Time series count data models: An empirical application to traffic accidents

Mohammed A. Quddus

Transport Studies Group, Department of Civil and Building Engineering, Loughborough University, Epinal Way/Ashby Road, Loughborough, Leicestershire LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 3 September 2007
Received in revised form 25 January 2008
Accepted 5 June 2008

Keywords:
Traffic accidents
Time series count data
Integer-valued autoregressive
Negative binomial
Accident prediction models

ABSTRACT

Count data are primarily categorised as cross-sectional, time series, and panel. Over the past decade, Poisson and Negative Binomial (NB) models have been used widely to analyse cross-sectional and time series count data, and random effect and fixed effect Poisson and NB models have been used to analyse panel count data. However, recent literature suggests that although the underlying distributional assumptions of these models are appropriate for cross-sectional count data, they are not capable of taking into account the effect of serial correlation often found in pure time series count data. Real-valued time series models, such as the autoregressive integrated moving average (ARIMA) model, introduced by Box and Jenkins have been used in many applications over the last few decades. However, when modelling non-negative integer-valued data such as traffic accidents at a junction over time, Box and Jenkins models may be inappropriate. This is mainly due to the normality assumption of errors in the ARIMA model. Over the last few years, a new class of time series models known as integer-valued autoregressive (INAR) Poisson models, has been studied by many authors. This class of models is particularly applicable to the analysis of time series count data as these models hold the properties of Poisson regression and able to deal with serial correlation, and therefore offers an alternative to the real-valued time series models.

The primary objective of this paper is to introduce the class of INAR models for the time series analysis of traffic accidents in Great Britain. Different types of time series count data are considered: aggregated time series count data where both the spatial and temporal units of observation are relatively large (e.g., Great Britain and years) and disaggregated time series data where both the spatial and temporal units are relatively small (e.g., congestion charging zone and months). The performance of the INAR models is compared with the class of Box and Jenkins real-valued models. The results suggest that the performance of these two classes of models is quite similar in terms of coefficient estimates and goodness of fit for the case of aggregated time series traffic accident data. This is because the mean of the counts is high in which case the normal approximations and the ARIMA model may be satisfactory. However, the performance of INAR Poisson models is found to be much better than that of the ARIMA model for the case of the disaggregated time series traffic accident data where the counts is relatively low. The paper ends with a discussion on the limitations of INAR models to deal with the seasonality and unobserved heterogeneity.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Road transport brings huge benefits to society, but it also has both direct and indirect costs. Direct costs include the costs of providing road transport services such as infrastructure, equipments, and personnel. Indirect costs include road transport accidents, travel delay due to road traffic congestion, and air pollution from road traffic. Among all of these costs, the cost associated with road traffic accidents is very high. According to the UK Department for Transport (DfT, 2003), the value of preventing a fatality (VFP) for the roads is £1.25 million (at 2002 price). Although UK is one of the safest countries in the world in terms of accident per veh-km travelled, the total number of fatalities from road traffic was 3201 in 2005. One of the best ways to understand the causes of road traffic accidents is to develop various accident prediction models which are capable of identifying significant factors related to human, vehicle, socio-economic, road infrastructure, land-use, and the environment. For instance, Noland and Quddus (2004) developed an accident prediction model and reported that the improvements in medical technology and medical care reduced UK traffic-related fatalities. Based on the outcomes of accident prediction models, different countermeasures are implemented to reduce the frequency of road traffic accidents. Accident-forecasting models are used to monitor the effectiveness of various road safety policies that have been introduced to minimise accident occurrences. For example,
Over the last few years, a new class of such time series models known as integer-valued autoregressive (INAR) Poisson models, has been studied by many authors in the fields of finance, public health surveillance, travel and tourism, forest sector, etc. This class of models is particularly applicable to the analysis of time series count data as these models hold the properties of the distribution of count data and are able to deal with serial correlation, and therefore offers an alternative to the real-valued time series models and general Poisson or NB models.

The key objective of this paper is to introduce the class of INAR models for the time series analysis of accident count data from Great Britain. Two types of time series accident count data are considered: (1) aggregated time series data where both the spatial and temporal units of observation are relatively large (e.g., Great Britain and year) and (2) disaggregated time series data where both the spatial and temporal units of observation are relatively small (e.g., congestion charging zone in Central London and month). Various econometric models such as ARIMA, NB, NB with a trend, and INAR(1) Poisson models are used to develop accident prediction models for each dataset. The performance of the INAR(1) Poisson model is compared with the other models.

The rest of the paper is organised as follows. The next section describes the class of INAR models used in this study. This is followed by a description of data sources used for the analysis. A presentation and interpretation of the results are then discussed in some detail. This paper ends with conclusions and limitations of this study.

2. Methodology

The model for continuous autoregressive pure time series data was introduced by Box and Jenkins (1970) and are now very well developed. The Box and Jenkins model such as the seasonal autoregressive integrated moving average (SARIMA) model is capable of taking into account the trend and seasonality (and hence the serial correlation) normally present in time series data. An extension of this model was proposed by Box and Tiao (1975) which has the ability to examine the effects of various exogenous factors as explanatory variables along with the usual trend and seasonal components. This model can be expressed as follows (Hipel and McLeod, 1994):

\[ y_t = \sigma d_t + \beta s_t + \varepsilon_t \]  

In which \( t \) is the discrete time (e.g., week, month, quarter, or year), \( y_t \) is the appropriate Box–Cox transformation of \( Y_t \), say \( \ln Y_t, Y_t^{1/2}, \) or \( Y_t \); itself (Box and Cox, 1964), \( Y_t \) is the dependent variable for a particular time \( t \), \( \varepsilon_t \) is the intervention component, \( X \) is the deterministic effects of independent variables known as control variables and \( \varepsilon_t \) is the stochastic variation or noise component which can be represented by a ARIMA model denoted as ARIMA \((p,d,q)\) for a non-seasonal time series or SARIMA \((p,d,q) \times (P,D,Q)\) for a seasonal time series. In these models, \( p \) is the order of the non-seasonal autoregressive (AR) process, \( P \) is the order of the seasonal AR process, \( d \) is the order of the non-seasonal differencing, \( D \) is the order of the seasonal differencing, \( q \) is the order of the non-seasonal moving average (MA) process, \( Q \) is the order of the seasonal MA process and the subscript \( s \) is the length of seasonality (for example \( s = 12 \) with monthly time series data). The SARIMA model can be expressed as (Box et al., 1994):

\[ N_t = \frac{\theta(B)\phi(B)\varepsilon_t}{\phi(B)\phi(B)^s(1-B)^d} \]  

In which \( \phi \) and \( \Phi \) are the regular and seasonal AR operators, \( \theta \) and \( \Theta \) are the regular and seasonal MA operators, \( B \) and \( B^s \) are the back-
A natural idea of such models is to replace the deterministic term with a stochastic one (see Eq. (3)). The approach developed replaces the scalar multiplication between $\alpha$ and $Y_{t-1}$ by binomial thinning which is defined as follows. If $Y_{t-1}$ is a non-negative integer and $\alpha \in [0,1]$ then

$$\alpha \circ Y_{t-1} \equiv u_1 Y_{t-1} + u_2 Y_{t-1} + \ldots + u_{Y_{t-1} - 1} Y_{t-1} = \sum_{i=1}^{Y_{t-1} - 1} u_i \tag{4}$$

where $\{u_i\}$ is a sequence of independently and identically distributed (IID) Bernoulli random variables, independent of $N$, and for which $\Pr(u_i = 1) = 1 - \Pr(u_i = 0) = \alpha$. It is noticeable that conditional on $Y_{t-1}$, $\alpha \circ Y_{t-1}$ is a binomial random variable, the number of successes in $Y_{t-1}$ independent trials in each of which the probability of success is $\alpha$. Thus, the original real-valued AR(1) model of Eq. (3) is replaced by

$$Y_t = \alpha \circ Y_{t-1} + e_t \tag{5}$$

The thinning operation of $\alpha$ on $Y_{t-1}$ is independent of $e_t$. The second part of Eq. (5) consists of the elements which entered the system during the interval $[t - 1, t]$ known as innovations. The basic derivation of the INAR process is based on the assumption that the innovations, $e_t$, are independently and identically Poisson distribution, i.e., $e_t \sim \text{Poisson}(\lambda_t)$ where $\lambda_t$ is the Poisson mean denoted by

$$\lambda_t = \exp(b \cdot \beta_t + \sigma \cdot g_t) \tag{6}$$

The properties of the model in Eq. (5) can be found in Al-Osh and Alzaid (1987) and McKenzie (1988). The mean and variance of the process $\{Y_t\}$ are equal to $\lambda_t / (1 - \alpha)$. Eq. (5) is termed as the Poisson INAR(1) which assumes that the underlying time series process is a stationary (Al-Osh and Alzaid, 1987; McKenzie, 1988; Brännäs and Hellström, 2001; Hellstrom, 2002).

Extensions of this model includes the Poisson INMA(1), the Poisson INARMA(1,1), the NB INAR(1) model, and the INARMA(1,1) NB model. These may be able to deal with both non-stationary and over-dispersed count data (Al-Osh and Alzaid, 1988; Brännäs and Hall, 2001; Karlis, 2006). Eq. (5) can be estimated using the programmable exact maximum (EM) likelihood algorithm (Karlis, 2006). Other models for time series of counts such as the serially correlated error model (Zeger, 1988) and the Zeger–Qaqish model (Zeger and Qaqish, 1988) can be found in Hellstrom (2002) and Kedem and Fokianos (2002).

### 3. Data

Two datasets are used to investigate the appropriateness of different types of accident prediction models discussed above. One of these is a highly aggregated time series accident count and the other is a relatively disaggregated time series accident count.

The highly aggregated time series data considered in this study is the annual road traffic fatalities in GB between 1950 and 2005 obtained from the UK Department for Transport (DfT, 2006). The total number of observations is 55 and the mean and standard deviation of this time series process are 5769 and 1352, respectively. It is very well known that an accident model should contain an exposure variable to control for total road traffic movements within the road network. The literature suggests that a good exposure variable to accident variable is vehicle-kilometres travelled (VKT). The annual VKT data of GB are then collected from the DfT (DfT, 2006).

Both annual road traffic fatalities and VKT data are shown in Fig. 1. It is interesting to note that annual road traffic fatalities increase with the increase in VKT until 1966. Fatalities are then reduced with the increase in VKT. This is largely due to the implementation of different road safety measures, legislations, and policies over the years. For instance, the UK government introduced the seat-belt safety law in 1983 to reduce the severity of accidents. Penalty points for careless driving, driving with insurance, and seat-belt wearing for child passengers became law in 1989.

The accident prediction model that will be developed using this dataset will also investigate the impact of these two interventions on road traffic fatalities while controlling for VKT.

The disaggregated time series data considered in this study is the monthly car casualties within the London congestion charging (CC) zone between January 1991 and October 2005 (Fig. 2). Casualty data for this zone were taken from the STATS19 national road accident database. The introduction of the congestion charge was postulated to reduce traffic casualties. According to Transport for London (TfL, 2006), there was an overall reduction of about 40–70 casualty crashes a year during the charging hours within the charging zone. This is also noticeable from Fig. 2 that the monthly car casualties reduce after the intervention. It is, therefore, our expectation that accident prediction models that will be developed in this study will discover this fact and will identify the impact of the introduction of the charge on car casualties. The total number of observations is 178 and the overall mean and variance of this time series process is 60.98 and 239.77. The total number of monthly road traffic accidents within greater London will be taken in all models as an exposure to risk of accidents for this dataset.

### 4. Results

Different accident prediction models are developed using the econometric models such as ARIMA or SARIMA, NB, NB with a time
trend, and INAR(1) Poisson models as described in Section 2 for both aggregated and disaggregated time series datasets. Our main objective is to identify the best accident model for each type of time series datasets. For this purpose, each of the datasets is divided into two parts. One part is used to estimate the model parameters and the other part is used to validate the corresponding model using the estimated model parameters. The results for each of the datasets are presented below.

4.1. Annual road traffic fatalities in GB (aggregated time series process)

The first part of the highly aggregated time series process representing the annual road traffic fatalities in GB contains observations from 1950 to 2000 resulting in a total of 51 observations. This part of this time series process, usually known as a training dataset, are used to develop accident prediction models based on ARIMA, NB, NB with a trend, and INAR Poisson models. The rest of the observations (from 2001 to 2005) of this time series process, normally known as a validation dataset, is used to validate the developed accident prediction models. It is obvious from Fig. 1 that this time series exhibits a downward trend suggesting that this dataset is non-stationary. This is also confirmed by the plot of the sample autocorrelation function (ACF) that clearly indicates serial correlation in the data as the autocorrelation coefficients at various lags fall outside the confidence limits (see Fig. 3).

From the plot of the road fatality series in Fig. 1, it is not obvious that a data transformation is required. However, the variance appears to be decreasing from 1988 of this series. A Box–Cox transformation was applied with $\lambda = 0$. Even after this transformation,
there was a downward trend in the series suggesting that the series is non-stationary. This was also confirmed by the Augmented Dickey–Fuller (ADF) test which did not reject the null hypothesis of non-stationarity. This points out that non-seasonal differencing is needed for removing the non-stationary behaviour. After both transformation and differencing, the series became stationary which was also confirmed by the ADF test. However, it was difficult to determine the ARIMA model parameters using both ACF and partial ACF plots. The Box–Jenkins methodology was, therefore, employed to identify the most suitable ARIMA model based on the estimation sample. The values of \( p \), \( q \) were considered up to three and the final model was selected based on the Schwarz information criterion (SIC) with the requirement that all parameters were significant at the 95% confidence level. The final model was ARIMA\((1,1,1)\) suggesting that this stationary time series also has only a non-seasonal AR(1) and a non-seasonal MA(1) components.

It is worthwhile to note that the other models considered in this study such as NB, NB with a time trend, and INAR(1) Poisson models assume that the underlying time series process is a stationary process and therefore, there is no need to manipulate the response variable of the process.

The results of ARIMA, NB, NB with a time trend variable, and INAR(1) Poisson models are presented in Table 1. In each of these models, two interventions and one control variable are used as the explanatory variables and the annual road traffic fatalities in GB is used as a response variable. The first intervention variable is the introduction of the seat-belt law in 1983 and the second intervention variable is the introduction of various safety legislations in 1989. Both of these intervention variables are dummy variables represented by the so-called step functions. This suggests that these interventions cause an immediate and permanent effect on road traffic fatalities in GB. The control variable is the annual VKT in GB.

It can be seen that both intervention variables are statistically significant in all models except in the ARIMA \((1,1,1)\) model. However, both AR1 and MA1 components of this ARIMA model are statistically significant at the 100% confidence level. The control variable, VKT, is also statistically significant in all models expect in the NB with a time trend model. This is due to the fact that the trend variable (linear) and the control variable (i.e., VKT) are highly correlated showing a correlation coefficient of 0.99.

The performance of each of the models presented in Table 1 can be found from the different “measures of accuracy” of the fitted models. These are the mean absolute percentage error (MAPE), the mean absolute deviation (MAD), the mean squared deviation (MSD), and the root mean squared error (RMSE). For all four measures, the smaller the value, the better the fit of the model. It can be seen that the best fitted model is the ARIMA\((1,1,1)\) model in terms of all “measures of accuracy”. The performance of the INAR(1) Poisson model is also good relative to the ARIMA model. The worst performance model is found to be the NB model with a trend model for this dataset.

The validation dataset that contains observations from 2001 to 2005 is used to estimate the relative forecast error, \( RFE \) (%) of each model using the following equation:

\[
RFE = \frac{1}{5} \sum_{i=1}^{5} \left( \frac{\text{abs}(y_i - \hat{y}_i)}{y_i} \right) \times 100
\]

where \( y_i \) is the observed annual road traffic fatalities in GB and \( \hat{y}_i \) is the forecasted annual road traffic fatalities using the developed model.

The results are shown in the last row of Table 1. The lowest \( RFE \) (2.79%) is also found in the ARIMA \((1,1,1)\) model suggesting that the best performance model is the ARIMA \((1,1,1)\) model both in terms of the forecasted values associated with the out of sample observations.

In terms of the significant variables in the models, the two best performance models provide dissimilar results. Both intervention variables are found to be insignificant in the ARIMA model but found to be significant in all other models including the INAR(1) model. Both the seat-belt wearing law in 1983 and the different

---

\( A \) A SARIMA model is not applicable as this dataset is a non-seasonal time series.
safety legislations in 1989 have a negative impact on road traffic fatalities in the UK in the INAR(1) model. This finding is consistent with the finding of other studies on seat-belt safety law (e.g., Houston and Richardson, 2002). However, the application of NB and INAR(1) Poisson models are in-line with the observed fatalities for the ARIMA, NB with a trend, and INAR(1) Poisson models from 1985 to 2005. It can be seen that the predicted fatalities of this differenced series exhibits a downward trend (Fig. 6). This pattern, a seasonal differencing is applied to the series and a plot of the sample autocorrelation function shown in Fig. 5 which indicates that the series is non-stationary as the first twenty autocorrelation coefficients fall outside the 95% confidence limits. Since the original series has a consistent seasonal pattern, a seasonal differencing is applied to the series and a plot of this differenced series exhibits a downward trend (Fig. 6). This suggests that a non-seasonal differencing is also required (Hipel and McLeod, 1994). The final series obtained by one seasonal and one non-seasonal differenced is presented in Fig. 7 which implies that the data are seasonal. This is logical given that the exposure to accidents, the demand for travelling, is highest during the warmer summer months and lowest during the winter months. In addition, the decreasing trend component (from October 2002) indicates that the car casualty data in each month of the year are decreasing over time. This implies that monthly car casualty data within the congestion charging zone have both seasonal and trend components. This is also confirmed with a plot of the sample autocorrelation function shown in Fig. 5 which indicates that the series is non-stationary as the first twenty autocorrelation coefficients fall outside the 95% confidence limits. Since the original series has a consistent seasonal pattern, a seasonal differencing is applied to the series and a plot of this differenced series exhibits a downward trend (Fig. 6). This suggests that a non-seasonal differencing is also required (Hipel and McLeod, 1994). The final series obtained by one seasonal and one non-seasonal differenced is presented in Fig. 7 which implies that the data are stationary. This was also confirmed by the ADF test which rejected the null hypothesis of non-stationarity at the 5% level. However, a closer inspection of Fig. 7 indicates that there is a pattern of excessive changes in sign from one observation to the next, i.e., up-down-up-down. This may mean that the resulting series has been over-differenced. The examination of autocorrelation coefficient at lag 1 of this final series was found to be −0.42 which is negative but not more negative than −0.5 indicating that the resulted series may be “mild over-differenced” which can be compensated for by adding MA terms in the model (Sanchez, 2002).

Table 1
Accident prediction models for annual road traffic fatalities in GB

<table>
<thead>
<tr>
<th></th>
<th>ARIMA (1,1,1)</th>
<th>NB</th>
<th>NB with a time trend</th>
<th>INAR(1) Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seat-belt wearing law</td>
<td>−0.0449</td>
<td></td>
<td>−0.3176</td>
<td>−0.3336</td>
</tr>
<tr>
<td>Veh-km (billion)</td>
<td>0.0031</td>
<td></td>
<td>0.0067</td>
<td>0.0022</td>
</tr>
<tr>
<td>Trend (linear)</td>
<td>−</td>
<td></td>
<td>−0.0017</td>
<td>0.0023</td>
</tr>
<tr>
<td>Non-seasonal AR1</td>
<td>0.9736</td>
<td>14.80</td>
<td>−</td>
<td>8.5157</td>
</tr>
<tr>
<td>Non-seasonal MA1</td>
<td>0.8251</td>
<td>4.97</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

Descriptive statistics

Mean absolute % error (MAPE) 4.16 11.28 11.94 4.73
Mean absolute deviation (MAD) 246.13 636.11 642.23 251.00
Mean squared deviation (MSD) 95475 57104.90 572092.00 101231.10
Root mean square error (RMSE) 308.99 755.71 756.37 318.16
Relative forecast error [%] (Out of sample, 2001–2005) 2.79 23.27 23.52 5.97

The training dataset for this time series process contains observations from January 1991 to December 2004 resulting a total of 168 observations over the 14 years. The validation dataset contains observations from January 2005 to October 2005. The plot of monthly car casualties versus time (Fig. 2) reveals important characteristics about the observations. The ADF test was applied to the original series to investigate whether the series is a stationary series. The hypothesis test does not reject the null hypothesis of random walk without drift at the 5% level suggesting that the series is non-stationary. Therefore, the first and most important step in fitting an ARIMA model is the determination of the order of differencing needed to stationarise the original series.

The sinusoidal curve (albeit relatively weak) in Fig. 2 indicates that the data are seasonal. This is logical given that the exposure to accidents, the demand for travelling, is highest during the warmer summer months and lowest during the winter months. In addition, the decreasing trend component (from October 2002) indicates that the car casualty data in each month of the year are decreasing over time. This implies that monthly car casualty data within the congestion charging zone have both seasonal and trend components. This is also confirmed with a plot of the sample autocorrelation function shown in Fig. 5 which indicates that the series is non-stationary as the first twenty autocorrelation coefficients fall outside the 95% confidence limits. Since the original series has a consistent seasonal pattern, a seasonal differencing is applied to the series and a plot of this differenced series exhibits a downward trend (Fig. 6). This suggests that a non-seasonal differencing is also required (Hipel and McLeod, 1994). The final series obtained by one seasonal and one non-seasonal differenced is presented in Fig. 7 which implies that the data are stationary. This was also confirmed by the ADF test which rejected the null hypothesis of non-stationarity at the 5% level. However, a closer inspection of Fig. 7 indicates that there is a pattern of excessive changes in sign from one observation to the next, i.e., up-down-up-down. This may mean that the final series has been over-differenced. The examination of autocorrelation coefficient at lag 1 of this final series was found to be −0.42 which is negative but not more negative than −0.5 indicating that the resulted series may be “mild over-differenced” which can be compensated for by adding MA terms in the model (Sanchez, 2002).

If the seasonal pattern of the series was ignored and one non-seasonal differenced was applied to the original series then the resulting series also looked stationary (see Fig. 8). This was also confirmed by the ADF test which rejected the null hypothesis of random walk at the 5% level. There may be two different models which fit the data almost equally well.
Therefore, two ARIMA models were estimated based on the data series presented in Fig. 7 (for the case of one seasonal and one non-seasonal differenced) and Fig. 8 (for the case of one non-seasonal differenced). Both ACF and partial ACF plots of the two final differenced series do not exhibit a pattern to identify a suitable SARIMA model. The Box–Jenkins methodology was then employed to identify the most suitable ARIMA models based on the estimation sample. The values of \( p, P, q \) and \( Q \) were considered up to three and the final model for the series shown in Fig. 7 was selected based on the SIC with the requirement that all parameters are significant (at the 95% confidence level). The final model for this series with the lowest SIC value was SARIMA(0,1,1) \( \times (0,1,2)_{12} \) suggesting that this stationary time series also has only a non-seasonal MA(1) and two seasonal MA components such as SMA1 and SMA2. The same methodology was applied to the series shown in Fig. 8. The final model for this series was ARIMA(0,1,1) suggesting that this station-
The seasonally differenced monthly car casualty series for the congestion charging zone (from January 1992 to December 2004, series length = 156).

The SARIMA(0,1,1) × (0,1,2)_{12} model was found to be the most appropriate for the time series of monthly total car casualties within the London congestion charging zone. The SIC values for the SARIMA(0,1,1) × (0,1,2)_{12} and ARIMA(0,1,1) models were found to be 1230.5 and 1278, respectively. Therefore, the SARIMA(0,1,1) × (0,1,2)_{12} model may be considered as the appropriate model for the time series of monthly total car casualties within the London congestion charging zone.

The results of ARIMA(0,1,1), SARIMA(0,1,1) × (0,1,2)_{12}, NB, NB with a time trend, and INAR(1) Poisson models are presented in Table 2. It is worthwhile to note that the sum of the MA coefficients in the SARIMA model is not exactly 1 suggesting that there is no presence of a unit root in the MA part of the model. Each of these models has an intervention variable and a control variable. The intervention variable is the introduction of the London congestion charge in February 2003 which is assumed as a step function. The control variable is the total monthly road traffic accidents in greater London which is a direct measure of exposure to risk (Noland et al., 2006). It can be seen that the intervention variable, the introduction of the congestion charge, is statistically significant in all models. The coefficient value of this variable is found to be −0.31 in the INAR(1) model suggesting that the introduction of the congestion charging zone within central London reduces car casualties by about 27% if all other factors remain constant. The coefficient of the intervention variable in the preferred SARIMA model suggests that there are 13 fewer fatalities (an average 33% reduction) in each month after the introduction of the charge. The control variable is statistically significant in all models. Based on the various “Measures of Accuracy” and “Relative Forecast Error” of the developed models, it can be said that the best performance model is the INAR(1) Poisson model. The RFE calculated using Eq. (7) for the INAR(1) Poisson model is only 0.91%. The worst performance model is the SARIMA model for which the RFE is 1.36%.

In summary, it can be said that for the case of the aggregated time series count data the best accident prediction model is obtained when the real-valued ARIMA model is used and for the case of the disaggregated time series count data the best accident prediction model is achieved when the INAR(1) Poisson model...
model is employed. It should be noted that both time series count datasets used in this study exhibits serial correlation and hence it is not surprising that none of the NB models (with a trend and without a trend) is found to be a suitable model for serially correlated time series count data as these models are unable to take into account the effects of serial correlation. This suggests that the integer-valued discrete property of count data is not so important if the mean of the counts associated with a time series process are high. However, if the counts associated with a time series process exhibit low values, the distribution of count data follows a Poisson distribution and the properties of integer-valued count data becomes important. The INAR(1) Poisson model provides relatively good results for both datasets. Further research is required to fully understand the differences in the performance between ARIMA and INAR models for disaggregated time series data.

### 5. Conclusions

Accident prediction models for time series count data were developed employing a range of econometric models such as ARIMA, NB, NB with a time trend, and INAR(1) Poisson models. Two time series accident count datasets were used to develop the accident models in this study. One of the datasets was a highly aggregated time series process of annual road traffic fatalities in GB and the other dataset was a disaggregated time series process of monthly car casualties within the congestion charging zone. Both of the datasets had a problem of serial correlation. Each of these datasets was used to develop four accident prediction models while controlling for exposure to risk of accidents. The performance of the fitted models was investigated using various “Measures of Accuracy” for within sample observations and “Relative Forecast Error” for out
of sample observations. The results implied that the best accident prediction model for the aggregated time series count data was achieved when the ARIMA model was used. This is due to the fact that this model is able to take into account both serial correlation and non-stationarity normally found in a time series dataset. The performance of INAR(1) Poisson model was also found to be good for this dataset compared with NB models. On the other hand, the best accident prediction model for the disaggregated time series count data was achieved when the INAR(1) Poisson model was used. This largely suggests that the preserving of integer structure of the count data together with the controlling of serial correlation is important if the mean of the counts is relatively low. INAR(1) Poisson model is capable of controlling both properties of time series count data. This suggests that one should consider to employ an INAR model when developing accident prediction models for serially correlated time series count data, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. Further research is needed to fully understand the differences in performance between ARIMA and INAR models when dealing with time series count data exhibiting low mean. However, the ARIMA model has to be correctly specified and other forms of INAR models should be considered.

The INAR(1) Poisson process is a stationary time series process that has a limitation to deal with the presence of over-dispersion commonly found in accident data. The extensions of this model are an INAR(1) NB model or an INARMA(1,1) NB model that could potentially control for both non-stationary time series process and over-dispersion. However, the methods of estimating parameters for such models are very complex and are not readily available to the author to investigate in this study.

Acknowledgement

The author would like to thank Dimitris Karlis from Athens University of Economics and Business and Charles Lindveld from Imperial College London for their invaluable help in estimating the INAR model.

References


