**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 were 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).
and weather effects on the basis of data from the City of Edmonton, Alberta, Canada, from 2000 to 2010 with multivariate Poisson log-normal 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

<table>
<thead>
<tr>
<th>Quantitative Variable (unit)</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of daily collisions</td>
<td>12</td>
<td>305</td>
<td>72.29</td>
<td>34.75</td>
</tr>
<tr>
<td>Daily mean temperature (°C)</td>
<td>−39.7</td>
<td>22.8</td>
<td>1.18</td>
<td>12.24</td>
</tr>
<tr>
<td>Daily total rainfall (mm)</td>
<td>0</td>
<td>44.8</td>
<td>0.77</td>
<td>2.92</td>
</tr>
<tr>
<td>Daily total snowfall (cm)</td>
<td>0</td>
<td>16.8</td>
<td>0.40</td>
<td>1.48</td>
</tr>
</tbody>
</table>
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:

\[ y_t = T_t + S_t + R_t \]  

(1)

where

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

**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.
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).

\[
\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \right) \left(1 - \beta_1 B - \beta_2 B^2 - \cdots - \beta_s B^s \right) \left(1 - \Psi_1 B - \Psi_2 B^2 - \cdots - \Psi_q B^q \right) y_t = C + \left(1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_q B^q \right) e_t
\]

where

\[
B = \text{backward shift operator (such that } B \cdot y_t = y_{t-1}), \quad \text{AR}(p) = \text{autoregressive part of order } p, \quad \text{MA}(q) = \text{moving average part of order } q, \\
\text{AR}(P) = \text{seasonal autoregressive part of order } P, \quad \text{MA}(Q) = \text{seasonal moving average part of order } Q, \\
I(d) = \text{differencing of order } d, \quad I(D) = \text{seasonal differencing of order } D, \\
\text{AR}(P), \text{MA}(Q) = \text{al seasonal moving average part of order } Q, \\
C = \text{intercept term}, \\
t = \text{time (e.g., the day for daily collision data)}, \\
s = \text{period of seasonal pattern appearing (e.g., 7 days for daily collision data)}, \\
\phi, \beta, \psi, \text{and } \theta = \text{model parameters to be estimated, and } e_t = \text{error term.}
\]

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, \(\hat{y}_{t, reg}\), can be modified as follows (25, 26):

\[
\hat{y}_{t, reg} = \hat{y}_t + \gamma \cdot \phi_p(B) \cdot \beta_p(B') \cdot (1 - B)^d \cdot (1 - B^s)^D x_t
\]

where

\[
\phi_p(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \right) \\
\beta_p(B') = \left(1 - \beta_1 B - \beta_2 B^2 - \cdots - \beta_s B^s \right)
\]

and

\gamma = \text{parameter of } x_{reg} \text{ to be estimated.}

Conventional Regression Model to Predict Time Series Data

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

\[
y_t = T_t + S_t + R_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i U_i + \sum_{j=1}^{n} \beta_j V_j + e_t
\]

where

\[
T_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i U_i
\]

and

\[
S_t = \sum_{j=1}^{n} \beta_j V_j
\]

\[
U_i = \text{trend-cycle variables}, \\
V_j = \text{seasonal variable}, \\
t = \text{time}, \\
\alpha_0, \alpha_i, \text{and } \beta_j = \text{parameters to be estimated}, \\
m = \text{total number of trend-cycle variables}, \\
k = \text{total number of seasonal variables, and } e_t = \text{error term.}
\]

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):

\[
y_t = T_t + S_t + R_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i t^i + \sum_{j=1}^{n} \gamma_j \cos \left(\frac{2\pi jt}{s}\right) + \varepsilon_t
\]

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):

\[
f(y_t; \mu_t) = \frac{1}{\Gamma(y_t + \alpha^{-1})} \left( \frac{\alpha^{-1}}{\mu_t + \alpha^{-1}} \right)^{y_t + \alpha^{-1}} \left( \frac{\alpha^{-1}}{\mu_t + \alpha^{-1}} \right)^{-\alpha^{-1}}
\]

where

\[
y_t = \text{response variable for observation } i \text{ and period } t, \\
\mu_t = \text{mean response for observation } i \text{ and period } t, \text{ and } \\
\alpha = \text{dispersion parameter of Poisson–gamma distribution.}
\]

The \(\mu_t\) is structured as shown in Equation 7:

\[
\mu_t = \exp(x_t \beta + e_t) = f(X; \beta) \times e_t
\]

where \(f(\cdot)\) is function of covariates \((X \text{ or } x_t)\).

By embedding the independent variables of Equation 5 into Equations 6 and 7 as \(x_t\), 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).
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.

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 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) | 0.0659 | 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.

### Table 4 ANOVA Test Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Linear Regression Chi-square</th>
<th>Degree of Freedom</th>
<th>Pr(&gt;Chi-square)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>180.66</td>
<td>1</td>
<td>&lt;2 e–16***</td>
</tr>
<tr>
<td>$\sin t$</td>
<td>161.78</td>
<td>1</td>
<td>&lt;2 e–16***</td>
</tr>
<tr>
<td>Weekday</td>
<td>780.25</td>
<td>6</td>
<td>&lt;2 e–16***</td>
</tr>
<tr>
<td>Holiday</td>
<td>214.87</td>
<td>2</td>
<td>&lt;2 e–16***</td>
</tr>
<tr>
<td>Mean temperature</td>
<td>848.3</td>
<td>1</td>
<td>&lt;2 e–16***</td>
</tr>
<tr>
<td>Rainfall</td>
<td>14.62</td>
<td>1</td>
<td>.0001314***</td>
</tr>
<tr>
<td>Snowfall</td>
<td>311.04</td>
<td>1</td>
<td>&lt;2 e–16***</td>
</tr>
</tbody>
</table>

p < .1; *p < .05; **p < .01; ***p < .001.
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):

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

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) 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff.</th>
<th>SE</th>
<th>$p$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR1</td>
<td>0.2873</td>
<td>0.0316</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR2</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AR3</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AR4</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AR5</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AR6</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AR7</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>MA1</td>
<td>−0.9444</td>
<td>0.0162</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SAR1</td>
<td>1.0000</td>
<td>0.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SAR2</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>SAR3</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>SAR4</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>SMA1</td>
<td>−0.9926</td>
<td>0.0050</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Intercept</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>Month</td>
<td>0.0047</td>
<td>0.0048</td>
<td>.0812</td>
</tr>
<tr>
<td>Holiday</td>
<td>−0.5364</td>
<td>0.0329</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Mean temperature</td>
<td>−0.0142</td>
<td>0.0012</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Rainfall</td>
<td>0.0052</td>
<td>0.0021</td>
<td>.0343</td>
</tr>
<tr>
<td>Snowfall</td>
<td>0.0601</td>
<td>0.4410</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

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.

*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.
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 SARMAX (7, 0, 0)(4, 0, 1).

FIGURE 6  Fitted versus actual daily collisions.
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).

### TABLE 6  MAPE of Fitted Models

<table>
<thead>
<tr>
<th>Statistic</th>
<th>SARIMAX (1, 1, 1)(1, 0, 1)</th>
<th>SARIMAX (4, 0, 1)(1, 0, 1)</th>
<th>SARIMAX (7, 0, 0)(4, 0, 1)</th>
<th>GLM</th>
<th>Recommended MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>23.1</td>
<td>17.1</td>
<td>16.5</td>
<td>19.7</td>
<td>15.6</td>
</tr>
<tr>
<td>Day 1–30</td>
<td>20.9</td>
<td>17.6</td>
<td>16.9</td>
<td>19.8</td>
<td>16.2</td>
</tr>
<tr>
<td>Day 31–60</td>
<td>25.2</td>
<td>16.7</td>
<td>16.1</td>
<td>19.5</td>
<td>15.1</td>
</tr>
</tbody>
</table>

![FIGURE 7  Absolute and relative high-collision alarms raised from collision predictions.](image-url)
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


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