Daily Collision Prediction with SARIMAX and Generalized Linear Models on the Basis of Temporal and Weather Variables

Yongsheng Chen and Stevanus Tjandra

Short-term collision prediction is a relatively new area of research in the field of traffic safety because of the high randomness of data and the methodological complexity. Motivated by requirements from frontline traffic operations and enforcement services, the authors conducted this study to develop models that predicted daily total collisions. The study started with decomposition analysis of time series data to determine trends, seasonality, and randomness of daily collisions before it proceeded with an investigation of potential collision contributors. Temporal factors (i.e., months, weekdays, and holidays) and weather forecasts (i.e., daily mean temperature, amount of rainfall, and amount of snowfall) were selected as predictive factors. Accordingly, the seasonal autoregressive integrated moving average model with external regressors (SARIMAX) was identified, and a series of SARIMAX models of different orders was estimated and diagnosed. A generalized linear model (GLM) was also developed and compared with the SARIMAX models by validation measures. Finally, a calibration mechanism was recommended to optimize predictions. Model validations provide evidence that both SARIMAX and GLM are adaptable; however, the SARIMAX models are a viable and preferable option because they can provide greater accuracy than GLM in the short-term prediction of collisions. The models developed in this paper are now being applied (*a*) to support scheduling of traffic operations, maintenance and enforcement, and dispatch of material and personnel resources and (*b*) to provide situation awareness for all road users and stakeholders.

Adverse weather conditions (i.e., snow, rain, fog, sleet, or wind) are known to play a significant role in reducing pavement friction, lowering vehicle performance, impeding driver visibility, severely affecting traffic control device functions, and changing driver behavior. On average, 7,130 people are killed and more than 629,000 people injured each year in weather-related collisions in the United States (*1*). The economic toll of these deaths and injuries is estimated at $42 billion/year (*2*). In Canada, weather-related collisions have been estimated to cost Canadians an average of approximately Can$1 billion/year (Can$1.00 = $0.911499 in 2014) (*3*). Significant increases in collision rates during snowy months have been reported in many places around the globe, for example, in Canada (*4*) and the United Kingdom (*5*).

Many of these consequences may be avoided if weather-related collisions are predicted in a timely, accurate manner. This could be done by various means, including operation measures (e.g., dynamic messaging signs), roadway maintenance measures (e.g., snow removal), media (television, radio, and newspaper), web-based messages, and enforcement.

Weather-related prediction of collisions is, by nature, a short-term prediction. Unlike conventional predictions of traffic safety (*6–8*) that focused predominantly on long-term (annual) collisions, short-term (daily or even hourly) prediction of traffic collisions is still a new area. Short-term safety data are usually unavailable, incomplete, or insufficient for modeling. In addition, short-term prediction might be hindered by methodological difficulties caused by high randomness of the data.

All these challenges overshadow the importance of short-term prediction of collisions. In fact, all practices at the tactical level—including traffic enforcement, operation, control, and maintenance—rely heavily on either daily or hourly preset schedules. The success of these schedules stems from accurate future forecasts in the form of short-term predictions.

To cater to the front-line requests of traffic operations and enforcement services, the authors conducted a study of daily collisions to predict short-term citywide collisions to (*a*) scheduling traffic enforcement, operation, control, and maintenance groups; (*b*) deploy resources and personnel; and (*c*) create situational awareness for all road users and stakeholders.

LITERATURE REVIEW

Liu and Chen used the Holt–Winters algorithm and an autoregressive moving average model to predict 2002 and 2003 annual fatalities in motor vehicle collisions by using monthly data from 1975 to 2001 (*9*). Quddus used several integer-valued autoregressive (INAR) models to fit annual collisions from 1950 to 2005 and monthly collision casualties from January 1991 to October 2005 in the United Kingdom (*10*). Brijs et al. also applied INAR on the daily car collision data attributed to weather and traffic exposure (*11*). Vanlaar et al. applied an autoregressive integrated moving average [ARIMA (*p, d, q*), where *p* is the number of autoregressive terms, *d* is the number of nonseasonal differences, and *q* is the number of lagged forecast errors in the prediction equation] model to estimate the logarithm of monthly collisions from 1994 to 2008 at 48 intersections to investigate trends regarding collisions, injuries, and collision severity before and after the implementation of photo enforcement in the City of Winnipeg, Manitoba, Canada (*12*). El-Basyouny and Kwon attributed daily collision frequency to calendar features (weekend, weekday, holiday)

Office of Traffic Safety, City of Edmonton, Suite 200, 9304–41 Avenue NW, Edmonton, Alberta T6E 6G8, Canada. Corresponding author: Y. Chen, yongsheng.chen@edmonton.ca.

Transportation Research Record: Journal of the Transportation Research Board, No. 2432, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 26–36.
DOI: 10.3141/2432-04

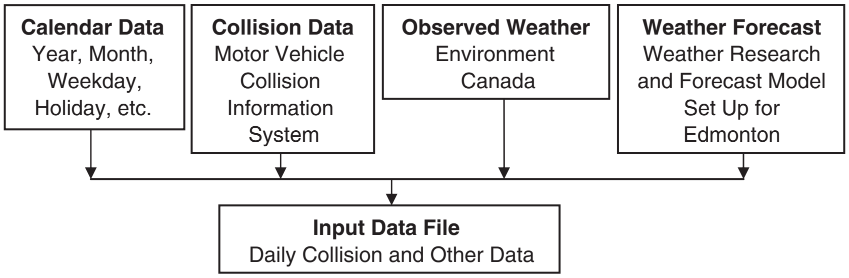

Calendar Data
Year, Month, Weekday, Holiday, etc.
Collision Data
Motor Vehicle Collision Information System
Observed Weather
Environment Canada
Weather Forecast
Weather Research and Forecast Model Set Up for Edmonton
Input Data File
Daily Collision and Other Data

FIGURE 1 Input data sources for prediction of daily collisions.

and weather effects on the basis of data from the City of Edmonton, Alberta, Canada, from 2000 to 2010 with multivariate Poisson lognormal models [the extensive form of a generalized linear model (GLM)] and estimated the parameters by the Markov chain Monte Carlo method (*13*). Halim and Jiang employed a negative binomial (NB) time series model to account for seasonality and weather factors on the basis of City of Edmonton daily collision data from September 2008 to December 2011 (*14*). They also used the Gaussian time series model to account for seasonality and weather factors to estimate collision severity.

This past research established a fundamental framework for short-term prediction of collisions and addressed many methodological issues from different perspectives. For example, the INAR model developed by Brijs et al. conceptually accommodated the discrete counts of daily traffic collisions (*11*). This type of model, however, was not seasonal and was unable to address daily collision seasonality. In addition, the applied traffic exposure variable defined by the "total amount of vehicle kilometers driven on the major road network of each city region" was not commonly measurable for larger areas such as the city of Edmonton (*11*).

After the synthesis of research on this topic, some methodological vacancies still existed and needed to be addressed through the study in this paper. Those include the following:

- Models to account for profound seasonality of daily collisions beyond its general trend,
- Temporal correlation in collisions and other external traffic safety factors like weather,
- Models to compensate for the lack of traffic exposure data for a large citywide area, and
- Models to fit daily-level collision data with high randomness.

Accordingly, this study aims to develop time series models with particular functionality such as embedding both temporal and external features, employing alternative measures to explain traffic exposure, and in particular adapting temporally microscopic (daily-level), spatially macroscopic (citywide) collision data.

DATA DESCRIPTION

Figure 1 demonstrates four input data sources used for this prediction. Among them, the Motor Vehicle Collision Information System was used to extract all collisions on public roadways that resulted in property-damage-only collisions of Can$2,000 or greater value, as well as any collisions that resulted in a minor or major injury or fatality (*15*). This database covers collision causes, temporal and spatial information, and injury severity information on collisions involving pedestrians, cyclists, and motorcyclists, and so on.

Historical weather data were downloaded from the Environment Canada website, while qualitative weather forecasts necessary for collision prediction were provided by the research team from the University of Alberta, who applied the Weather Research and Forecasting (WRF) Model calibrated to the city of Edmonton and finally generated a 7-day quantitative weather forecast (*16*).

Finally, all data sources were combined by date to form one unified data set. This data set was separated into three subsets on the basis of time slots applied, respectively, for model fitting, validating, and predicting. The model-fitting data set began on September 1, 2008, the first day that numerical weather data were available, and ended with the beginning of validating of the data set, which was 61 days and ended 2 months before the current date to allow thorough collision data entry for model testing. The predicting data set lasted 67 days and ended up to 6 days before the current date subject to a 7-day weather forecast. One sample-fitting data set [for September 1, 2008, to March 24, 2013 (1,666 days)] is summarized in Tables 1 and 2.

PRELIMINARY DATA INVESTIGATIONS

Daily Collisions Defined as Time Series Data

Daily collision data by nature are time series–type data. A time series is a sequence of observations that are ordered in time (or space). It is either continuous, with an observation at every instant of time, or

TABLE 1 Quantitative Variables and Their Summary Statistics of One Sample Data Set Used for Model Fitting

Quantitative Variable (unit)	Minimum	Maximum	Mean	Standard Deviation
Number of daily collisions	12	305	72.29	34.75
Daily mean temperature (°C)	–39.7	22.8	1.18	12.24
Daily total rainfall (mm)	0	44.8	0.77	2.92
Daily total snowfall (cm)	0	16.8	0.40	1.48

TABLE 2 Categorical Variables and Their Attributes of One Sample Data Set Used for Model Fitting

Categorical Variable	Attributes (Number of Samples)
Holiday (days)	Holiday (54), nonholiday (1,612)
Weekday (days)	Monday–Sunday (238 for each)
Month (days)	January (155), February (141), March (148), April (120), May (124), June (120), July (124), August (124), September (150), October (155), November (150), December (155)
Year (days)	2008 (122), 2009 (365), 2010 (365), 2011 (365), 2012 (366), 2013 (83)

discrete, with an observation at (usually regularly) spaced intervals (*17*). Collision data in this study are discrete time series data with 1 day as the interval.

Time Series Decomposition Analysis

Decomposing a time series means separating it into its components, which are usually a trend component and an irregular component, and if it is a seasonal time series, it has another seasonal component (*18*). Figure 2 shows how the daily collision data were decomposed into the three components; the seasonal part signifies seasonal time series data.

This time series consists of three components:

$$y_t = T_t + S_t + R_t \tag{1}$$

where

y_t = original daily collision data,
T_t = trend component,
S_t = seasonal component, and
R_t = random or error component.

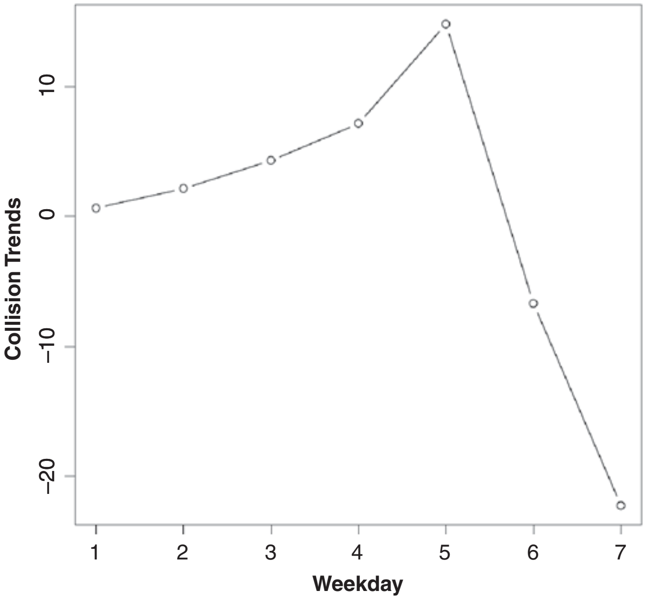


FIGURE 3 Weekly seasonality of daily collision data.

Further decomposition analysis, particularly in relation to seasonality (Figure 3), reveals that the cycle length for daily collision data's seasonality is seven days (i.e., it has a recurring seasonal pattern on weekdays: daily collisions increase on weekdays, peak on Friday, and then drop on the weekend).

MODEL SELECTION AND IDENTIFICATION

A variety of forecasting approaches are available to predict time series data, and no single model is universally applicable (*19*). Two approaches were identified for this study, as described in the following subsections.

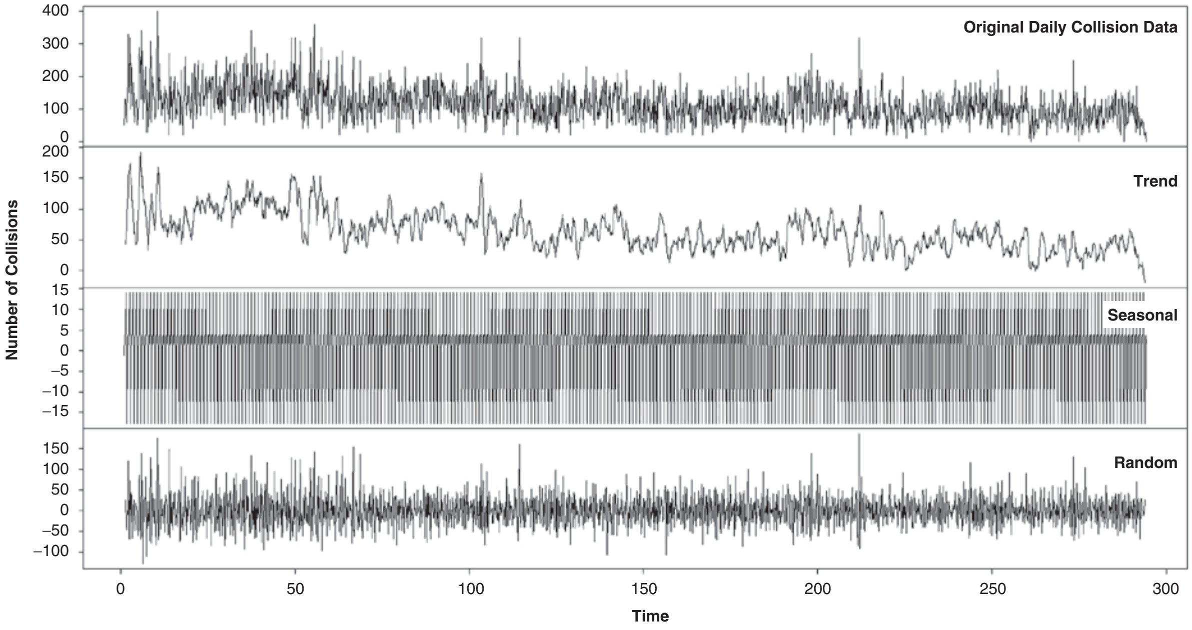


FIGURE 2 Time-serial decomposition analysis on daily collision data.

Seasonal ARIMA Model with External Regressors

For the development of a model that uses available data, a seasonal ARIMA (SARIMA) model is defined below (*19–22*).

$$\underbrace{\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)}_{\mathrm{AR}(p)}\underbrace{\left(1-\beta_1 B^s-\beta_2 B^{2s}-\cdots-\beta_p B^{Ps}\right)}_{\mathrm{AR}_s(P)}$$

$$\underbrace{(1-B)^d}_{I(d)}\underbrace{\left(1-B^s\right)^D}_{I_s(D)} y_t = C+\underbrace{\left(1-\psi_1 B-\psi_2 B^2-\cdots-\psi_q B^q\right)}_{\mathrm{MA}(q)}$$

$$\underbrace{\left(1-\theta_1 B^s-\theta_2 B^{2s}-\cdots-\theta_Q B^{Qs}\right)}_{\mathrm{MA}_s(Q)}\varepsilon_t \qquad (2)$$

where

B = backward shift operator (such that $B \cdot y_t = y_{t-1}$),
$\mathrm{AR}(p)$ = autoregressive part of order p,
$\mathrm{AR}_s(P)$ = seasonal autoregressive part of order P,
$\mathrm{MA}(q)$ = moving average part of order q,
$I(d)$ = differencing of order d,
$I_s(D)$ = seasonal differencing of order D,
$\mathrm{MA}_s(Q)$ = seasonal moving average part of order Q,
C = intercept term,
t = time (e.g., the day for daily collision data),
s = period of seasonal pattern appearing (e.g., 7 days for daily collision data),
ϕ, β, ψ, and θ = model parameters to be estimated, and
ε_t = error term.

Equation 2 is composed of six parts that are, respectively, classified by nonseasonal orders (*p, d, q*), seasonal orders (*P, D, Q*), and seasonal periods. Some circumstances have additional exogenous variables, for example, weather features within daily traffic safety data. These variables are not reflected in any of the six parts in Equation 2.

The SARIMA model containing external factors is abbreviated in this paper as SARIMAX. The SARIMA model can be modified into a SARIMAX in different ways. One approach is to apply the autoregressive and differencing parts of SARIMA exactly the same way to the external variables. Suppose the estimation without external factors, given by Equation 2, is $\hat{y}_t$, and the external time series x_{reg} is x_t, the model response with external regressors, $\hat{y}_{t,x\text{reg}}$, can be modified as follows (*23, 24*):

$$\hat{y}_{t,x\text{reg}} = \hat{y}_t + \gamma \cdot \phi_p(B) \cdot \beta_p\left(B^s\right) \cdot (1-B)^d \cdot \left(1-B^s\right)^D x_t \qquad (3)$$

where

$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$

$\beta_P(B^s) = (1 - \beta_1 B^s - \beta_2 B^{2s} - \cdots - \beta_P B^{Ps})$

and

γ is the parameter of x_{reg} to be estimated.

Conventional Regression Model to Predict Time Series Data

Equation 1 can be rewritten as a regression model (*20*):

$$y_t = T_t + S_t + R_t = \alpha_0 + \sum_{i=1}^{m}\alpha_i U_{it} + \sum_{j=1}^{k}\beta_j V_{jt} + e_i \qquad (4)$$

where

$$T_t = \alpha_0 + \sum_{i=1}^{m}\alpha_i U_{it}$$

and

$$S_t = \sum_{j=1}^{k}\beta_j V_{jt}$$

U_{it} = trend-cycle variables,
V_{jt} = seasonal variable,
t = time,
α_0, α_i, and β = parameters to be estimated,
m = total number of trend-cycle variables,
k = total number of seasonal variables, and
e_i = error term.

The trend-cycle variables can be written as an *m*th-order polynomial in time. In addition, if the seasonal period is *s*, S_t can be a linear function of sine–cosine of various frequencies. As a result, Equation 4 can be rewritten as follows (*20*):

$$y_t = T_t + S_t + R_t = \alpha_0 + \sum_{i=1}^{m}\alpha_i t^i + \sum_{j=1}^{\left[\frac{s}{2}\right]}\left[\begin{array}{l}\beta_j \sin\left(\dfrac{2\pi jt}{s}\right)\\ + \gamma_j \cos\left(\dfrac{2\pi jt}{s}\right)\end{array}\right] + e_i \qquad (5)$$

where [s/2] is the integer portion of s/2.

Use of GLM with NB distribution is a common practice for the regression time series model. Its probability density function is given in Equation 6 (*25, 26*):

$$f\left(y_{it};\alpha,\mu_{it}\right) = \frac{\Gamma\left(y_{it}+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)y_{it}!}\left(\frac{\alpha^{-1}}{\mu_{it}+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu_{it}}{\mu_{it}+\alpha^{-1}}\right)^{y_{it}} \qquad (6)$$

where

y_{it} = response variable for observation i and period t,
μ_{it} = mean response for observation i and period t, and
α = dispersion parameter of Poisson–gamma distribution.

The μ_{it} is structured as shown in Equation 7:

$$\mu_{it} = \exp\left(x_i\beta + e_i\right) = f\left(X;\beta\right) \times \varepsilon_i \qquad (7)$$

where $f(\cdot)$ is function of covariates (X or x_i).

By embedding the independent variables of Equation 5 into Equations 6 and 7 as x_i, a time series GLM model can be developed. This is a common generalized linear estimation that conventional statistical software can handle (*27, 28*). The final model includes temporal variables representing trend and cycle (statistically selected components of Equation 5) and other explanatory variables (results shown in Tables 3 and 4).

MODEL FITTING AND DIAGNOSTICS

Open source statistical software R was applied to run model fittings and diagnostics and to export predictions (*27, 28*).

TABLE 3 GLM Model-Fitting Results

Variable	Coeff.	SE	*Z*-Value	Pr(>| *z* |)
Intercept	4.1160	0.0463	88.9520	<2 e–16
t (time)	–0.0002	0.0000	–13.4750	<2 e–16
sin *t* (sine of *t*)	–0.1215	0.0095	–12.7850	<2 e–16
Weekday (Friday)	0.0000			
Weekday (Monday)	–0.1757	0.0242	–7.2500	4.18 E–13
Weekday (Tuesday)	–0.3559	0.0242	–14.6970	<2 e–16
Weekday (Wednesday)	–0.6333	0.0247	–25.6430	<2 e–16
Weekday (Thursday)	–0.1265	0.0239	–5.2940	1.19 E–07
Weekday (Saturday)	–0.1862	0.0240	–7.7740	7.61 E–15
Weekday (Sunday)	–0.1663	0.0239	–6.9540	3.56 E–12
Holiday (holiday)	0.0000			
Holiday (days in lieu of holidays)	0.3689	0.0991	3.7230	1.97 E–04
Holiday (nonholiday)	0.6059	0.0401	15.1170	<2 e–16
Mean temperature	–0.0160	0.0006	–27.8180	<2 e–16
Rainfall	0.0087	0.0023	3.8020	1.43 E–04
Snowfall	0.0749	0.0043	17.4670	<2 e–16

NOTE: Coeff. = coefficient; SE = standard error; Pr = probability; dispersion parameter = 0.0550.

SARIMAX Model Fitting and Diagnostics

The R function arima was applied to fit SARIMAX models. Relevant diagnostics packages were also used to test model performance (*27, 28*).

The logarithm of the number of daily collisions was applied as a response variable for the model fittings to accommodate the requirement of a continuous outcome of SARIMAX and at the same time adhere to common practices of safety modeling (*6–8*).

Model-Fitting Results

The SARIMAX model was classified by different selections of orders, denoted in this paper as SARIMAX$(p, d, q)(P, D, Q)_s$ (from Equations 2 and 3). Daily collision data have obvious weekly seasonality, so s constantly equals 7. Different combinations of other orders were also attempted.

TABLE 4 ANOVA Test Results

Variable	Linear Regression Chi-square	Degree of Freedom	Pr(>Chi-square)
t	180.66	1	<2.2 e–16***
sin *t*	161.78	1	<2.2 e–16***
Weekday	780.25	6	<2.2 e–16***
Holiday	214.87	2	<2.2 e–16***
Mean temperature	848.3	1	<2.2 e–16***
Rainfall	14.62	1	.0001314***
Snowfall	311.04	1	<2.2 e–16***

$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

The rule for deeming a model significant was stipulated as requiring all parameter estimates to be significant as judged by their p-values at the 10% level. Multiple significant models were ranked by Akaike's information criterion (AIC) (*25, 26, 29*), defined in Equation 8:

$$\mathrm{AIC}(M) = 2\log - \mathrm{likelihood}_{\max}(M) - 2\dim(M) \tag{8}$$

where dim is the number of estimated parameters for model M.

Some preliminary attempts were statistically significant while others were not. Table 5 shows three fittings with different orders but all statistically significant and with the best AIC values. Table 5 shows just one example of a variety of model applications; together with the long-time model applications, the model with the best AIC switched from one to another between these three models, so all three models were considered for next-step studies.

Model Diagnostics

Diagnostic checks for SARIMAX were conducted to analyze the residuals from the model fit to identify any signs of nonrandomness. The R software has the function tsdiag, which, as shown in Figure 4, produces output containing a plot of the residuals, the autocorrelation function of the residuals (ACF), and the p-values of the Ljung–Box statistic for the first 10 lags (*30, 31*).

Figure 4 demonstrates a good model by the following standards (*30–32*):

- The standardized residuals do not show clusters of volatility.
- The ACF shows no significant autocorrelation between the residuals.
- The p-values for the Ljung–Box statistics are all large, an indication that the residuals have no patterns (i.e., residuals are independently distributed).

TABLE 5 SARIMAX Model-Fitting Results

	SARIMAX (1, 1, 1)(1, 0, 1)$_7$			SARIMAX (4, 0, 1)(1, 0, 1)$_7$			SARIMAX (7, 0, 0)(4, 0, 1)$_7$		
Variable	Coeff.	SE	*p*-Value	Coeff.	SE	*p*-Value	Coeff.	SE	*p*-Value
AR1	0.2873	0.0316	<.0001	1.2789	0.0294	<.0001	0.3095	0.0251	<.0001
AR2	na	na	na	–0.2198	0.0402	<.0001	0.0839	0.0255	.0003
AR3	na	na	na	0.0236	0.0399	<.0001	0.0973	0.0255	<.0001
AR4	na	na	na	–0.0881	0.0260	.0002	0.0336	0.0258	.0481
AR5	na	na	na	na	na	na	0.0009	0.0258	.2432[a]
AR6	na	na	na	na	na	na	0.0509	0.0258	.012
AR7	na	na	na	na	na	na	0.1842	0.0664	.0014
MA1	–0.9444	0.0162	<.0001	–0.9626	0.0149	<.0001	na	na	na
SAR1	1.0000	0.0001	<.0001	0.9998	0.0002	<.0001	0.8348	0.0650	<.0001
SAR2	na	na	na	na	na	na	0.1309	0.0482	.0016
SAR3	na	na	na	na	na	na	–0.0472	0.0361	.0476
SAR4	na	na	na	na	na	na	0.0807	0.0270	.0007
SMA1	–0.9926	0.0050	<.0001	–0.9832	0.0081	<.0001	–0.9697	0.0100	<.0001
Intercept	na	na	na	4.1091	0.4966	<.0001	4.1404	0.2549	<.0001
Month	0.0047	0.0048	.0812	0.0073	0.0046	.0282	0.0082	0.0050	.0254
Holiday	–0.5364	0.0329	<.0001	–0.5243	0.0329	<.0001	–0.5257	0.0330	<.0001
Mean temperature	–0.0142	0.0012	<.0001	–0.0133	0.0011	<.0001	–0.0133	0.0011	<.0001
Rainfall	0.0052	0.0021	.0034	0.0055	0.0021	.0023	0.0054	0.0021	.0024
Snowfall	0.0601	0.4410	<.0001	0.0586	0.0040	<.0001	0.0581	0.0040	<.0001

NOTE: na = not applicable. AIC: SARIMAX (1, 1, 1) (1, 0, 1)$_7$ = 47.56; SARIMAX (4, 0, 1) (1, 0, 1)$_7$ = 26.73; SARIMAX (7, 0, 0) (4, 0, 1)$_7$ = 30.14.

[a] Because this is the only coefficient with *p*-value greater than but closer to 10%, the model was still chosen as an overall statistically significant model.

Figure 5, one more test for normality of the residuals, shows that the residuals of the SARIMAX model are normally distributed with a 0 mean, which is a signal that the fitted SARIMAX model was a correctly specified model (*30–32*).

All these diagnostics led to the conclusion that models included in Table 5 were sound and valid.

GLM Time Series Model Fitting and Diagnostics

The R software function glm was applied to fit GLM models with an NB distribution. Relevant diagnostics packages were also used to test model performance (*27, 28*).

Model-Fitting Results

Preliminary fitting attempts included all variables from Table 1. Insignificant variables were eliminated, and models were rerun until the final model was obtained with all statistically significant variables. Table 3 displays time series GLM model-fitting results. A model is considered to be significant if all the parameter estimates are significant at the 10% confidence level. Furthermore, a small dispersion parameter to prove the NB distribution of the response variable is generally expected (*25*).

Model Diagnostics

The function analysis of variance (ANOVA) performs sequential likelihood ratio tests for NB generalized linear models (*33*). ANOVA test results for the fitted GLM model are shown in Table 3. These results suggest that all variables have a strong significant effect on the response variable (i.e., the GLM model was well fitted).

MODEL VALIDATION AND CALIBRATION

As explained in the earlier section on data description, a 61-day period of historical data was applied to validate the four models fitted above. The fitted and actual observed daily collision values are shown in Figure 6.

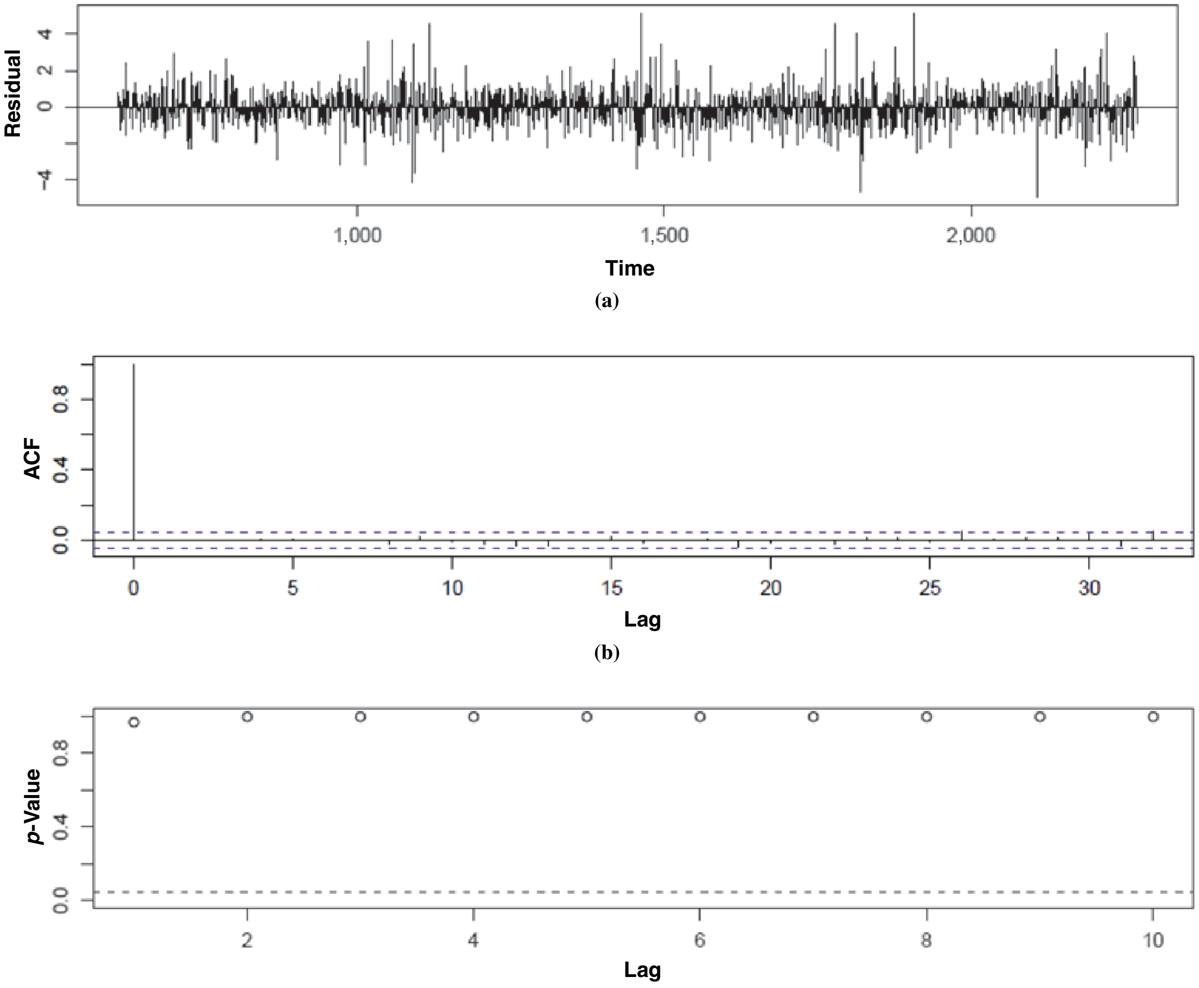
The mean absolute percentage error (MAPE) was applied as a measure to validate model accuracy. It is calculated as shown in Equation 9 (*19*):

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{\text{actual}_t - \text{forecast}_t}{\text{actual}_t}\right| \tag{9}$$

where

actual_t = number of observed collisions on day t,
forecast_t = number of predicted collisions of day t by models, and
n = number of days to be used for validation.

Table 6 enlists all MAPE values from the four fitted models. Among them, SARIMAX (7, 0, 0)(4, 0, 1)$_7$ showed the best performance. However, validation along different time spans led to different results, and no single model consistently gave the best prediction. More generally, Figure 6 shows that, if a horizontal line of collisions = 50 is drawn, the largest predictions are closer to the



(a)

(b)

(c)

FIGURE 4 Diagnostic outputs for SARIMAX (7, 0, 0)(4, 0, 1)$_7$: (*a*) standardized residuals, (*b*) ACF, and (*c*) *p*-values for Ljung–Box statistic.

actual in the top half of the chart, while below this line, the smallest predictions are closer to the actual. Subsequently, a model calibration rule was set up as follows: once all four predicted values are larger than 50, recommend the maximum one as the final prediction; when all four predicted values are smaller than 50, recommend the minimum one as the final prediction; otherwise, the average of the four predicted values is the final recommendation. The MAPE following this calibration rule is also listed in Table 6.

MODEL APPLICATION

The models are used to predict citywide total collisions that are expected to occur in the upcoming 7 days.

Some synthesized predictions, as shown in Figure 7, raised two alarms about high collision predictions: one is the absolute high-collision alarm, and the other is the relatively high-collision alarm (e.g., a weekend day, which would typically have low collisions), both attributable to heavy snowfalls.

Figure 8 shows one extensive application that predicted collisions for two police divisions in the City of Edmonton. During weekdays, Police Division A had higher predicted collisions, but during the weekend, Police Division C had higher predicted collisions. According to this temporal–spatial collision prediction, the police department can prepare its patrol and enforcement schedules and make advanced adjustments to mitigate this safety concern proactively rather than reactively.

CONCLUSION AND FUTURE WORK

After disaggregating daily collision data and analyzing their seasonality, this study explored and developed SARIMAX and GLM time series models with calendar and weather as attributes. The fitted models were validated, and an optimization mechanism was established to recommend predictions for daily collisions. From the methodological perspective, this paper provided some advances beyond those in past research in this area. The contributions include these:

- This paper successfully developed SARIMAX and GLM models with statistical significance and sufficient predictive accuracy despite high randomness of daily collision data. The diversity of

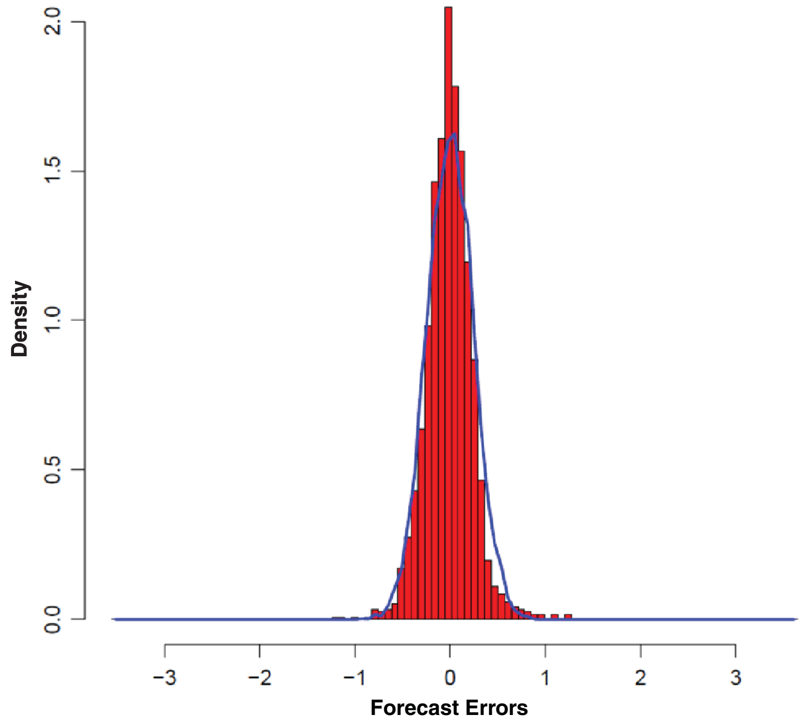


FIGURE 5 Test for normality of residuals for SARIMAX (7, 0, 0)(4, 0, 1)$_7$.

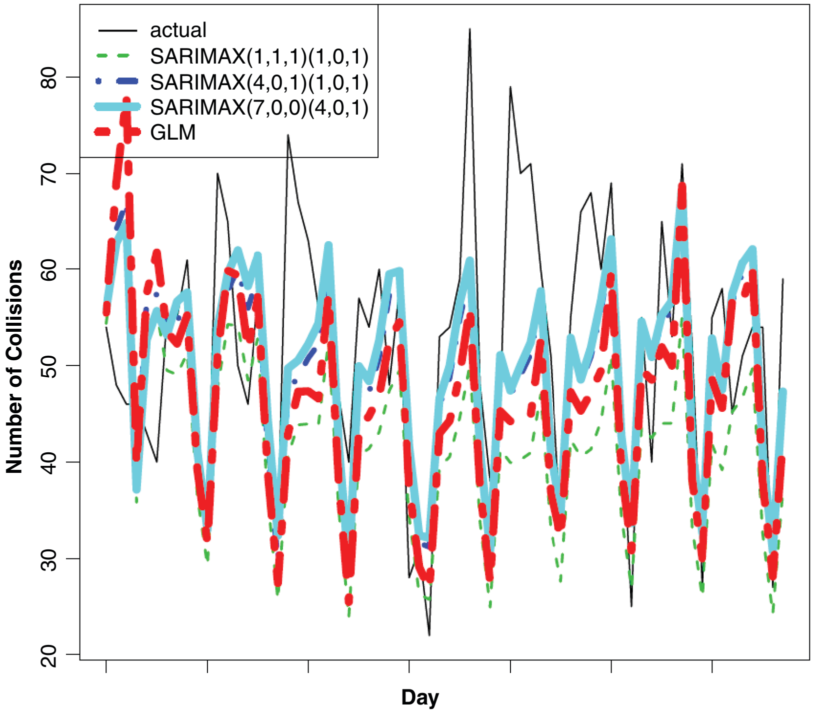


FIGURE 6 Fitted versus actual daily collisions.

TABLE 6 MAPE of Fitted Models

	MAPE, by Model				
Statistic	SARIMAX (1, 1, 1)(1, 0, 1)$_7$	SARIMAX (4, 0, 1)(1, 0, 1)$_7$	SARIMAX (7, 0, 0)(4, 0, 1)$_7$	GLM	Recommended MAPE
Average	23.1	17.1	16.5	19.7	15.6
Day 1–30	20.9	17.6	16.9	19.8	16.2
Day 31–60	25.2	16.7	16.1	19.5	15.1

model selection secures the nonexclusivity of all information and enjoys the applicative flexibility.

- SARIMAX models with weather as external regressors developed in this paper provided the advantage of accounting for seasonality and collision trends as well as the capability of capturing dedicated collision fluctuations from quantitative weather changes.
- All SARIMAX and GLM models developed in this paper yielded promising results; particularly, two of three SARIMAX models displayed better predictive accuracy than the GLM model.

The models developed in this paper are being applied to several City of Edmonton departments. The Edmonton Police Service receives collision-predictive intelligence that is based on both citywide and geographic areas of its deployment to inform scheduling, enforcement, and response. The Transportation Services Department uses collision-predictive models to inform drivers of severe weather and road conditions through a digital messaging sign (DMS) system. The models provide insight into designing the DMS messages, phrases, and contents for notifying drivers of changing conditions. Further model applications will be developed and implemented with more front-line transportation operation areas, communication, stakeholders, law enforcement, and emergency response areas.

This research has limitations and needs future work for improvements. During the data processing stage of this study, the authors failed to incorporate any traffic exposure variables into models for a large area like the city of Edmonton. In the future, traffic exposure data should be collected and embedded into the models for the sake of more robust model performance. Current model calibration is based purely on observation of model outputs. A more robust approach should be established in the future, such as Bayesian model averaging that calibrates forecast ensembles from numerical weather models (*34, 35*).

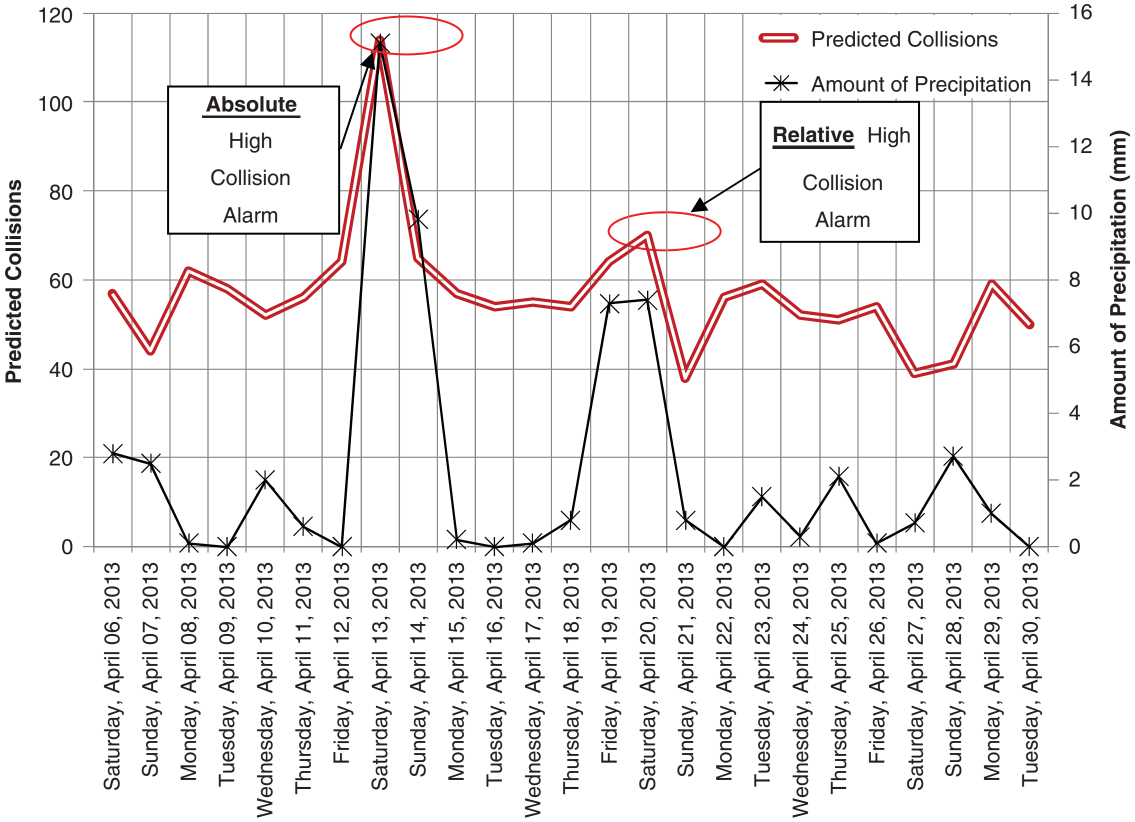


FIGURE 7 Absolute and relative high-collision alarms raised from collision predictions.

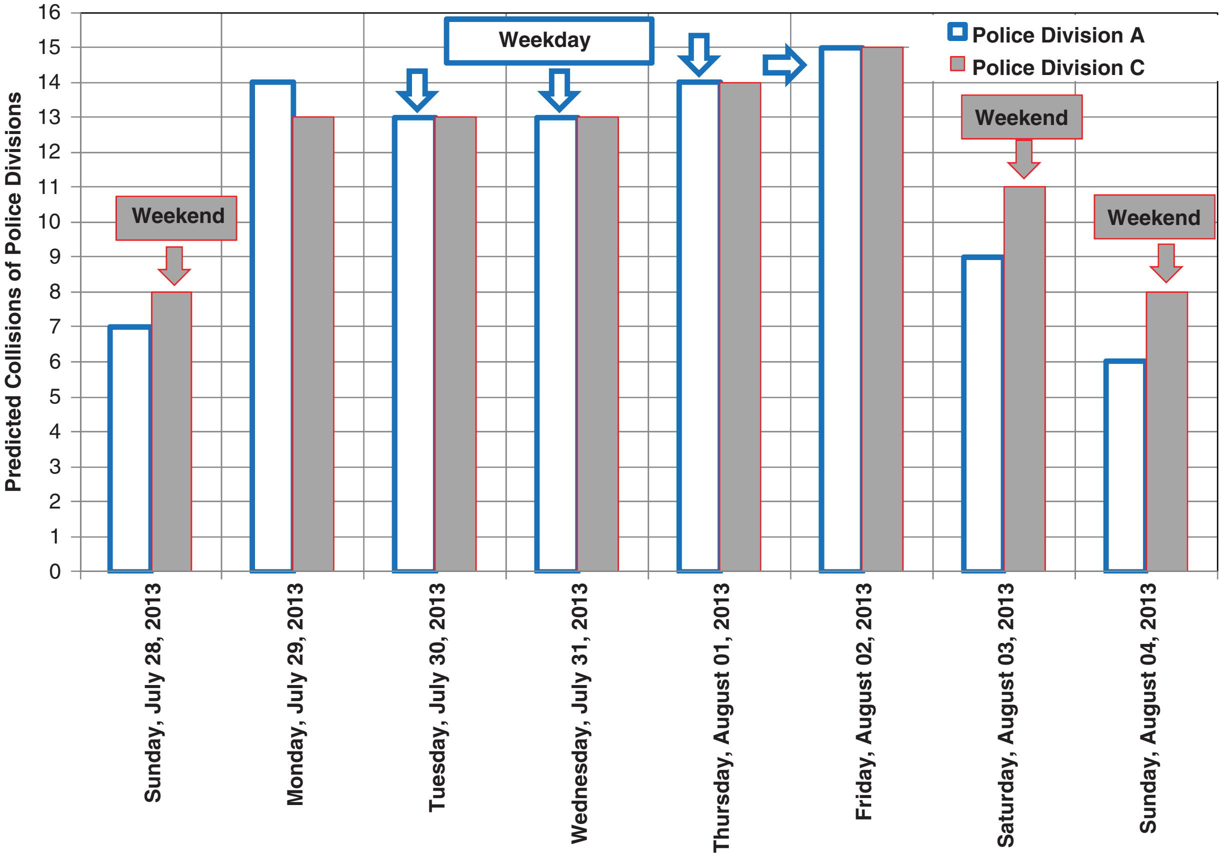


FIGURE 8 Collision predictions by police division.

ACKNOWLEDGMENTS

The authors thank Gerhard Reuter and Clark Pennelly of the Department of Earth and Atmospheric Sciences, University of Alberta, for providing weather forecast data for this study. The authors' gratitude also goes to Karim El-Basyouny of the Department of Civil and Environmental Engineering, University of Alberta, and Shewkar E. Ibrahim of the Office of Traffic Safety, City of Edmonton, for their contributions on paper revisions and to all reviewers for their efforts to improve this paper.

REFERENCES

1. *Road Weather Management Program.* FHWA, U.S. Department of Transportation. http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm. Accessed July 23, 2013.
2. *Surface Transportation Weather Research and Development.* National Center for Atmospheric Research, Boulder, Colo. http://www.ral.ucar.edu/general/press/brochures/surface_transport_3_11.pdf. Accessed July 23, 2013.
3. Andrey, J., B. Mills, and J. Vandermolen. *Weather Information and Road Safety: Final Report.* Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada, 2001.
4. Andreescu, M.-P., and D.B. Frost. Weather and Traffic Accidents in Montreal, Canada. *Climate Research,* Vol. 9, 1998, pp. 225–230.
5. Perry, A.H., and L. Symons. *Highway Meteorology.* E & FN Spon, London, 1991.
6. Hauer, E., and J. Bamfo. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. *Proc., ICTCT 1997 Conference,* Lund, Sweden, Lund University, Sweden, 1997.
7. Hauer, E. *Observational Before–After Studies in Road Safety.* Emerald Group Publishing, Bradford, United Kingdom, 2007.
8. *Highway Safety Manual.* AASHTO, Washington, D.C., 2010.
9. Liu, C., and C.-L. Chen. Time Series Analysis and Forecast of Annual Crash Fatalities. *Research Note DOT HS 809 717,* NHTSA, U.S. Department of Transportation, 2004.
10. Quddus, M.A. Time Series Count Data Models: An Empirical Application to Traffic Accidents. *Accident Analysis and Prevention,* Vol. 40, No. 5, 2008, pp. 1732–1741.
11. Brijs, T., D. Karlis, and G. Wets. Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model. *Accident Analysis and Prevention,* Vol. 40, 2008, pp. 1180–1190.
12. Vanlaar, W., R. Robertson, and K. Marcoux. *Evaluation of the Photo Enforcement Safety Program of the City of Winnipeg: Final Report.* Traffic Injury Research Foundation, Ottawa, Ontario, Canada, 2011.
13. El-Basyouny, K., and D.W. Kwon. Assessing Time and Weather Effects on Collision Frequency by Severity in Edmonton Using Multivariate Safety Performance Functions. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
14. Halim, S., and H. Jiang. The Effect of Operation 24 Hours on Reducing Collisions in the City of Edmonton. *Accident Analysis and Prevention,* Vol. 58, 2013, pp. 106–114.
15. *Motor Vehicle Collisions 2012.* Office of Traffic Safety, City of Edmonton, June 2013. http://www.edmonton.ca/transportation/OTS_Motor_Vehicle_Collisions_2012_Annual_Report.pdf. Accessed July 25, 2013.
16. *Daily Data Reports.* Environment Canada, Government of Canada, Ottawa, Ontario, Canada, July 2013. http://climate.weather.gc.ca/climateData/dailydata_e.html?timeframe=2&Prov=ALTA&StationID=50149&dlyRange=2012-09-01%7C2012-11-08&cmdB1=Go&Year=2013&Month=6&cmdB1=Go. Accessed June 24, 2013.
17. Easton, V. J., and J.H. McColl. Time Series Data. *Statistics Glossary,* Ver. 1.1. http://www.stats.gla.ac.uk/steps/glossary/time_series.html. Accessed July 10, 2012.

18. Coghlan, A. *Time Series 0.2 Documentation.* http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/index.html. Accessed July 3, 2012.
19. Mohamed, N., M. H. Ahmad, Z. Ismail, and Suhartono. Double Seasonal ARIMA Model for Forecasting Load Demand. *MATEMATIKA,* Vol. 26, No. 2, 2010, pp. 217–231.
20. Wei, W. W. S. *Time Series Analysis—Univariate and Multivariate Methods,* 2nd ed. Pearson Education, Inc., New York, 2006.
21. Wilson, S., and R. Dahyot. *Chapter 15: Seasonal ARIMA (p,d,q) $(P,D,Q)_s$*, Lecture Notes of ST 7005: Time Series, Hilary Term, Trinity College, Dublin, Ireland, 2012. https://www.scss.tcd.ie/Rozenn.Dahyot/ST7005/15SeasonalARIMA.pdf. Accessed Aug. 4, 2012.
22. Chatfield, C. *Time-Series Forecasting.* Chapman and Hall/CRC, Baton Rouge, Fla., 2000.
23. Savelainen, A. Construction of SARIMAX—Models Using Matlab. Mat-2.4108. *Independent Research Projects in Applied Mathematics,* Systems Analysis Laboratory, Helsinki, Finland, 2009.
24. Nau, R. F. *ARIMA Models with Regressors.* Decision 411 Forecasting. http://people.duke.edu/~rnau/arimreg.htm. Accessed October 10, 2012.
25. Chen, Y., B. Persaud, E. Sacchi, and M. Bassani. Investigation of Models for Relating Roundabout Safety to Predicted Speed. *Accident Analysis and Prevention,* Vol. 50, 2013, pp. 196–203.
26. Lord, D., P. F. Kuo, and S. R. Geedipally. Comparison of Application of Product of Baseline Models and Accident-Modification Factors and Models with Covariates: Predicted Mean Values and Variance. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2147,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 113–122.
27. Ricci, V. *R Functions for Time Series Analysis,* R.0.5 26/11/04. http://stat.ethz.ch/CRAN/doc/contrib/Ricci-refcard-ts.pdf. Accessed July 27, 2012.
28. ARIMA Modeling of Time Series. *R Documentation for Package "Stats,"* Ver. 2.15.3. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/arima.html. Accessed Aug. 20, 2012.
29. Claeskens, G., and N. L. Hjort. *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, United Kingdom, 2009.
30. Yurekli, K., and A. Kurunc. Testing the Residuals of an ARIMA Model on the Cekerek Stream Watershed in Turkey. *Turkish Journal of Engineering and Environmental Sciences,* Vol. 29, 2005, pp. 61–74.
31. Zucchini, W., and O. Nedadic. *Time Series Analyst with R—Part I.* http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/zeitreihen/sommer03/ts_r_intro.pdf. Accessed April 03, 2012.
32. Running Diagnostics on an ARIMA Model. *Interactive ebooks for iPad, iPhone and the Web.* https://www.inkling.com/read/r-cookbook-paul-teetor-1st/chapter-14/recipe-14-20. Accessed July 30, 2012.
33. Likelihood Ratio Tests for Negative Binomial GLMs. *R Documentation for Package "MASS,"* Version 7.2-35. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/arima.html. Accessed Aug. 22, 2012.
34. Chen, Y., B. Persaud, and E. Sacchi. Improving Transferability of Safety Performance Functions by Bayesian Model Averaging. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2280,* Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 162–172.
35. Vrugt, J. A., C. G. H. Diks, and M. P. Clark. Ensemble Bayesian Model Averaging Using Markov Chain Monte Carlo Sampling. *Environmental Fluid Mechanics,* Vol. 8, No. 5, 2008, pp. 579–595.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.