Traffic Safety Facts
Research Note

NCSA National Center for Statistics & Analysis

March 2004 DOT HS 809 718

Time Series Analysis and Forecast of Crash Fatalities during Six Holiday Periods

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 motor vehicle crash fatalities during six holiday periods. Based on the data from 1983 to 2001, the estimated fatalities in 2002 are: 564 and 564 (New Year), 501 and 514 (Memorial), 748 and 755 (4th of July), 486 and 498 (Labor), 572 and 565 (Thanksgiving), and 156 and 145 (Christmas). Incorporating the actual 2002 fatality counts in the data series, the forecasts for the 2003 holiday periods are: 141 and 141 (New Year), 508 and 504 (Memorial), 537 and 564 (4th of July), 526 and 545 (Labor), 547 and 564 (Thanksgiving), and 548 and 576 (Christmas).

1. Introduction

Generally, there are higher fatality rates during holiday periods than during non-holiday periods. In 2001 and 2002, the overall average fatalities were 116 and 117 per day respectively. In comparison, the average fatalities during six major holiday periods were 153 and 156 per day, respectively [1, 2]). Analysis and forecasting of the fatality rates during holiday periods are useful for providing warnings that may reduce fatalities. In this research note, time series techniques are employed to analyze the fatality data during six holiday periods. The fatalities in 2002 are forecasted using the data from 1982 to 2001 and then compared with the actual observations in 2002. Fatalities in 2003 during six holiday periods are also forecasted when the observations in 2002 are included in the analysis. Two forecasting techniques are used: Holt-Winters algorithm and ARMA (autoregressive moving average) models. Data from the Fatality Analysis Reporting System (FARS) were used. The FARS database is a national census of police-reported motor vehicle crashes resulting in fatal injuries. It is conducted by the National Center for Statistics and Analysis (NCSA) in the National Highway Traffic Safety Administration (NHTSA).

2. Methodologies

In the current study, the extrapolation method will be used. That is, the forecast is based on an inferred study of past general data behavior over time (time series).

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 4.1.

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. In this process, two smoothing parameters, α and β with $\alpha, \beta \in [0, 1]$ are needed. They can be fixed or be chosen in a way to minimize the sum of squares of the one-step errors. See Appendix 4.2 for description of this technique.

2.2. ARMA Models

The family of ARMA processes plays a key role in the modeling of time series. Appendix 4.3 shows us the definition and some properties of ARMA (p, q) process. In this study,

visualizations of six time series are shown to be stationary and confirmed by statistical criteria. The ARMA models are employed for the analysis.

3. Results

3.1. Data Manipulation

Fatalities during six holiday periods over 1982-2002 are listed in Tables 1 and 2. For the fatalities, i.e. New Year, Fourth of July and Christmas, which have different numbers of whole days during the holiday period, we first obtain the average number of fatalities for each day of the period, and used these numbers to perform a time series analysis and forecast, in the end, multiplied by whole days of the holiday period to get the total fatalities for each holiday period. Figure 1 shows the time series of the killed persons during six holiday periods over 1982-2001. From Figure 1, we can see there is a weak downtrend in Memorial Day and Labor Day series. Indeed, different ARMA models for those two data series from four other holidays' series can be found.

Table 1

Number of Killed Persons during New Year, Memorial, and Fourth of July Holiday Periods and Forecast Values in 2002 and 2003

Year		New Year		Memorial	Fourth of July	
		Number (Days)	One Day	Number (Days)	Number (Days)	One Day
1982		NA	NA	498(3)	600(3)	200
1983		375(3)	125	539(3)	620(3)	207
1984		346(3)	116	527(3)	223(1)	223
1985		496(4)	124	557(3)	689(4)	173
1986		223(1)	223	616(3)	611(3)	204
1987		535(4)	134	519(3)	556(3)	186
1988		407(3)	136	529(3)	631(3)	211
1989		443(3)	148	594(3)	748(4)	187
1990		421(3)	141	589(3)	268(1)	268
1991		441(4)	111	533(3)	718(4)	180
1992		164(1)	164	438(3)	535(3)	179
1993		370(3)	124	454(3)	525(3)	175
1994		372(3)	124	482(3)	519(3)	173
1995		392(3)	131	483(3)	661(4)	166
1996		420(3)	140	514(3)	627(4)	157
1997		190(1)	190	511(3)	508(3)	170
1998		545(4)	137	393(3)	479(3)	160
1999		354(3)	118	500(3)	509(3)	170
2000		469(3)	157	466(3)	717(4)	180
2001		357(3)	119	515(3)	206(1)	206
2002		570(4)	143	491(3)	683(4)	171
2002 Forecast	**HW**	564(4)	141	501(3)	748(4)	188
	ARMA	564(4)	141	514(3)	755(4)	189
		95% C.L.	95%C.L.	95% C.L.	95% C.L.	95%C.L.
		(352,772)	(88,193)	(421,607)	(556, 956)	(139, 239)
2003 Forecast	**HW**	141(1)	141	508(3)	537(3)	179
	ARMA	141(1)	141	504(3)	564(3)	188
		95%C.L.	95%C.L.	95% C.L.	95% C.L.	95%C.L.
		(89, 192)	(89, 192)	(413, 595)	(417, 711)	(139, 237)

Source: NCSA FARS 1982-2002

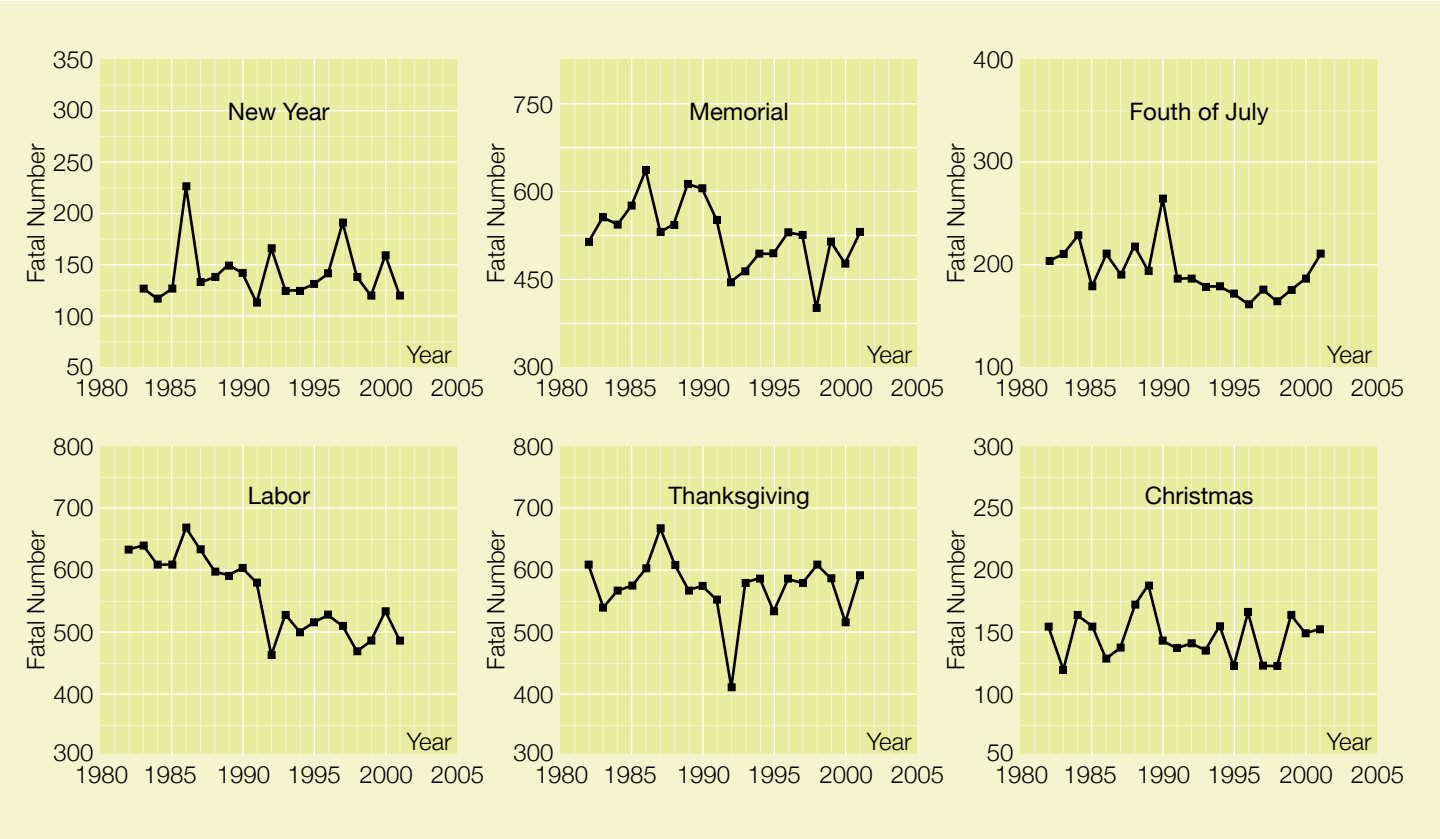
Table 2

Number of Killed Persons during Labor, Thanksgiving, and Christmas Holiday Periods and Forecast Values in 2002 and 2003

Year		Labor	Thanksgiving	Christmas	
		Number (Days)	Number (Days)	Number (Days)	One Day
1982		628(3)	601(4)	458(3)	153
1983		636(3)	533(4)	352(3)	118
1984		609(3)	558(4)	643(4)	161
1985		605(3)	566(4)	152(1)	152
1986		663(3)	598(4)	508(4)	127
1987		630(3)	659(4)	409(3)	137
1988		592(3)	601(4)	511(3)	171
1989		588(3)	561(4)	553(3)	185
1990		599(3)	563(4)	567(4)	142
1991		577(3)	546(4)	135(1)	135
1992		460(3)	403(4)	410(3)	137
1993		522(3)	569(4)	402(3)	134
1994		494(3)	575(4)	455(3)	152
1995		511(3)	527(4)	358(3)	120
1996		525(3)	579(4)	166(1)	166
1997		507(3)	571(4)	480(4)	120
1998		464(3)	602(4)	364(3)	122
1999		485(3)	581(4)	485(3)	162
2000		529(3)	509(4)	442(3)	148
2001		482(3)	585(4)	601(4)	151
2002		541(3)	543(4)	130(1)	130
2002 Forecast	**HW**	486(3)	572(4)	156(1)	156
	ARMA	498(3)	565(4)	145(1)	145
		95% C.L.	95% C.L.	95% C.L.	95% C.L.
		(421, 575)	(470, 660)	(109, 181)	(109, 181)
2003 Forecast	**HW**	526(3)	547(4)	548(4)	137
	ARMA	545(3)	564(4)	576(4)	144
		95% C.L.	95% C.L.	95% C.L.	95% C.L.
		(467, 623)	(470, 657)	(436, 720)	(109, 180)

Source: NCSA FARS 1982-2002

Figure 1:
Fatalities during Six Holiday Periods over 1982-2001



Source: NCSA FARS 1982-2001

3.2. Forecasts by HW Algorithm

Using 1982-2001 data, the Holt-Winters algorithm is implemented to predict 2002 fatalities. Forecast values in 2003 are also obtained when actual observations in 2002 are included in the time series. Table 3 shows detailed information for the two optimized smoothing parameters α and β in the exponential smoothing recursive process and the forecast values by the Holt-Winters algorithm. For 2002 fatalities, the Holt-Winters algorithm performs forecasts very well from the comparisons of the actual observations with the forecast values listed in Tables 1 and 2.

3.3. Forecasts by ARMA Model

Sample autocorrelation function (ACF) and sample partial auto-correlation function (PACF), preliminary and maximum likelihood estimation procedures with minimum AICC criteria are used to find the best model for each data series. The ARMA models and the forecast values in 2002 and 2003 are shown as follows (mean-corrected). For the forecast value, the 95% confidence limit (C.L.) can be constructed assuming approximate normality in terms of mean-squared error (MSE) (i.e. C.L = measure ± 1.96 $\sqrt{MSE}$). For most cases, the value forecasted by ARMA model is a little bit higher than the one obtained by Holt-Winters algorithm.

Table 3: **Forecasted Fatalities of Six Holidays in 2002 and 2003 by Holt-Winters Algorithm.**

Holiday		α	β	Forecast
New Year	2002	0.24	0.27	140.15
	2003	0.24	0.27	140.50
Memorial	2002	0.30	0.78	500.30
	2003	0.30	0.78	507.08
Fourth of July	2002	0.46	0.14	187.08
	2003	0.42	0.16	178.97
Labor	2002	0.53	0.20	485.85
	2003	0.39	0.50	525.34
Thanks-Giving	2002	0.76	0.24	571.08
	2003	0.75	0.25	546.66
Christmas	2002	0.64	0.40	155.70
	2003	0.65	0.38	136.14

New Year

2002:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 726.094)$.
Forecast: 140.106, $\sqrt{MSE} = 26.946$.

2003:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 690.188)$.
Forecast: 140.25, $\sqrt{MSE} = 26.271$.

Memorial

2002:
AR (1): $X(t) = 0.4005\ X(t\text{-}1) + Z(t)$,
$Z(t) \sim WN(0, 2243.63)$.
AICC = 215.955.
Forecast: 513.711, $\sqrt{MSE} = 47.367$.

2003:
AR (1): $X(t) = 0.4003\ X(t\text{-}1) + Z(t)$,
$Z(t) \sim WN(0, 2160.137)$.
AICC = 225.673.
Forecast: 503.48, $\sqrt{MSE} = 46.677$.

Fourth of July

2002:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 646.887)$.
Forecast: 188.75, $\sqrt{MSE} = 25.434$.

2003:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 630.37)$.
Forecast: 187.905, $\sqrt{MSE} = 25.11$.

Labor

2002:
AR (1): $X(t) = 0.7861\ X(t\text{-}1) + Z(t)$,
$Z(t) \sim WN(0, 1559.86)$.
AICC = 209.473.
Forecast: 497.6784, $\sqrt{MSE} = 39.495$.

2003:
AR (1): $X(t) = 0.7493\ X(t\text{-}1) + Z(t)$,
$Z(t) \sim WN(0, 1580.429)$.
AICC = 219.761.
Forecast: 544.414, $\sqrt{MSE} = 39.755$.

Thanksgiving

2002:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 2358.03)$.
Forecast: 564.35, $\sqrt{MSE} = 48.56$.

2003:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 2266.41)$.
Forecast: 563.33, $\sqrt{MSE} = 47.61$.

Christmas

2002:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 326.64)$.
Forecast: 144.65, $\sqrt{MSE} = 18.24$.

2003:
MA (0): $X(t) = Z(t), Z(t) \sim WN(0, 326.71)$.
Forecast: 143.95, $\sqrt{MSE} = 18.08$.

We can also see that those models perform forecasts pretty well in 2002 as compared to the actual observations.

4. Appendix

4.1. 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, ±1, …} is said to be stationary if it has statistical properties similar to those of the "time-shifted" series {X(t+h), t=0, ±1, …} 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 σ^2 (= $E[X(t)^2]$), specified as $\{X(t)\} \sim IID\ (0, \sigma^2)$. If {X(t)} is a sequence of uncorrelated random variables, each with mean 0 and variance σ^2, then it is referred to as white noise, specified as $\{X(t)\} \sim WN\ (0, \sigma^2)$. Every IID $(0, \sigma^2)$ sequence is WN $(0, \sigma^2)$ but not conversely.

For a 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.

4.2. Holt-Winters (HW) Algorithm

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

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

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,⋯⋯, n can be computed in terms of exponential smoothing recursions schemes. In this process, two smoothing parameters, α and β with α, β ∈ [0, 1] are needed (no seasonal component). Here, they are chosen in a way to minimize the sum of squares of the one-step errors $\sum_{j=3}^{n}$

$(X(j)-P_{j-1} X(j))^2$, Pj is predictor operator. Details of the HW algorithm can be found in references [3-6]. The HW algorithm is one of the best-known forecasting techniques in time series theories [7].

4.3. ARMA Models

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

$$X(t) = \phi_1 X(t-1) + \ldots + \phi_p X(t-p) + Z(t) + \theta_1 Z(t-1) + \ldots + \theta_q Z(t-q), \quad (2)$$

where $\{Z(t)\} \sim WN(0, \sigma^2)$. {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.(2). A stationary solution {X(t)} of the Eq.(2) exists if and only if $\phi(z)=1- \phi_1 z - \ldots - \phi_p z^p \neq 0 , \forall |z| =1$.

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,…,n. Refer to [3-6] for details. ITSM and SAS are used in the calculations.

In this work, the stationarity issue of the data series can be confirmed by sample ACF and sample PACF of each data series and hence the ARMA models are used for the analysis. For non-stationary process, ARIMA (auto-regressive integrat ed moving average) models should be used. For the data series in this study, there is no evidence to use ARIMA process.

5. 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.

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

NHTSA People Saving People www.nhtsa.dot.gov

MATHEMATICAL ANALYSIS DIVISION

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/AvalInf.html