U.S. Department
Of Transportation
National Highway Traffic Safety Administration

NCSA

DOT HS 809 717

March 2004

Research Note

Time Series Analysis and Forecast of Annual Crash Fatalities

Cejun Liu* and Chou-Lin Chen

Summary

This research note uses two Time Series techniques, Holt-Winters (HW) Algorithm and Autoregressive Moving Average Model (ARMA), to predict annual motor vehicle crash fatalities. Bases on the monthly Fatality Analysis Reporting System (FARS) data from 1975 to 2001, the estimated fatalities are 42,675 and 42,876 respectively in 2002. These estimates are very close to the true counts, as compared to the 2002 fatalities of 42,815. Incorporating the actual 2002 fatalities in the data series, the forecast values in 2003, 41,349 and 41,876, show a decline from the fatalities of 2002.

1. Introduction

Each year, the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA) estimates fatalities in highway vehicle crashes. In this research note, we use an alternative method, time series technique, to estimate the fatalities in 2002 using FARS data from 1975 to 2001 [1, 2]. This forecast value is then compared with the actual observation from FARS 2002. The annual fatalities in 2003 are also forecasted when the actual observation in 2002 is included in the analysis. Two time-series forecasting techniques are used: Holt-Winters algorithm and ARMA (autoregressive moving average) models. The FARS database is a national census of police-reported motor vehicle crashes resulting in fatal injuries, conducted by NCSA.

2. Methodologies

A time series model for the observed data $\{x(t)\}$ is a specification of a sequence of random variables $\{X(t)\}$ of which $\{x(t)\}$ is postulated to be a realization. In this work, the stationary time series model is an appropriate model to be used to perform the analysis and forecast. Definitions and properties of stationary time series models can be found in Appendix 5.1.

2.1. The Holt-Winters (HW) Algorithm

The Holt-Winters algorithm is an effective forecasting technique that has less emphasis on the construction of a model for the time series data. Three smoothing parameters, a, b and $?(\in [0, 1])$ are needed in this process. They can be fixed or be chosen in a way to minimize the sum of squares of the one-step errors. See Appendix 5.2 for description of this technique.

2.2. ARMA Models

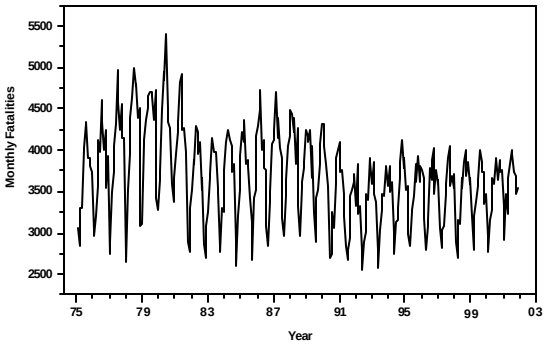
The family of ARMA processes plays a key role in the modeling of time series. In this work, the monthly fatality data from FARS 1975-2002 are used. Since the seasonality exists in the data, we first use lag-d differencing operator to eliminate the seasonal component and then fit an ARMA model. Appendix 5.3 shows us the definition and some properties of ARMA (p, q) process.

3. Results

3.1. Forecasts by HW Algorithm

Figure 1 shows the monthly fatalities over the period of 1975-2002. Figure 1 and sample autocorrelation function (ACF), sample partial auto-correlation function (PACF) in Figure 2 indicate the existence of seasonality. The Holt-Winters algorithm is implemented to predict 2002 fatalities. The forecast value in 2003 is also obtained when actual observations in 2002 are included in the data series. The predicted fatalities are 42,585 and 41,349 for 2002 and 2003 respectively. Three optimized smoothing

Figure 1: Monthly Fatality Series during 1975-2002



Source: NCSA FARS 1975-2002

Table 1: Observed and Forecast Values of Annual Fatalities in 2002 and 2003 by Holt-Winters Algorithm and ARMA Models

Yr.	Actual Fatalities	Forecast				
		Holt-Winters				ARMA
		a	ß	?	Value	Value (95% C.L.)
2002	42815	.25	.00	.32	42585	42675 (38346, 47004)
2003		.25	.00	.33	41349	41876 (37501, 46251)

parameters in the exponential smoothing recursive processes are also shown in Table 1.

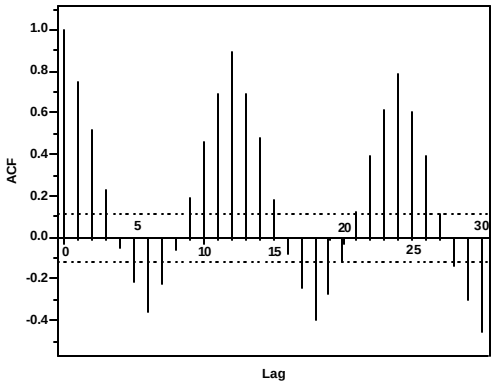
3.2. Forecasts by ARMA Model

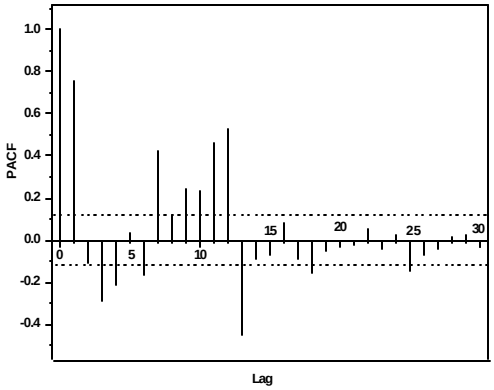
The sample ACF and PACF for the 1975-2001 monthly fatalities are shown in Figure 2, which clearly display a pattern of seasonality with period d=12. After applying the difference operator to $X(t)$ (i.e. $X(t)$-$X(t-d)$ = (1- B^d) $X(t)$ = $Y(t)$, B is backward shift operator), we choose the MA(17) model (E.q. (1)) to this new differenced time series $Y(t)$ (mean-corrected). The residual ACF and PACF and other tests (Ljung and Box test, the McLeod and Li test and the Turning Point test, etc.) show that this model adequately fits the time series $Y(t)$ and the coefficients are significantly different from zero (see Appendix 5.3 for ARMA models),

$$Y(t) = Z(t) + \sum_{i} C_i Z(t-i) \qquad (1)$$

with $Z(t)$ ~ WN(0, 23106) and AICC=4077. The non-zero coefficients are C_1 =.33, C_2 =.27, C_3 =.39, C_4 =.14, C_5 =.25, C_6 =.32, C_7 =.28, C_8 =.26, C_9 =.30, C_{10} =.20, C_{11} =.27, C_{12} =-.58, C_{14} =.17, C_{16} =.29 and C_{17} =.21.

Figure 2: Sample ACF and PACF for Monthly Fatality Series over 1975-2001





The forecast fatalities in 2002 can then be obtained in terms of this fitted model as

$$42675,\ 95\%\ \text{C.L.} = (38346, 47004). \qquad (2)$$

where C.L. = Confidence Limit. We also get an adequately fitted MA (17) model (mean-corrected) to the monthly fatalities when the actual observation in 2002 is included,

$$Y(t) = Z(t) + \sum_{i} C_i Z(t-i) \qquad (3)$$

with $Z(t)$ ~ WN(0, 23120) and AICC=4233. The non-zero coefficients are C_1 =.38, C_2 =.32, C_3 =.38, C_4 =.17, C_5 =.25, C_6 =.29, C_7 =.22, C_8 =.23, C_9 =.27, C_{10} =.19, C_{11} =.24, C_{12} =-.59, C_{13} =-.095, C_{14} =.11, C_{16} =.23 and C_{17} =.16.

Then the forecast fatalities in 2003 is

$$41876,\ 95\%\ \text{C.L.} = (37501, 46251). \qquad (4)$$

4. Conclusions

In this work, time series techniques are used to analyze the annual crash fatalities. The fatalities in 2002 are predicted and then compared with the actual observation. The forecast in 2003 is also implemented when the observations in 2002 is included. The values predicted by ARMA models are a little bit higher than the ones obtained by Holt-Winters algorithm. In 2002, both forecast values are pretty good when the relative forecast errors are examined. Based on these two forecasts, the annual fatalities in 2003 will decrease as compared to 2002.

5. Appendix

5.1. Stationary Time Series

Loosely speaking, a time series model for the observed data $\{x(t)\}$ is a specification of a sequence of random variables $\{X(t)\}$ of which $\{x(t)\}$ is postulated to be a realization. A time series $\{X(t), t = 0, \pm 1, \ldots\}$ is said to be stationary if it has statistical properties similar to those of the "time-shifted" series $\{X(t+h), t=0, \pm 1, \ldots\}$ for each integer h.

Two simple but very useful stationary models are IID (independently and identically distributed) noise and White noise. For IID noise, random variables $X(t)$ are mean 0 and variance s^2 (= $E[X(t)^2]$), specified as $\{X(t)\}$~IID (0, s^2). If $\{X(t)\}$ is a sequence of uncorrelated random variables, each with mean 0 and variance s^2, then it is referred to as white noise, specified as $\{X(t)\}$~WN (0, s^2). Every IID (0, s^2) sequence is WN (0, s^2) but not conversely.

For a stationary time series $\{X(t)\}$, sample autocorrelation function (ACF) and sample partial auto-correlation function (PACF) are used in choosing an appropriate model to the observed time series.

5.2. Holt-Winters (HW) Algorithm

The Holt-Winters algorithm is an effective forecasting technique that has less emphasis on the construction of a model for the time series. Giving time series $\{X(t)\}$, $t=1,\ldots,n$ from the following classical decomposition model

$$X(t) = m(t) + s(t) + Y(t), \; t=1,\ldots n, \qquad (5)$$

where $m(t)$ is a trend component, $s(t)$ is a seasonal component with known period d (i.e. $s(t+d)=s(t)$ and $\sum_{j=1}^{d} s(j) = 0$) and $Y(t)$ is a random noise component which is stationary with $E(Y(t))=0$. The estimated component $m(t)$ and $s(t)$ at times $t=1, 2, \cdots, n$ can be computed in terms of exponential smoothing recursions schemes. In the current study, three smoothing parameters, a, b (for trend component) and $?$ (for seasonal component) with $a, b, ? \in [0, 1]$ are needed. Here, they are chosen in a way to minimize the sum of squares of the one-step errors $\sum_{j=d+2}^{n} (X(j)-P_{j-1} X(j))^2$, P_j is predictor operator. Details of the HW algorithm can be obtained in references [3-6]. This method allows the seasonal pattern to adapt over time. It is one of the best-known forecasting techniques in time series theories [7].

5.3. ARMA Models

A stationary time series $\{X(t)\}$ is called an ARMA(p, q) process if for every t

$$X(t) = f_1 X(t-1) + \ldots + f_p X(t-p) + Z(t) + q_1 Z(t-1) + \ldots + q_q Z(t-q), \qquad (6)$$

where $\{Z(t)\} \sim WN(0, s^2)$. $\{X(t)\}$ is said to be an ARMA (p, q) model with mean μ if $\{X(t)-\mu\}$ is an ARMA (p, q) process defined by Eq.(6). A stationary solution $\{X(t)\}$ of the Eq.(6) exists if and only if $f(z)=1- f_1 z - \ldots - f_p z^p \neq 0$, $\forall \; |z| =1$.

For pure autoregressive (AR) models, the Yule-Walker algorithm is used to implement the preliminary estimation of the models. For pure moving average (MA) models or mixed ARMA models, the Innovations algorithm is used to implement the preliminary estimation of the models. Final decisions with respect to order selection of the models are made on the basis of the Maximum Likelihood Estimator and the minimum AICC (bias-Corrected Information Criterion of Akaike) criterion.

Once a well-fitted model for a time series is

obtained, it can then be employed to predict $X(n+h)$ ($h>0$) with known mean and auto-covariance function in terms of the values $\{X(t)\}$, $t=1,\ldots,n$. Refer to [3-6] for details. ITSM and SAS are used in the calculations.

For non-stationary time series (e.g. trend or seasonality), ARIMA (auto-regressive integrated moving average) models are used. For ARIMA process, the classical decomposition model (i.e. Eq.(5)) or lag-d differencing techniques are employed to eliminate the trend or seasonality component and then a stationary time series is generated. In this work, we use a differencing scheme to eliminate the seasonal component of the time series and then adopt an approach of fitting a subset ARMA model to this differenced series as suggested in ACF. We did not use the structure $(p,d,q)\times(P,D,Q)_s$ (here d is the order of difference to the time series and s the period of seasonality) and corresponding identification procedure for the seasonal ARIMA model (SARIMA) in this study [3-6]. In addition, there is no evidence against the stationarity of the time series in variance and hence no transformation was needed.

6. References

[1] National Highway Traffic Safety Administration. Traffic Safety Facts 2001. Washington, DC: US Department of Transportation.

[2] 2002 Annual Assessment. National Center for Statistics and Analysis, NHTSA, http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/Rpts/2003/Assess02.pdf

[3] D.C. Montgomery, L.A. Johnson and J.S. Gardiner, *Forecasting and Time Series Analysis*, 2nd, McGraw-Hill, Inc., New York (1990).

[4] W.S. Wei, *Time Series Analysis*, Addison-Wesley Publishing Company, Inc. New York (1990).

[5] P.J. Brockwell and R.A. Davis, *Introduction to Time Series and Forecasting*, New York: Springer (1996).

[6] H.Arsham, *Time Series Analysis and Forecasting Techniques*. http://ubmail.ubalt.edu/~harsham.

[7] L. Richard, *How Should Additive Holt-Winters Estimates be Corrected?* International Journal of Forecasting, 14: 393 (1998).

---

*__Cejun Liu__ is a Program Analyst employed by Rainbow Technology Inc., a contractor working for the Mathematical Analysis Division, National Center for Statistics and Analysis, NHTSA.

__Chou-Lin Chen__ is a Mathematical Statistician and team leader in the Mathematical Analysis Division, National Center for Statistics and Analysis, NHTSA.

Very useful suggestions and comments from Santokh Singh and Dennis Utter at NCSA, other reviewers at NHTSA and helpful communication with Professor, Dr. Peter J. Brockwell are greatly appreciated.

For additional copies of this research note, please call 1-800-934-8517 or fax your request to (202) 366-3189. For questions regarding the data reported in this research, contact Cejun Liu [202-366-5354] or Chou-Lin Chen [202-366-1048]. Internet users may access this research note and other general information on highway traffic safety at: http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/AvailInf.html

U.S. Department
of Transportation
National Highway
Traffic Safety
Administration
400 Seventh Street, S.W., NPO-100
Washington, D.C. 20590