NCRST
NATIONAL COMMISSION
ON RESEARCH SCIENCE & TECHNOLOGY

Namibian Journal for
Research, Science and
Technology

Volume 4 Issue 1
May 2022
P-ISSN: 2026-8548

Original Research Article

Application of Longitudinal Analysis to Crime Data: Windhoek Case study (2011-2016)

Lazarus Unandapo[1,*], Augustinus Ntjamba[1]

[1]Department of Statistics and Population Studies, University of Namibia, Windhoek, Namibia
*Corresponding author: lunandapo@unam.na

ARTICLE INFO

Received: March 2021
Accepted: April 2022

Keywords:
Longitudinal data analysis,
Crime analysis,
Windhoek,
Namibia

ABSTRACT

Crime in the Windhoek municipal area continue to be on an increase trend as the population grow over time. Despite past effort done to reduce crime, crime seems to be on a continuously increasing trend; mostly in area regarded as crime hotspots by Windhoek municipal Police. Past study done to analyse the crime record have concentrated mostly on cross-sectional analysis, which does not take correlation into account, thus makes it difficult to compare snapshots of crime over time. The main aim of this research was to analyse reported crime data for the period (2011-2016), using a more robust method known as longitudinal data analysis. This method helped us to visualise the different crime frequencies at different time points (month, day or time of day) in all identified police zones. Furthermore, the use of Generalised Estimating Equations (GEE) was also done, to model these crime data, where the best correlation structure was identified to be the exchangeable correlation structure, which assume constant correlation over time.

1. Introduction

Windhoek, the capital city of Namibia, is developing and the locations are expanding due to people migrating from several rural areas and other towns in the country to seek for better living. This is mainly due to the fact that the city is a big attraction to foreign investment and tourist which attribute to better living condition and employment opportunities as compared to many other towns in the country. As the population increases, so is the unequal distribution of wealth, which then leads to higher poverty rate. As a developing country, crime in Namibia fluctuates overtime. According to Neema and Böhning (2012) developing countries tend to have quite a high rate of crime due to unfavourable prevailing socio-economic conditions, high unemployment levels, and lack of organised policy and justice systems among others. The crime in Namibia especially in the capital city appears alarming. It is hard to note if crime is on increase, decrease or constant.

Research conducted on crime in Namibia have often bear unsatisfactory results due to the inability to compare types of crime on a longitudinal platform. The Namibian Police Force (Nampol) reported an upward trend in overall crime statistics for Wanaheda police station for a period of five year (2014/15-2018/19) (Lilungwe 2020).

Lilungwe (2020) further augured that the most prevalent crime reported in Khomas region are theft, assault, common assault, housebreaking and robbery. Crime such as theft, housebreaking, and robbery are referred to as property crime while assault, common assault is referred to as violent crime in the study. It was found that bad economies at times lead to more property crime as criminals steal popular items that they cannot afford (Kathena and Sheefeni, 2017). They further argue that hard economic times results in more domestic violence and greater consumption of mind- altering substances such as drugs and alcohol and in return this result even in more crime. The study of crime data is not only unique to Namibia, as an example, Tang (2011) researched on the dynamic relationship between tourist arrivals, inflation, unemployment and crime rates in Malaysia. Throughout almost two decades since 1990, reducing crime in Malaysia has been viewed as an urgent task for the policymakers and the royal Malaysia policy (RMP), because of the deep impact on socio-economic development.

It has been a challenge to measure change of a particular variable of interest over time. Longitudinal research employs continuous or repeated measures over time to follow a group of subjects or samples from the same population. In most cases longitudinal studies, time can be measure as calendar time, day, months, years, decades, whichever suit the situation better. Two important characteristics of change need to be captured in longitudinal study, namely within-unit change across time, or growth trajectory, and inter-unit differences in change. It is vital to be precise about variables that are expected to change and the reason for changing overtime (Ployhart and Vandenberg 2010).

To study how crime would relate to the hotspot overtime, using repeated measurement, the substantive meaning of change and the time need to be theorised, prior to the actual crime analysis. Furthermore, it is clear that crime does not change, revolve, or develop because of time; rather it does so overtime. The longitudinal approach attempts to describe the form of change overtime, and this can be either linear or nonlinear. Some features distinguishing longitudinal studies includes correlated observations (due to the variable measurements at multiple time points), high possibility of missing data (due to the rigorous follow-up needed for each subject) and the existence of multiple covariates. Generalised Estimating Equation (GEE) is a general statistical approach to fit a marginal model for longitudinal/clustered data analysis (Wang and Carey, 2014). The method is popular into the field of medicine and psychology, for instance, orthodontic measurements on children of different ages and response is the measurement of the distance from the centre of the pituitary to the pterygomaxillary fissure, measure at a different age repeatedly. The primary goal was to understand the patterns of crime over the study period and investigate how crime has change. The GEE model makes use of a population- averaged estimates where the quasi-likelihood function approach applies. The method develops as a means of testing hypothesis about the effect of factors on binary ad other exponentially distributed response variables.

2. Methods

A quasi-likelihood estimator, as defined by Zeger and Liang (1986) is a solution to the score-likelihood equation system given below:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right]^T V(\hat{\alpha})_i^{-1}\left(y_i - \mu_i(\boldsymbol{\beta})\right) = 0 \qquad (1)$$

where $\boldsymbol{y} = (y_1, \dots, y_n)$ is a vector of outcomes (y_i) variable decomposed into $\boldsymbol{n}$ strata with μ_i an expected value of y_i given as:

$$\mathrm{E}(y_i) = h^{-1}\left(X_i(\boldsymbol{\beta})\right) \qquad (2)$$

X = ($\boldsymbol{X}_1$, ..., $\boldsymbol{X}_n$) is an *n x p* design covariate matrix of predictor variables decomposed into *n* strata and $\boldsymbol{\beta}$ its an *k x* 1 vector of regression parameters. Here *p* is the dimension of each of the strata and *k* is the dimension of the vector of regression parameters. According to Zeger and Liang (1986), Zorn (2001), $\boldsymbol{h}$ is a link function, which specifies the relation- ship between *E*($\boldsymbol{y}_i$) and the $\mathbf{X}_i$. This function transforms the expectation of the response variable μ_i to linear predictors, e.g. $h(\mu_i) = X_i(\boldsymbol{\beta})$. $V(\hat{\alpha})_i$ is the variance of $\boldsymbol{y}_i$ given as a known function *g* of *E*($\boldsymbol{y}_i$), e.g. $V(\hat{\alpha})_i = g(\mu_i)\varphi$ where φ is a scale parameter and $\hat{\alpha}$ is a consistent estimate of α (Zorn 2001). The solution to Equation (1) can be obtained by the method of iteratively re-weighted least squares (IRWLS) as stated by Zorn (2001), Zeger and Liang (1986), and Millar (2011). According to Crowder (1995) specifications of the correlation between the $\boldsymbol{y}_i$ can be avoided by assuming a prior working correlation matrix (working correlation structure) $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ when re- peated measurements are analysed using GEE models. Here $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ is a fully specified vector of unknown regression parameters (Weiss, 2005). The choices of working correlation matrix include independent working correlation matrix, exchangeable working correlation matrix, first order auto-regressive (AR1) working correlation matrix and unstructured working correlation matrix among others. Each $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ has its own assumptions, for example, the independent $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ assumes zero correlation between the subsequent measurements, exchangeable $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ assumes constant correlation across all observations in a strata (in this case seasons), while AR1 $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ assumes that two measurements taken one time point away within a strata tend to be highly correlated than two observations taken far apart in the same strata. See Weiss (2005), Pan and Connett (2002), Zorn (2001), Cui and Qian (2007), Wang and Carey (2003) for more choices of $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$.

Given $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ for response vector $\boldsymbol{y}$, Pan and Connett (2002), Zorn (2001), Zeger and Liang (1986), expressed the covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\alpha}})$ in terms of the correlation matrix $\boldsymbol{V}(\hat{\boldsymbol{\alpha}})$ as:

$$V = V(\widehat{\alpha}) = A^{1/2} R(\widehat{\alpha}) A^{1/2} \qquad (3)$$

where **A** = *diag*(V(y_1), V(y_2), ..., V(y_p)) better link A and $V(y_i)$ is a diagonal matrix with $V(y_i) = V(\mu_i)$. The extension of Equation (1) to longitudinal data is expressed as:

$$S(\beta) = \sum_{i=1}^{n} D_i^T \, V(\alpha)_i^{-1}(y_i - \mu) = 0 \qquad (4)$$

with $D_i = D_i(\beta)$ the partial derivative of μ_i with respect to β. When n = 1, Zeger and Liang (1986) note that Equation (4) reduces to the quasi-likelihood estimation. They further state that when the link function h is correctly specified, the GEE (4) give consistent regression coefficients. Equation (4) is a score equation for $\boldsymbol{\beta}$, and depends on both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (Zorn, 2001; Zeger and Liang, 1986).

Zeger and Liang (1986) replaced $\boldsymbol{\alpha}$ with some $K^{1/2}$ consistent estimator, $\hat{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{\beta}, \varphi)$, in Equation (3) and (4) to express the two equations as functions of $\boldsymbol{\beta}$ only. They also replaced the scale parameter φ in $\hat{\boldsymbol{\alpha}}$ by $K^{1/2}$ consistent estimator, $\varphi(\boldsymbol{y}, \boldsymbol{\beta})$, so that the estimate $\hat{\beta}$ of $\boldsymbol{\beta}$ is expressed as a solution to:

$$\sum_{i=1}^{n} \mathrm{m} U_i\{\beta, \hat{\alpha}[\beta, \hat{\phi}(\beta)]\} = 0. \qquad (5)$$

with $\boldsymbol{U} = \boldsymbol{D}^{\boldsymbol{T}} \; \boldsymbol{V}^{\boldsymbol{-1}} \; \boldsymbol{S}$ as a function of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When K increases to infinity, $\hat{\beta}$ becomes a consistent estimator of β *and* $K^{1/2}(\hat{\beta} - \beta)$ becomes a multivariate Gaussian with covariate matrix V_β, which consistently estimate the variance (Zeger and Liang 1986; Oh et al. 2008):

$$V_\beta = \underbrace{K(\Omega)^{-1}\left[\sum_{i=1}^{n} D_i^T \, V_i^{-1} cov(y_i) V_i^{-1} D_i\right](\Omega)^{-1}}_{limit\ as\ K \to \infty}$$

solving the GEE for $\hat{\beta}$, one first has to solve for the regression coefficients, the correlation $\boldsymbol{\alpha}$ and scale parameter φ. If we are given an estimate of working correlation matrix $\boldsymbol{R}(\hat{\boldsymbol{\alpha}})$ and scale parameters φ, then $\boldsymbol{\beta}$ can be calculated by IRWLS method. If the V_i is reasonably approximated, then the estimates of $\boldsymbol{\beta}$ is efficiently relative to ML estimates.

3. Results

Figure 1 shows the crime trends across each Zone for the combined period 2011-2016. Overall, property crime was highest in Zones 1 and 14, while Zones 1 and 6 showed an increasing trend in violent crime as shown in figure 1. Moreover, there remaining crime types recorded low values across all 19 Zones, with the exception of property crime that had low values for Zones 2, 8, 12, 15, and 19.

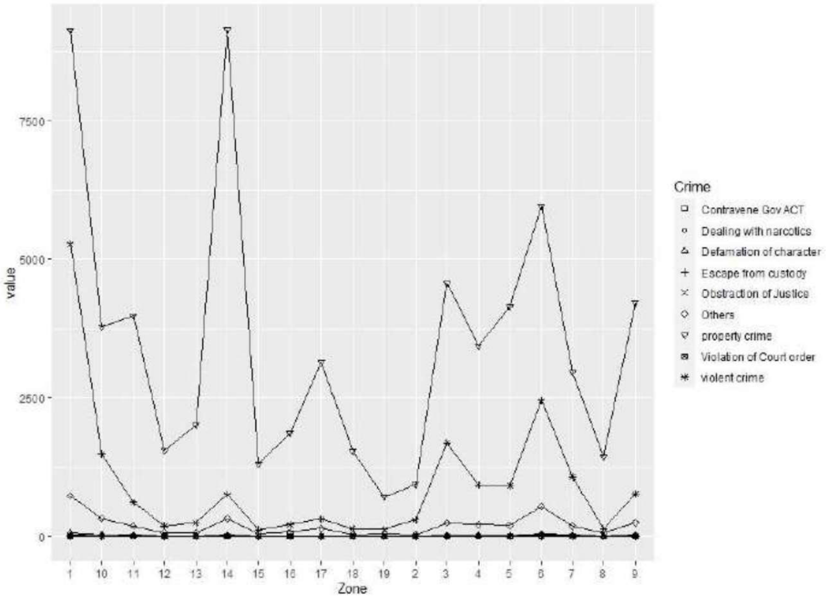


Figure 1: Crime trends across each Zone for the combined period 2011-2016

Figure 2 on the other hand, shows the overall crime trends for each hour of the day for the combined period 2011-2016. It can be observed that crime mostly happened from 09 AM to midnight, most probably due the fact that the city gets busy during these hours and criminals take advantage of this to commit crime. It can be noted that the frequencies of violent crimes and property crime is at its highest during this time period when most people are at work. The graph also showed that from midnight to early hour in the morning (01:00- 09:00) crime were very low. This may be due to police patrol and shutting down shebeens and other public gathering where crime is likely to occur and also everyone else being home with families.

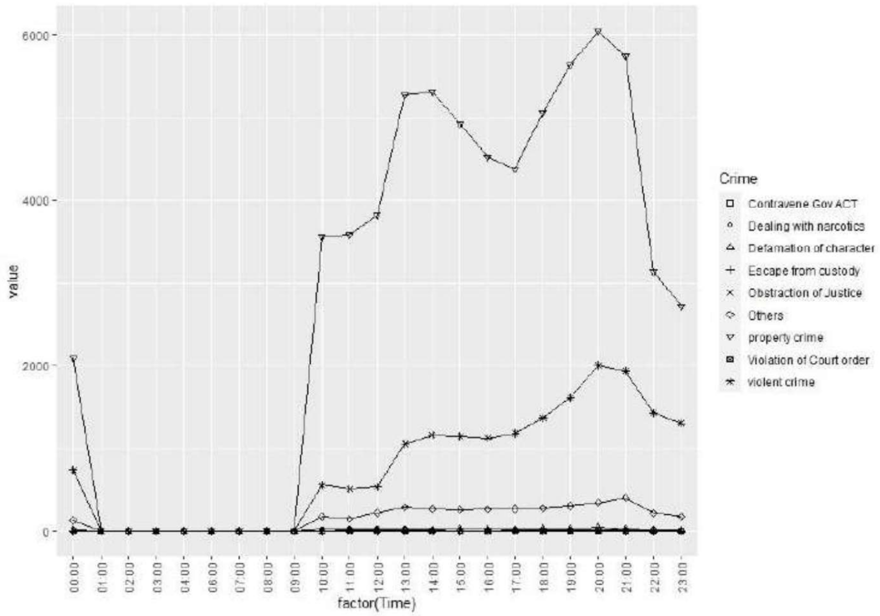


Figure 2: Overall crime trends for each hour of the day for the combined period 2011-2016

Finally the figure 3 shows the snapshots of this crime behaviours over months for the period 2011-2016. From this figure it can be seen that both property crime and violent crime have been fluctuating across the months but remained high in December, compared to the remaining types of crime reported in 2011-2016. During this vacation period, a lot of houses stay unoccupied hence a perfect opportunity for perpetrator. A lot of social events also take place during this time period, making it easier for things to be stolen from vehicles while individuals are at gatherings, thus leading to high increase of violent and property crime. The 'others' crime category remained flat from January to April and then peaked up from there and remained a bit high until December.

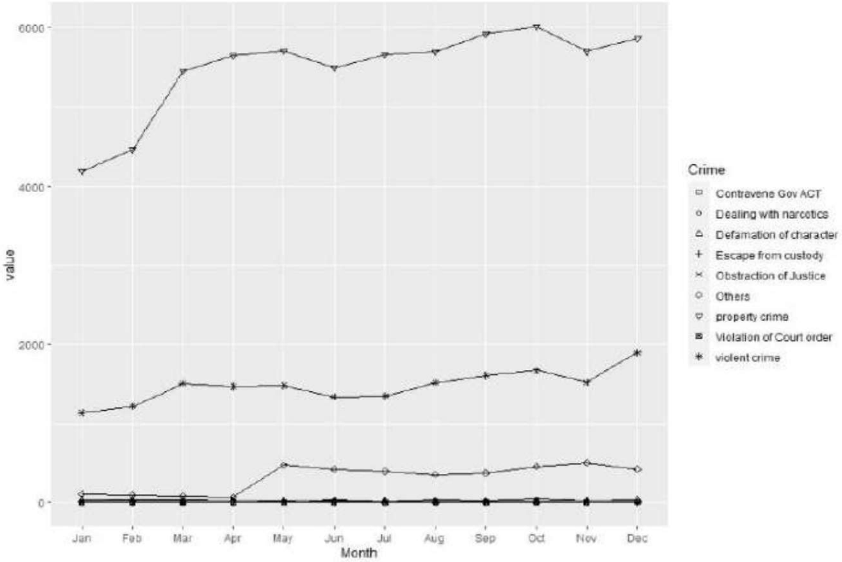
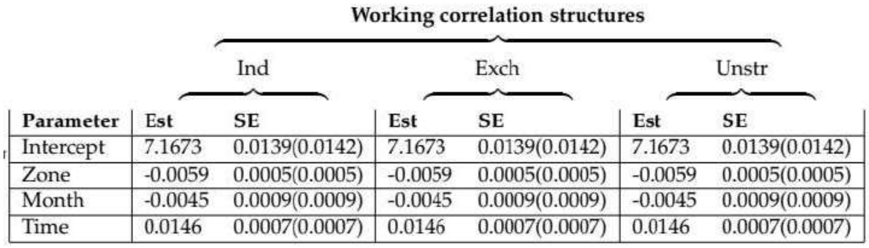


Figure 3: Snapshots of same crime behaviours over months for the period 2011-2016

Table 1 shows the output summaries of the fitted model for different correlation structure. The presents output from model fit with intercept and slope parameter only for crime data. The same results were produced for all the three (Exchangeable, Independent, and Unstructured) working correlation structure. This occurred, perhaps, due to the redundant in the data. Consequently, the researcher was unable to select the best model for the data since the QIC value were the same as well as the standard errors.

Table 1: Output summaries of the fitted model for different correlation structure



	Working correlation structures					
	Ind		Exch		Unstr	
Parameter	Est	SE	Est	SE	Est	SE
Intercept	7.1673	0.0139(0.0142)	7.1673	0.0139(0.0142)	7.1673	0.0139(0.0142)
Zone	-0.0059	0.0005(0.0005)	-0.0059	0.0005(0.0005)	-0.0059	0.0005(0.0005)
Month	-0.0045	0.0009(0.0009)	-0.0045	0.0009(0.0009)	-0.0045	0.0009(0.0009)
Time	0.0146	0.0007(0.0007)	0.0146	0.0007(0.0007)	0.0146	0.0007(0.0007)

In theory the best working correlation structure is the unstructured one but the challenge with that is that if the data set is too big, then it estimates as many parameters as we can accommodate in any report. hence the exchangeable correlation structure could be adopted for this research. A different data set on medicare longitudinal data was used to test the problem with results obtain in this report and confirm that the theory is working fine for the medicare data, hence the conclusion that out crime data need serious revisitation to figure out the error.

4. Discussion

This study has demonstrated the analysis of longitudinal crime data in Windhoek municipality using Generalised Estimating Equations model fit. The study also looked at some factors that lead to crime. It was found that there are some similarities in the literature and the study results.

During the study period, the number of property and violent crime was high in Windhoek; these could be due to unemployment factor. This result was similar to the findings of Lilungwe (2020) on the relationship between youth and unemployment, which showed that there was a direct relationship between youth and unemployment. The number of property crime remained higher.

The number of property and violent crime were high in the month of January to April, while the remaining crimes were low. During December, property and violent crime increased, because residents get to celebrate social event such as Christmas leaving their houses unattended while some get drunk and start to have conflicts.

It was found that there was a correlation between variables (crime, Months, Zone, and Time), it is important to note that correlation does not mean interconnection (cause). This would not mean that there more people in a certain location during the day the more crime being committed, but also this could be the possibility that there is lack of police patrolling in the locations at that time. Furthermore, not all crime incidents that took place in various locations were reported and this lead to bias. It was found that for any change in Zone, crime decrease while for every hour increase in time, crime increased.

References

M. Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika, 82(2):407–410, 1995.

J. Cui and G. Qian. Selection of working correlation structure and best model in gee analyses of longitudinal data. Communications in Statistics Simulation and Computation, 36 (5):987–996, 2007.

Neema I and Böhning D. Monitoring murder crime in Namibia using Bayesian space-time models. Journal of Probability and Statistics, 2012, 2012.

Ignatius N Kathena and Johannes PS Sheefeni. The relationship between eco- nomic growth and crime rates in Namibia. European Journal of Basic and Applied Sciences Vol, 4(1), 2017.

Mayumbelo Lilungwe. Youth unemployment and property crime in Namibia: A case of Samora Machel constituency. PhD thesis, 2020.

R. B. Millar. Maximum likelihood estimation and inference: with examples in R, SAS and ADMB, volume 111. John Wiley & Sons, 2011.

S. Oh, K. C. Carriere, and T. Park. Model diagnostic plots for repeated measures data using the generalized estimating equations approach. Computational Statistics & Data Analysis, 53(1):222–232, 2008.

W. Pan and J. E. Connett. Selecting the working correlation structure in generalized estimating equations with application to the lung health study. Statistica Sinica, 12(2): 475–490, 2002.

Robert E Ployhart and Robert J Vandenberg. Longitudinal research: The theory, design, and analysis of change. Journal of management, 36(1):94–120, 2010.

Chor Foon Tang. An exploration of dynamic relationship between tourist arrivals, inflation, unemployment and crime rates in Malaysia. International Journal of Social Eco- nomics, 2011.

Y. G. Wang and V. Carey. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. Biometrika, 90(1):29–41, 2003.

R. E. Weiss. Modeling Longitudinal Data: With 72 Figures. Springer Science & Business Media, 2005.

S. L. Zeger and K. Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. Biometrics, pages 121–130, 1986.

C. J. W. Zorn. Generalized estimating equation models for correlated data: A review with applications. American Journal of Political Science, pages 470–490, 2001.