey and Peters, 1990). Unknown parameters are estimated by maximizing the likelihood function with respect to the unknown parameters.

Another state space form of (3.1) and (3.3) is possible. The observation equation (3.1) is same but the system equation (3.3) has different form as

$$\alpha_{t+1} = T_t\alpha_t + \eta_t, \qquad t = 1, 2, \ldots, T \qquad (3.4)$$

Outcomes of using the state space form of (3.1) and (3.3), and the form of (3.1) and (3.4) in the Kalman filter are only different when the disturbances ε_t and η_t are correlated at time t. The form of (3.1) and (3.4) specification is useful for setting up the ARMAX models.

For analyzing our data by Kalman filter, we used a R function, *fkf* in a R package, FKF. Given the estimates of unknown parameters in matrices, Z_t, T_t, h_t and Q_t, the function *fkf* provides prediction, filtering and smoothing for univariate and multivariate time series based on the state space form of (3.1) and (3.4). To estimate unknown parameters in matrices, Z_t, T_t, h_t and Q_t, we used a R function, *optim* in a R package, stats. For the method for optimization in *optim*, *L-BFGS-B* is used. *L-BFGS-B* is that of Byrd *et. al.* (1995) which allows box constraints, that is each variable can be given a lower and/or upper bound. The initial value must satisfy the constraints. The method uses a limited-memory modification of the BFGS, which is a quasi-Newton method.

4. Analysis of data

U.S. Birth data: In the United States, State laws require birth certificates to be completed for all births, and Federal law mandates national collection and publication of births and other vital statistics data. The National Vital Statistics (NVS) System is a joint work of National Center for Health Statistics (NCHS) and U.S. States. The NVS system provides statistical information of U.S. birth counts based on birth certificates. Standard form for the collection of the data and model procedures for the uniform registration of the events are developed and recommended for State use through joint activities of the U.S. States and NCHS. NVS report, for example, Martin et. al (2018) presents detailed data on numbers and characteristics of births for year 2017, birth and fertility rates, maternal demographic and health characteristics, medical and health care utilization, source of payment for the delivery, and infant health characteristics.

In this paper, U.S. Birth data from 1946 to 2017 is used to analyze. We separated data by female, male and total. Figure 4-1 shows the time series plot of three data sets of birth from 1946 to 2017.

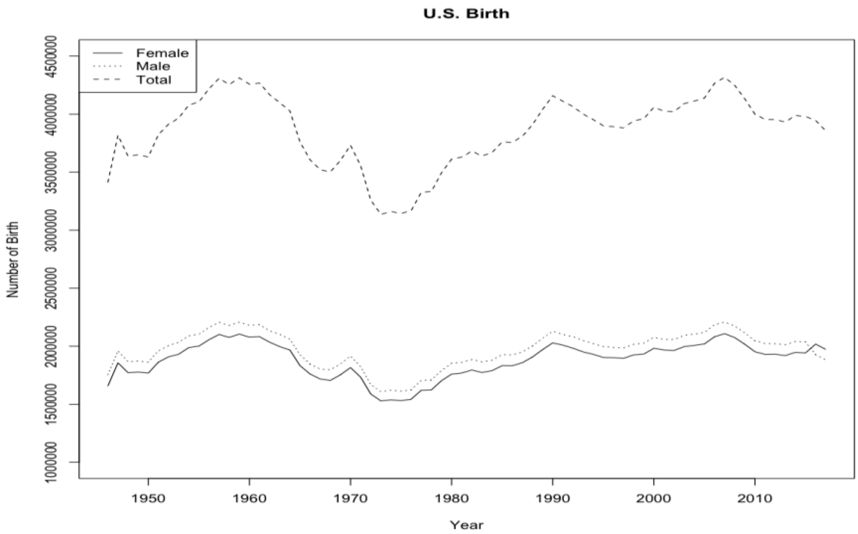

Figure 4.1: U.S. Birth for Female, Male and Total

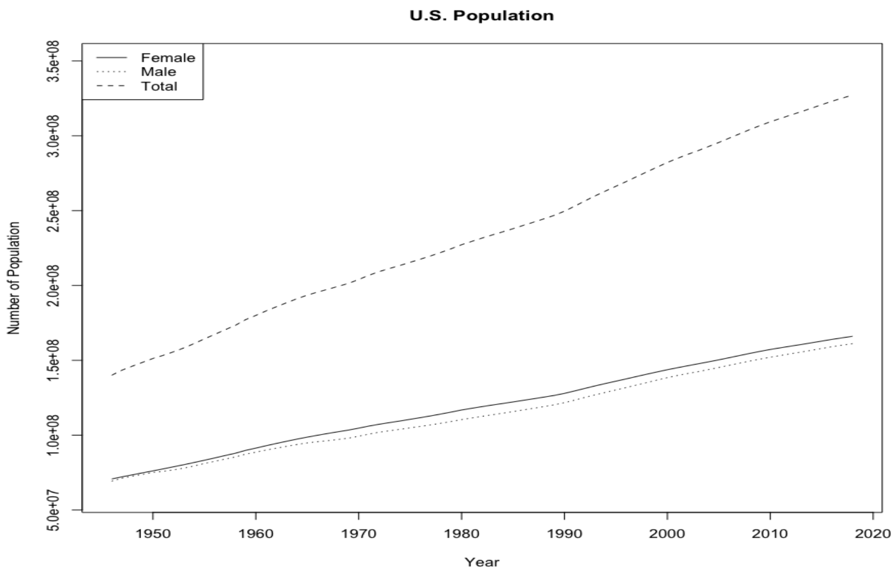

Figure 4.2: U.S. Population for Female, Male and Total

***U.S. Population data*:**

The U.S. Census Bureau provides statistical information of U.S. people counts based on decennial censuses and several annual surveys such as the American Community Survey, the Current Population Survey, and the periodic Survey of Income and Program Participation. In this paper, U.S. Population data from 1946 to 2018 is used to analyze. We separated data by female, male and total as U.S. Birth data. Figure 4-2 shows the time series plot of three data sets of U.S. Population from 1946 to 2018.

***Stochastic models*:**

Three structural time series models are fitted to both U.S. Birth and U.S. Population data: local level (LL) model, linear trend (LT) model and fixed trend (FT) model. State space form for LL model, LT model and FT model is given (2.1), (2.3) and (2.5), respectively. For all three models, it is assumed that errors terms are normally distributed with mean of zero and unknown variances. Thus, unknown parameters of LL model are variances of error terms of η_t and ε_t, namely (varN, varE). For LT model, unknown parameters are variances of error terms of η_t, ζ_t, and ε_t, namely (varN, varK, varE). For FT model, unknown parameters are variances of η_t and ε_t, namely (varN, varE). For FT model, a simple regression of data on time is applied to get the estimate of fixed slope and the estimate is used for starting value for Kalman filter.

For the distributions of components in the initial state, α_0 for Kalman filter, diffuse priors are used since α_t is a nonstationary in all three models. That is, initial values of means for components in α_0 are zero vector and initial value of covariance matrix is kI where k is a large number and I is an identity matrix with a corresponding dimension, namely the dimension of LL model is 1 and the dimension of both LT model and FT model is 2.

***Outputs of U.S. Birth data*:**

Table 4.1 shows the estimates of unknown variances, log likelihood and AIS (Akaike Information Criterion) for female, male and total data using three models.

	VAR N	VAR E	LOGLIKE	AIC
FEMALE	2.77938 e+9	0.341	-892.4731	1788.946
MALE	3.10869 e+9	0.012	-896.4731	1796.946
TOTAL	1.12812 e+10	2.959 e-04	-942.9052	1889.810

Table 4.1(a): LL model for U.S. Birth

	VAR N	VAR E	LOGLIKE	AIC
FEMALE	2.76565 e+9	9.603 e-05	-892.295	1788.590
MALE	3.10548 e+9	0.064	-896.436	1796.873
TOTAL	1.12489 e+10	0.01	-942.804	1889.609

Table 4.1(b): FT model for U.S. Birth

	VAR N	VAR K	VAR E	LOGLIKE	AIC
FEMALE	2.77952 e+9	3.9374 e-4	6.349 e-5	-880.538	1767.077
MALE	3.10869 e+9	2.6535 e-2	1.371 e-4	-884.511	1775.033
TOTAL	1.12812 e+10	1.0015 e-7	2.449 e-8	-930.269	1866.538

Table 4.1(c): LT model for U.S. Birth

Among three models, the LT model has the smallest AIC values for all three types of data, female, male and total. Thus, for U.S. Birth data, the LT model is the best fit among three structural models entertained in terms of AIC.

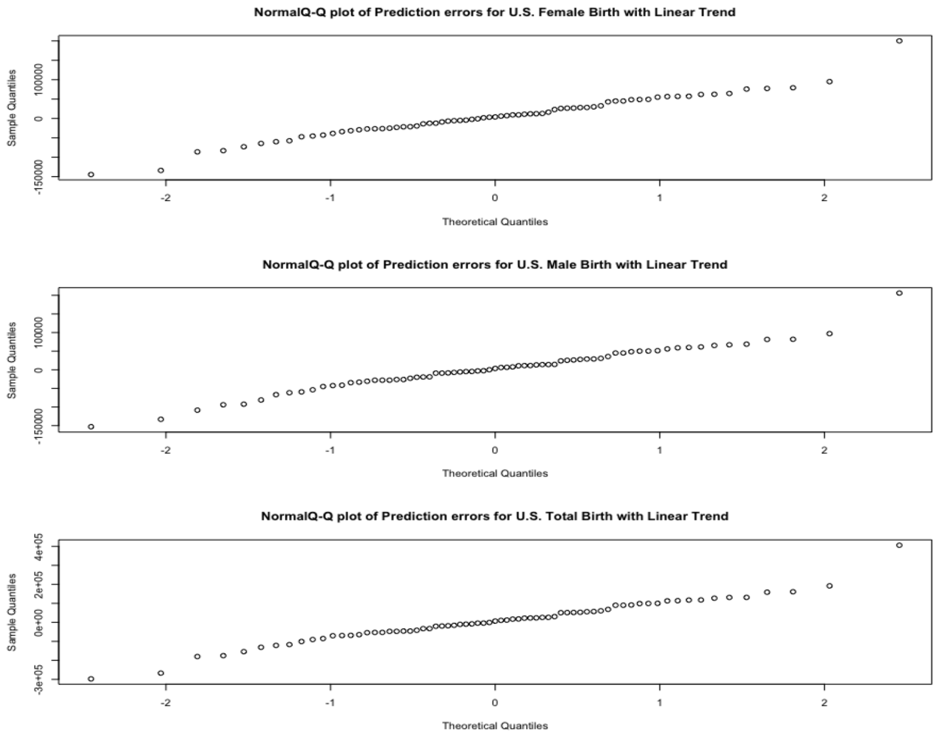

Figure 4.3: Normal plots of one step ahead prediction errors of U.S. Birth with LT model

To check the validity of the LT model, one step ahead prediction errors are used. One step ahead prediction errors in Kalman filter play a role in checking validity of a model as residuals in general statistical models. That is, if a model is valid, one step ahead prediction errors are normally distributed. Figure 4.3 shows the normal plot of one step ahead prediction errors for three types of data with the LT model. The normal plots for all three data show that the one step ahead prediction errors are not away from the normal distribution.

Also, Figure 4.4 shows Kalman filter outputs. It shows observed data, one step ahead prediction data, and filtered data. It is noted that since varE is a lot smaller than varN, the observed data and filtered data are almost overlapped.

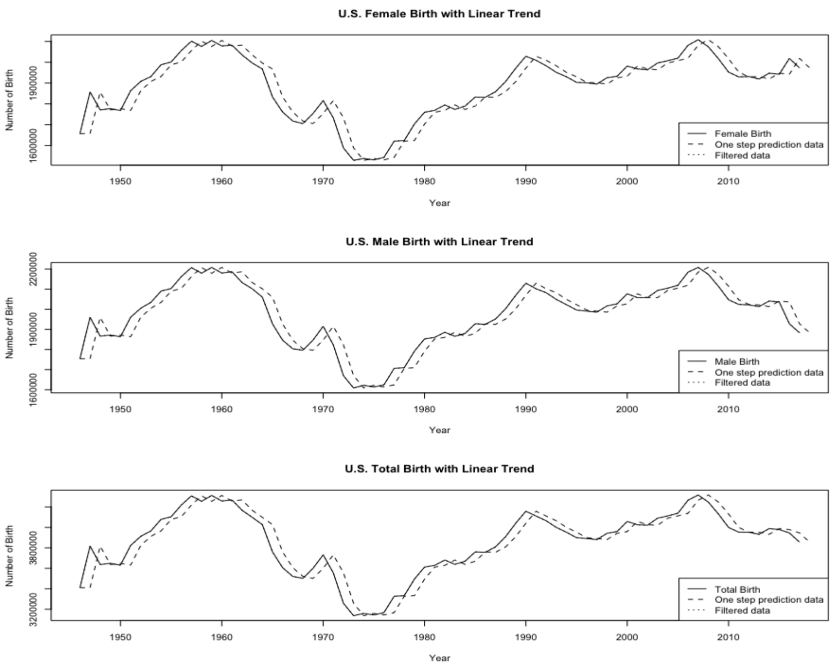

Figure 4.4: Plots of U.S. Birth data, one step ahead prediction data, and filtered data

Outputs of U.S. Population data: Table 4.2 shows the estimates of unknown variances, log likelihood and AIS (Akaike Information Criterion) for female, male and total data using three models.

	VAR N	VAR E	LOGLIKE	AIC
FEMALE	1.79134 e+12	3.504	-1142.843	2289.686
MALE	1.70093 e+12	1.601 e-05	-1140.939	2285.878
TOTAL	6.95357 e+12	5.894	-1192.343	2388.686

Table 4.2(a): LL model for U.S. Population

	VAR N	VAR E	LOGLIKE	AIC
FEMALE	3.98684 e+10	0.307	-1005.858	2015.716
MALE	7.44283 e+10	0.013	-1028.292	2060.584
TOTAL	1.99915 e+11	0.451	-1064.574	2133.148

Table 4.2(b): FT model for U.S. Population

	VAR N	VAR K	VAR E	LOGLIKE	AIC
FEMALE	5.90811 e+10	3.14163 e+10	9.454 e-05	-1028.015	2062.030
MALE	1.05944 e+11	2.94657 e+10	0.012	-1042.171	2090.342
TOTAL	3.18219 e+11	1.20076 e+11	8.769 e-07	-1084.746	2175.492

Table 4.2(c): LT model for U.S. Population

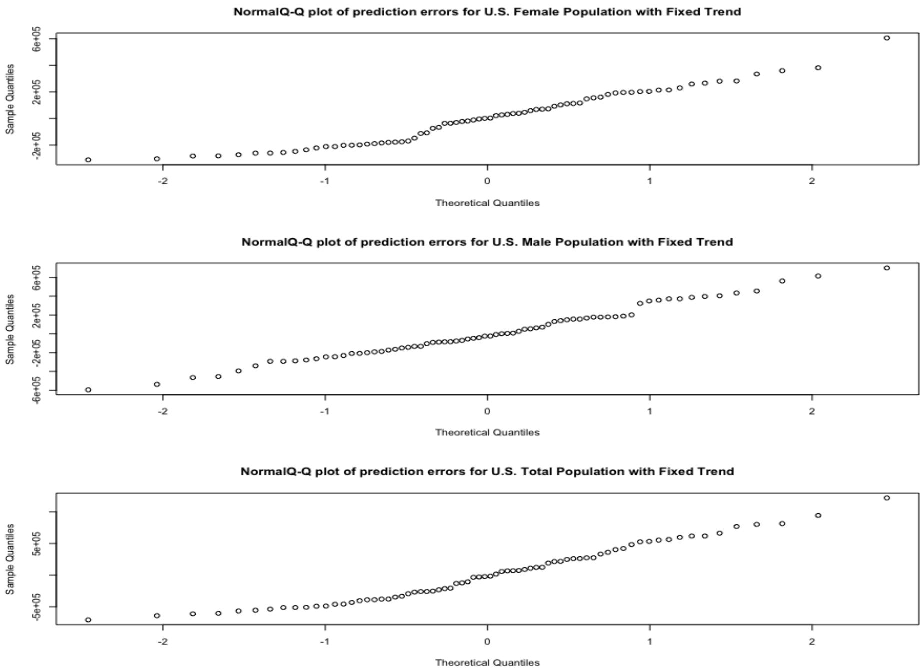

Figure 4.5: Normal plots of one step ahead prediction errors of U.S. Population with FT model

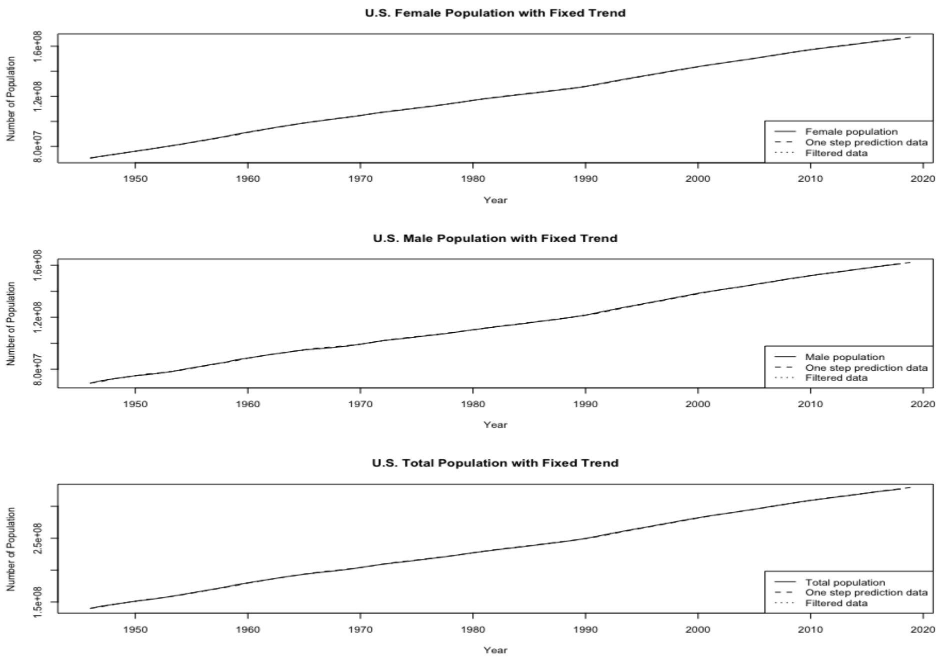

Figure 4.6: Plots of original U.S. Population data, one step ahead prediction data, and filtered data

Table 4.2 shows that among three models, the FT model has the smallest AIC values for all three types of data, female, male and total. Thus, for U.S. Population data which shows a constant slope, the FT model is the best fit among three models entertained in terms of AIC.

To check the model validity of the FT model, one step ahead prediction errors are used as Birth data. Figure 4.5 shows the normal plot of one step ahead prediction errors for three types of data with the FT model. The normal plots for all three data show that the one step ahead prediction errors are not away from the normal distribution.

Also, Figure 4.6 shows observed U.S. Population data, one step ahead prediction data, and filtered data. It is noted that since varE is a lot smaller than varN as birth data, the observed data and filtered data are almost overlapped.

4. Conclusions

This paper analyzes two data, U.S. Birth and U.S. Population data using three structural time series models with Kalman filter. A function, optim in stats package is used to estimate parameters and a function, fkf in FKF package is used to predict one stop ahead and to do filtering data. For U.S. Birth data which does not show any fixed trend pattern, the LT model is the best fit in terms of AIC and for U.S. Population data which shows a fixed trend, the FT model is the best fit in terms of AIC.

Works Cited

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

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995), “A limited memory algorithm for bound constrained optimization”, *SIAM Journal on Scientific Computing*, 16, 1190–1208.

Chukhrova, N. and Johannssen, A. (2017), “State Space Models and the Kalman-Filter in Stochastic Claims Reserving: Forecasting, Filtering and Smoothing”, *Risks* 2017, 5, 30.

Durbin, J. and Koopman, S.J. (2012), *Time Series Analysis by State Space Methods*, 2nd Edition, Oxford: Oxford University Press.

Gove, J.H. and Houston, D.R. (1996), “Monitoring the Growth of American Beech Affected by Beech Bark Disease in Maine Using Kalman Filter”, *Environmental and Ecological Statistics* 3, 167-187.

Harvey, A.C. (1981), *Time Series Models*, Deddington: Oxford.

Harvey, A.C., and Todd, P.H.J. (1983), “Forecasting economic time series with structural and Box-Jenkins models,” (with discussion), Journal of Business and Economic Statistics, 1, 299-315.

Harvey, A.C. (1989), *Forecasting, structural time series models and the Kalman filter*, Cambridge: Cambridge University Press.

Harvey, A.C., and Peters, S. (1990), “Estimation Procedures for Structural Time Series Models”,*Journal of Forecasting*, 9, 89-108.

Hannan, E.J., and Deistler, M. (1988), *The Statistical Theory of Linear System*, New York: Wiley.

Harrison, P.J., and Stevens, C.F. (1971), “A Bayesian Approach to Short-Term Forecasting”,*Operations Research Quarterly*, 22, 341-362.

Kalman, R.E. (1960), “A New Approach to Linear Filtering and Prediction Problems”, *Journal of Basic Engineering*, 82, 34-45.

Kalman, R.E., and Bucy, R.S. (1961), “New Results in Linear Filtering and Prediction Theory”, *Journal of Basic Engineering*, 83, 95-108.

Martin, J.A., Hamilton, B.E., Osterman, M. J. K., Driscoll A.K., and Drake, P. (2018), “Births: Final Data for 2017”, *National Vital Statistics Report,* November; 67 (8): 1-50.

Nobrega, J.P., and Oliveira, A.L.I. (2019), “A Sequential Learning with Kalman Filter and Extreme Learning Machine for Regression and Time Series Forecasting”, *Neurocomputing*, 337, 235-250.

Zulfi, M., Hasan, M., and Purnomo, K.D. (2018), “The development rainfall forecasting using Kalman filter”, *Journal of Physics: Conf. Series* 1008 (2018) 012006.