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.

This 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.

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 regressors and intervention variables as explanatory variables along with the usual trend and seasonal components. This model can be expressed as follows:

$$y_t = \varpi_0 I_t + \boldsymbol{\beta X} + \frac{\theta(B)\Theta(B)e_t}{\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D} \quad (1)$$

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 lnY_t, Y_t^2, or Y_t itself (Box and Cox, 1964), Y_t is the dependent variable for a particular time t, I_t is the intervention component, $\boldsymbol{X}$ is the deterministic effects of independent variables known as control variables ($\boldsymbol{X}$), d is the order of the non-seasonal difference, D is the order of the seasonal difference, the subscript s is the length of seasonality (for example s=12 in case of monthly time series data), ϕ and Φ are the regular and seasonal autoregressive (AR) operators, θ and Θ are the regular and seasonal MA operators, B and B^s are the backward shift operators, and e_t is an uncorrelated random error term with zero mean and constant variance (σ^2).

However, the model as shown in equation (1) is suitable for real-valued time series data as the error has to be normally distributed with zero mean and constant variance. Despite this assumption, this model are being used to investigate non-negative variate time series related to a number of applications including road traffic accidents (e.g., Houston and Richardson, 2002 ; Noland et al., 2006).

There are a few major problems with the application of ARIMA models to non-negative integer-valued variables such as monthly accident count data. The first problem is the definition of the model. A real-valued autoregressive process of order 1 can be expressed as follows:

$$Y_t = \alpha Y_{t-1} + e_t \quad (2)$$

In order to obtain an integer valued Y_t the following constraints have to be imposed on equation (2) such as (i) e_t is integer valued and (ii) α =-1, 0 , or 1. Such constraints limit the practical use of real-valued autoregression time series process in the framework of count variables. The second problem concerns the commonly made assumption of normality. For a count variable in which the mean of the counts is relatively high such as

yearly road traffic accidents, the distribution is usually found to be an approximate normal and hence, the use of SARIMA model may be satisfactory as the normality assumption is less questionable. However, for a count variable in which the mean of the count is close to zero such as monthly fatal road traffic accidents within a small geographic unit, the distribution is normally skewed to the right. Therefore, the assumption of normality, or of any other symmetric distribution, is unjustified.

The class of integer-valued autoregressive processes denoted by INAR have been studied by many authors (e.g., Al-Osh and Alzaid, 1987; McKenzie, E., 1988, Brännäs, Hellström, 2001, Karlis, 2006). A natural idea of such models is to replace the deterministic effect of lagged Y_t's by a stochastic one. The approach developed replaces the scalar multiplication between α 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

$$\alpha \circ Y_{t-1} \equiv u_{1,t-1} + u_{2,t-1} + ... + u_{Y_{t-1},t-1} = \sum_{i=1}^{Y_{t-1}} u_i \tag{3}$$

where $\{u_i\}$ is a sequence of independently and identically distributed 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 α. Thus, the original real-valued AR(1) model of equation (2) is replaced by

$$Y_t = \alpha \circ Y_{t-1} + e_t \tag{4}$$

The thinning operation of α on Y_{t-1} is independent of e_t. The second part of equation (4) 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 has an independently and identically Poisson distribution i.e., $e_t \sim Poisson(\lambda_t)$ where λ_t is the Poisson mean denoted by

$$\lambda_t = \exp(\beta X_t + \varpi_0 I_t) \tag{5}$$

The properties of the model in equation (3) can be found in Al-Osh and Alzaid (1987) and MaKenzie (1988). The mean and variance of the process $\{Y_t\}$ are equal to $\lambda/(1-\alpha)$. Equation (4) is termed as the Poisson INAR(1).

Extensions of this model includes the Poisson INMA(1), the Poisson INARMA(1,1), the NB INAR(1) model, and INARMA(1,1,) NB model which has the ability to deal with both under-dispersed and over-dispersed count data (Al-Osh and Alzaid, 1988; Brännäs and Hall, 2001, Karlis, 2006) . Equation (3) can be estimated using the programmable Exact Maximum (EM) algorithm (Karlis, 2006).

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 Nigeria between 1950 to 2005. The total number of observations is 55 and the mean and standard deviation of this time series process are 5,769 and 1,352 respectively. It is very well known that an accident model should contain an exposure to accident variable to control for total road traffic movements within the road network. The literature suggests that a good exposure to accident variable is vehicle kilometres travelled (VKT).

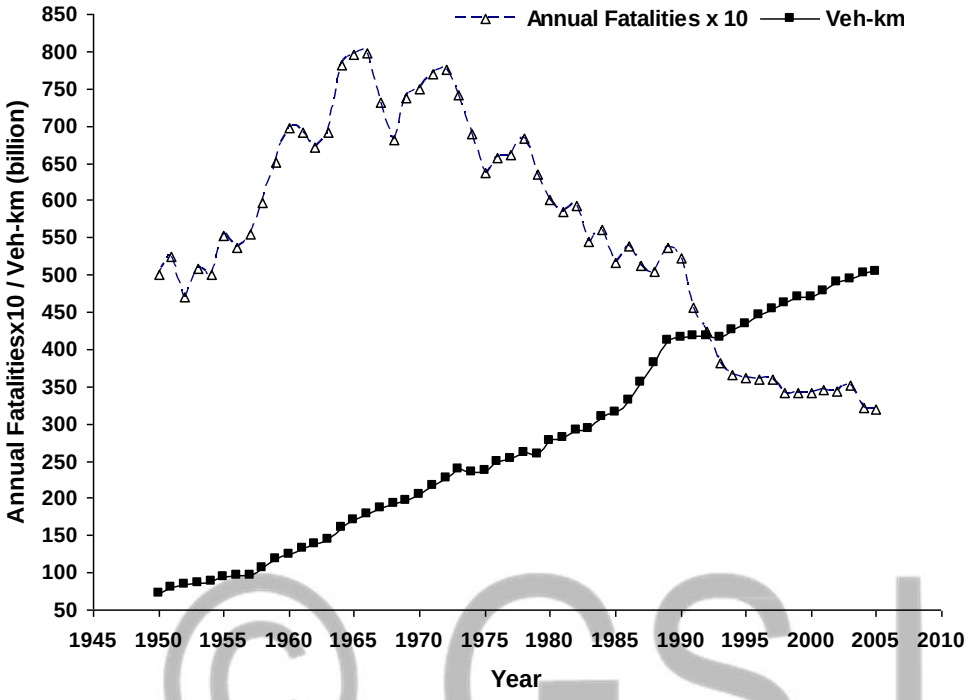
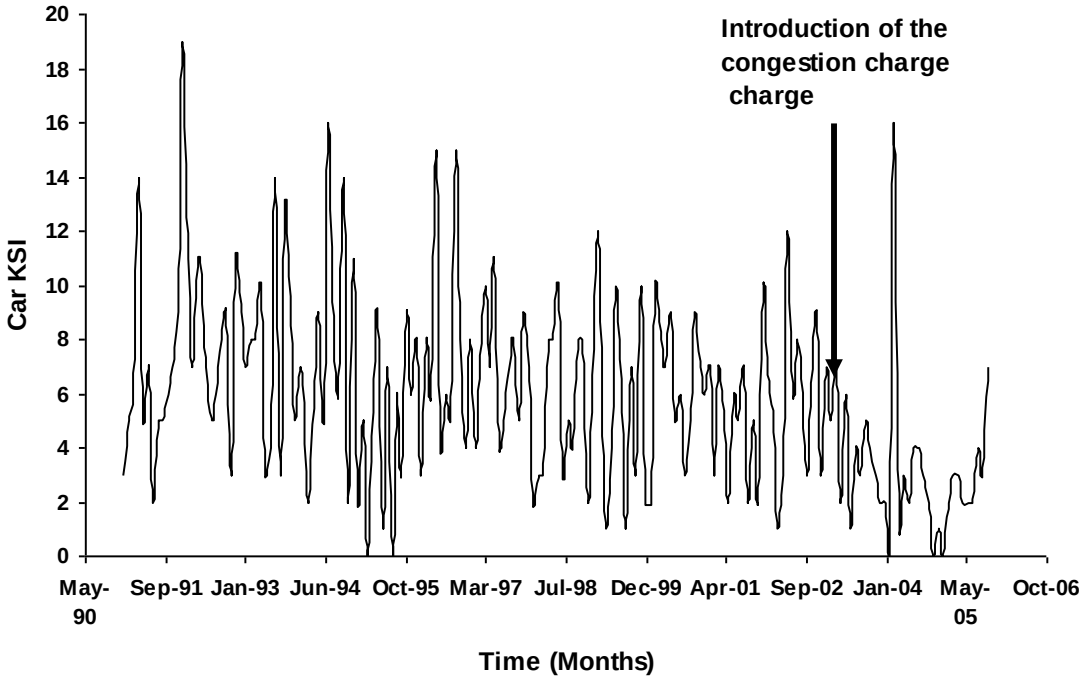


Figure 1: Annual road traffic fatalities and vehicle km travelled

The disaggregated time series data considered in this study is the monthly car KSI (Killed and seriously injured) within the congestion charging zone between January 1991 to October 2005. The time series plot of the data is shown in Figure 2. The introduction of the congestion charge (17th February 2003) is also highlighted within the plot. It is noticeable that the data exhibit both trend and seasonality. The total number of observations is 178 and the mean and standard deviation of this time series process is 6.07 and 3.54. The total number of monthly road traffic accidents within greater is taken as an exposure to risk of accidents for this dataset.



RESULTS

Different accident prediction models are developed using the econometric models such as ARIMA, NB, NB with a time trend, and INAR Poisson models as described in the methodology section 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 fit a model and the other part is used to validate the corresponding model. The results for each of the datasets are presented below.

Annual Road Traffic Fatalities in Nigeria (Aggregated Time Series Process)

It is worthwhile to note that the other models considered in this study such as NB, NB with a time trend, and INAR 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 Poisson models are presented in Table 1. In each of these models, two interventions and one control variables are used as the explanatory variables and the annual road traffic fatalities 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.

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.

Table 1: Accident prediction models for annual road traffic fatalities.

Aggregate Time Series Accident Count Data								
	ARIMA (1,1,1)		NB		NB with a time trend		INAR(1) Poisson	
Explanatory Variables	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Seat-belt wearing law	-0.0449	-0.84	**-0.3176**	**-3.94**	**-0.3336**	**-4.00**	**-0.3942**	**-3.65**
Newlegislation on safety	0.0273	0.46	**-0.3588**	**-4.65**	**-0.4186**	**-3.57**	**-0.4236**	**-2.95**
Veh-km (billion)	**0.0031**	**2.48**	**0.0007**	**2.12**	0.0022	1.01	**0.0023**	**1.89**
Trend (Linear)	-	-	-	-	-0.0107	-0.68	-	-
Constant	-	-	8.6481	131.14	8.5765	-0.68	8.5157	-1.44
Non-seasonal AR1	**0.9736**	**14.80**	-	-	-	68.97	-	-
Non-seasonal MA1	**0.8251**	**4.97**	-	-	-	-	-	-
Descriptive statistics								
Overdispersion parameter	-	-	0.0183	5.01	0.0181	5.01	-	-
Thinning parameter	-	-	-	-	-	-	0.1250	3.02
Series of length	51.00		51		51		51	
Number of residuals	50.00		51		51		51	
Log-likelihood	76.59		-410.94		-410.71		-406.21	
Accuracy of the fitted models (within sample)								
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)	**95475.05**		571104.90		572092.00		101231.10	
Root Mean Sqaure 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 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 models using the following equation:

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

where, y_i is the observed annual road traffic fatalities 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 the INAR(1) model. Both the seat-belt wearing law in 1983 and the different safety legislations in 1989 have a negative impact on road traffic fatalities in the NIGERIA 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).

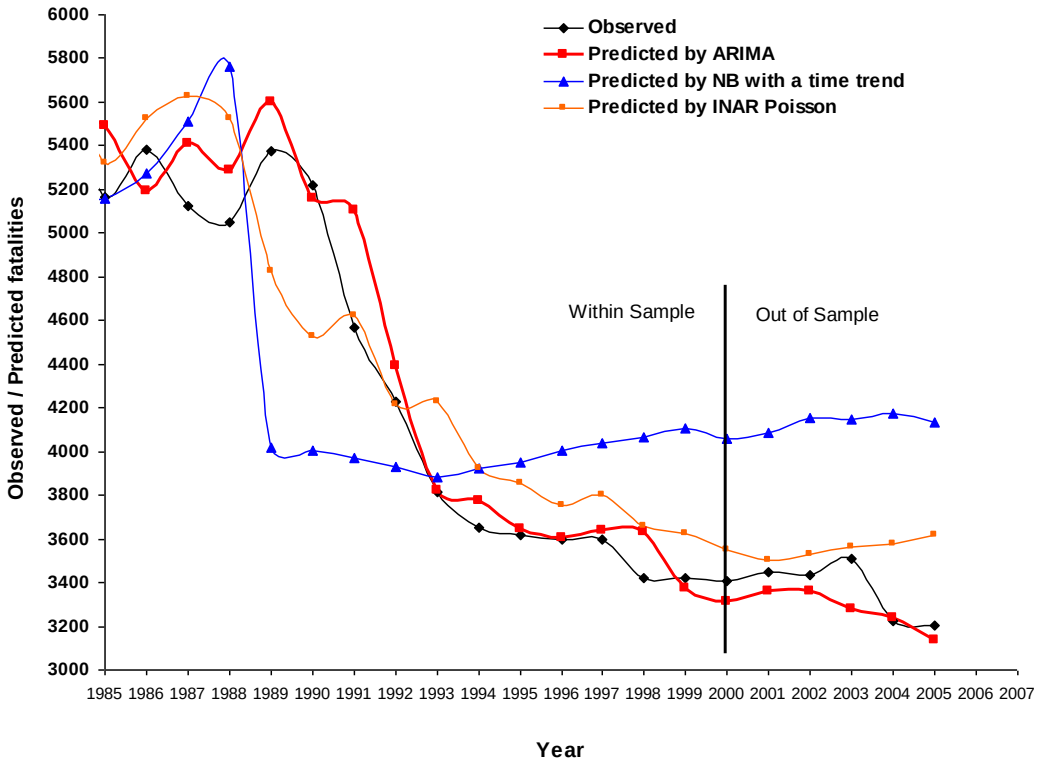


Figure 4 shows the graph of observed fatalities and predicted fatalities for the ARIMA, NB with a trend, and INAR(1) Poisson models from 1985 to 2005. It can be seen that the predicted fatalities of the ARIMA and INAR(1) Poison models are in-line with the observed fatalities for both within sample and out of sample observations. As expected, NB model with a time trend variable provides the worst fit.

The results of SARIMA, NB, NB with a time trend, and INAR(1) Poisson models are presented in Table 2. Each of these models has an intervention variable and a control variable. The intervention variable is the introduction of the congestion charge in February 2003 which is assumed as a step function. The control variable is the total monthly road traffic accidents which is a direct measure of exposure to risk. It can be seen that the intervention variable, the introduction of the congestion charge, is statistically significant in all models except in the SARIMA model. The coefficient value of this variable is found to be -0.41 in the INAR(1) model suggesting that the introduction of the congestion charging zone within central reduces car KSI by about 33% if all other factors remain constant. The control variable is statistically significant in the INAR(1) Poisson model only. 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 for the INAR(1) Poisson model is only 2.21%. The worst performance model is the SARIMA model for which the RFE is 9.03%.

Table 2: Accident prediction models for monthly car KSI within the congestion charging zone

Disaggregate Time Series Accident Count Data (Monthly Car KSI within the Congestion Charging Zone 1991 - 2004)								
	SARIMA		**NB**		**NB with a time trend**		**INAR(1) Poisson**	
Explanatory Variables	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Congestion charge	-1.4505	-1.12	**-0.4251**	**-2.25**	**-0.3191**	**-1.63**	**-0.4081**	**-2.50**
ln(Monthly Accidents)	1.5802	0.32	0.8363	1.59	0.5163	0.95	**0.8372**	**1.95**
Time trend (Linear)	-	-	-	-	**-0.0021**	**-1.9**	-	-
Constant	-		-4.8670	-1.15	-2.1394	-0.48	-4.9782	-1.44
Non-seasonal MA1	**0.9928**	**3.87**	-	-	-	-	-	-
Seasonal MA1	**0.8933**	**7.76**	-	-	-	-	-	-
Descriptive statistics								
Overdispersion parameter			0.1397	4.02	0.1326	3.9		
Thinning parameter							0.0973	2.03
Series of length	168		168					
Number of residuals	155		168					
Accuracy of the fitted models (within sample)								
Mean Absolute % error (MAPE)	22.38		16.98		17.21		**7.59**	
Mean Absolute Deviation (MAD)	1.75		0.96		1.01		**0.64**	
Mean Squared Deviation (MSD)	6.15		5.45		5.56		**2.33**	
Root Mean Sqaure Error (RMSE)	2.48		2.33		2.36		**1.53**	
Relative forecast error (%) (Out of sample, Jan 2005 to Oct 2005)	9.03		5.12		5.31		**2.21**	

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 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. This is confirmed by the

results of the disaggregated time series data while the real-valued time series model provides the worst performance among all models. The INAR(1) Poisson model provides good results for both datasets.

In terms of identifying the effects of interventions, the ARIMA model provides an unrealistic result for both time series datasets. The exact causes have not been identified. However, one of the reasons may be that the AR and MA components of this model weaken the impact of interventions.

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 and the other dataset was a disaggregated time series process of monthly car KSI 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 based on the four econometric 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. The performance of INAR(1) Poisson model was also found to be good for this dataset. 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 controlling of both serial correlation and non-negative discrete property of count data are important when the mean of the counts is relatively high. The preserving of integer structure of the count data is more important than the controlling of serial correlation if the mean of the counts is relatively low. Since INAR(1) Poisson model is capable of controlling both properties of time series count data, one should consider to employ this model when analysing time series accident count data.

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.

REFERENCES

Abdel-Aty, M., Radwan E., 2000, Modeling Traffic Accident Occurrence and Involvement. Accident Analysis and Prevention 32(5), 633-642.

Al-Osh, M., Alzaid, A.A., 1987, First-order integer-valued autoregressive (INAR (1)) process. Journal of Time Series Analysis 8, 261–75.

Al-Osh, M., Alzaid, A.A., 1988 Integer-valued moving average (INMA) process. Statistical Papers 29, 281–300.

Alzaid, A. A., Al-Osh, M.,1990, An integer-valued pth-order autoregressive structure (INAR(p)) process. Journal of Applied Probability 27, 314–23.

Box, G. and Jenkins, G., 1970, *Time series analysis: Forecasting and control*, San Francisco: Holden-Day.
Box, G.E.P., Tiao, G.C., 1975, Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70-74.

Brännäs, K., Hall, A., 2001, Estimation in integer-valued moving average models. Applied Stochastic Models in Business and Industry 17, 277–91.

Brännäs, K., Hellström, J., 2001, Generalized integer-valued autoregression. Econometric Reviews 20, 425–43.

Chin, H.C., Quddus, M.A., 2003, Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections, Accident Analysis & Prevention, 35(2), 253-259.

DfT (Department for Transport), 2003, Highways Economics Note No1. 2002 - Valuation of the benefits of prevention of road accidents and casualties. Department for Transport, NIGERIA.

DfT (Department for Transport), 2006, Transport statistics Great Britain, 32nd Edition, London: TSO.

Goh, B.H., 2005, The dynamic effects of the Asian financial crisis on construction demand and tender price levels in Singapore, Building and Environment, 40: 267-276.

Houston, D.J., Richardson, L.E., 2002, Traffic safety and the switch to a primary seat belt law: the California experience, Accident Analysis and Prevention 34: 743–751.

Karlis, D., 2006, Time series model for count data, Paper presented at the Annual Conference of the Transportation Research Board, Wahsington, D.C.

Kulmala, R., 1995, Safety at Rural Three-and Four-arm Junctions: Development and Application of Accident Prediction Models. VTT publications. Espoo: Technical Research Center at Finland.

Land, K.C., McCall, P.L., Nagin, D.S., 1996, A Comparison of Poisson, Negative Binomial and Semi-parametric Mixed Poisson Regressive Models with Empirical Applications to Criminal Careers Data. Sociological Methods and Research, 24, 387-442.

McKenzie, E. , 1988, Some ARMA models for dependent sequences of Poisson counts. Advances in Applied Probability 20, 822–35.

Noland, R. B., Quddus, M.A, 2004, Improvements in Medical Care and Technology and Reductions in Traffic-related Fatalities in Great Britain, Accident Analysis and Prevention, 36(1): 103-113.

Noland, R.B., Quddus, M.A. and Ochieng, W.Y., 2006, The effect of the congestion charge on traffic casualties in London: an intervention analysis, Presented at the Transportation Research Board (TRB) Annual Meeting, Washington, D.C., USA, January.

Sharma, P., Khare, M., 1999, Application of intervention analysis for assessing the effectiveness of CO pollution control legislation in India, Transportation Research Part D 4: 427-432.

Zimring, F., 1975, Firearms and Federal law: the Gun Control Act of 1968. Journal of Legal Studies 4, 2 (January): 133-198.