MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Forecasting homicides, rapes and counterfeiting currency: A case study in Sri Lanka

Abstract

Crimes have been disturbing threats to all the Sri Lankans all over the country. Finding the main variables associated with crimes are very vital for policymakers. Our main goal in this study is to forecast of homicides, rapes and counterfeiting currency from 2013 to 2020 using auto-regressive conditional Poisson (ACP) and auto-regressive integrated moving average (ARIMA) models. All the predictions are made assuming that the prevailing conditions in the country affecting crime rates remain unchanged during the period. Moreover, multiple linear regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis were used to identify the key variables associated with crimes. Profiling of districts as safe or unsafe was performed based on the overall total crime rate of Sri Lanka which is to compare with individual district's crime rates. Data were collected from the Department of Police and Department of Census and Statistics, Sri Lanka. It is observed that there are 14 safe and 11 unsafe districts in Sri Lanka. Moreover, it is found that the total migrant population and percentage of urban population is positively correlated with total crime. Besides, total migrant population, unemployment rate, mean household income and percentage of the urban population are significant variables for total crimes, and total migrant population, Gini index, mean household income and percentage of the urban population are significant variables for homicides. Random K-nearest neighbour (RKNN) algorithm classified districts as safe and unsafe with 84% of prediction accuracy.

Keywords: autoregressive conditional poisson model, autoregressive integrated moving average, crime analysis, gini index, random k-nearest neighbor algorithm.

Volume 9 Issue 6 - 2020

Chathura B. Wickrama,[1] Lakshika S. Nawarathna[2]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka

[2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka

Correspondence: Lakshika S. Nawarathna, Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka, Tel +940767552223, Email lakshikas@pdn.ac.lk

Received: November 03, 2020 | **Published:** December 31, 2020

Abbreviations: ACP, auto-regressive conditional Poisson; ARIMA, auto-regressive integrated moving average; LASSO, least absolute shrinkage and selection operator; RKNN, random K-nearest neighbour; CID, criminal investigating department

Introduction

Crime is one of the issues from which countries are suffered from the existence of mankind. These crimes have been disturbing threats to personalities, properties and lawful authorities of mankind. Reviews of the literature on this topic can be found in Louis et al.[1] Crime began in the primitive days as a simple and less organized problem. Nowadays, due to the technological advancements, crimes are well organized and difficult to investigate and hence the situation is more complex.

The wave of crime is a key social problem in Sri Lanka and caused by the rising population and advancement of modern technology than the earlier. Crimes such as homicides, rapes, child abuses, hitting, thefts, and illegal money printings are still threatening the Sri Lankan society. Due to this condition, a vast amount of harms have been occurred to the people all over the country. Threats, suspicions, revenging, fear of the people, suicides are the major calamities resulting from the crimes.[2] Crimes continue to attract the attention of all stakeholders, including the government and political leaders, the management and leadership of the Sri Lanka Police, individual citizens as well as the international community. Criminal Investigating Department (CID), criminal justice and law enforcement agencies exist to guarantee personal safety and security of property in Sri Lanka. The level of effectiveness of these agencies can be improved by information gained by crime analysis.

Crimes can be controlled by introducing new punishments such as the death penalty and finding the key factors affecting overall crimes and adjust those factors for positive changes by policy altering.[3] It has been found that when an opportunity for crime is blocked, an offender has several other types of displacement. Therefore, this study facilitates for policy altering by identification of criminal factors. In order to find those factors, multivariate statistical tools can be applied and proved to be effective in many criminological explanations.[4]

Identification of trends in crimes is very important for policy makers to change their policies, for that, we look for possible trends of homicides, rapes and counterfeiting currency incidents. This study can answer the question of what factors significantly affect the total crimes and homicides by developing a model. Inorder to minimize crimes, it is important to know which factors mainly affect the crimes to determine what type of policy changes can be made. With the developed model, we predict the crimes for each district using significant factors. Moreover, associations between different crime types which can be used to lower the crimes will be assessed. Using the Random K-nearest neighbor (RKNN) algorithm, we profile districts of Sri Lanka as safe or unsafe without using the actual number of crimes committed in Sri Lankan districts. Furthermore, this article will provide effective guidance to help individuals better understanding of the factors associated with crimes and thus will be helpful in crime prevention.

The rest of this article is organized as follows. Section 2 presents the proposed methodology. In Section 3, we rank and classify the districts of Sri Lanka based on total crimes, land area and overall crime rate. Besides time series analysis is used to forecast crimes. Moreover, we propose a model for predicting total crimes and homicides. Further, the classification of crimes is performed using variables associated with the safeness. Finally, Section 4 concludes with a discussion.

Methodology

In this study, the required data are collected from the Department of Police and the Department of Census and Population, Sri Lanka. All the statistical analysis was done by using R statistical software version 3.5.1.[5]

The crime rate varies across individual districts and could be more or less than the overall crime rate of Sri Lanka. Therefore, districts are ranked and categorized as safe and unsafe districts. If a crime rate of a district is below the overall crime rate, it is considered as a safe district and if crime rate of a district is more than the overall crime rate, it is considered as an unsafe district. The Crime rate is calculated based on population and land area of a district.

$$\text{Crime rate per } 100{,}000 \text{ population} = \frac{\text{Total crimes in a district}}{\text{Total population in that district}} * 100{,}000 \quad (2.1)$$

and

$$\text{Crime rate per 1 km}^2 = \frac{\text{Total crimes in a district}}{\text{Total area of a district in 1 km}^2}. \quad (2.2)$$

Data from different crime types in 2012 were analyzed for each district. Further, annual total crime data ranges from 1973 to 2014 are used for time series analysis to predict homicides, rapes and counterfeiting currency. In ARIMA technique, the future value of a variable is a linear combination of past values and past errors, expressed as follows.

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \quad (2.3)$$

where Y_t is the actual value, ε_t is the random error at time t, φ_i and θ_j are the coefficients, p and q are integers that are often referred to as autoregressive and moving average, respectively. Optimal values of p, q and difference term (d) are determined using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Given a time series of counts, $Y_1,...,Y_t$ where Y_{t-1} denote the information on the time series up to time $t-1$, then for the ACP(1,1) model, the counts, conditioned on past observations, are modeled as

$$Y_t \mid Y_{t-1} \sim Poisson(\mu_t) \quad (2.4)$$

with an autoregressive conditional mean given as

$$\mu_t = \omega + \alpha Y_{t-1} + \beta \mu_{t-1} \quad (2.5)$$

for $\omega > 0$ and $\alpha, \beta \geq 0$. This can be extended to include additional lags.[6] Provided the ACP (1,1) is stationary and has an unconditional mean and variance given by

$$E[y_t] = \mu = \frac{\omega}{(1-(\alpha+\beta))} \quad (2.6)$$

and

$$Var[y_t] = \frac{\mu(1-(\alpha+\beta)^2+\alpha^2))}{(1-(\alpha+\beta)^2} \quad (2.7)$$

Two Ordinary Least Squares (OLS) models are built as OLS total crime model and OLS homicide model. Total crime and homicide are dependent variables in OLS total crime model and OLS homicide model respectively. With OLS regression and LASSO regression analysis, this study can answer the question of what factors affect the total crimes and homicides and predict the future crimes for each district. A statistical model is created to predict total crimes for each district. All the variables utilized for the analysis are listed in the Table 1. Let

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_k X_k + \varepsilon \quad (2.8)$$

where X_k^0 s are the final selected variable using stepwise variable selection method, ε is the error term, β_0 is the intercept and $\beta^0 s$ are coefficients for selected variables. In fitting a multiple regression model, it is much more convenient to express the mathematical operations using matrix notation. Suppose that there are k independent variables and n observations.

Table 1 Details of variables

Variable No	Variable	Variable Description
1	Y	Total crimes
2	X1	Percentage of people between 15 and 24
3	X2	Total migrant population
4	X3	Unemployment rate
5	X4	Gini coefficient which describe income inequality
6	X5	No schooling percentage
7	X6	Mean household income
8	X7	Population density (People per square kilometer)
9	X8	Percentage of urban population
10	X9	Percentage of people below the poverty line
11	X10	Percentage of people divorced and separated
12	X11	Percentage difference between male and female

This model is a system of n equations that can be expressed in matrix notation as,

$$Y = X\beta + \varepsilon \quad (2.9)$$

where $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_{25} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & . & . & . & x_{1k} \\ 1 & x_{12} & . & . & . & x_{2k} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 1 & x_{251} & . & . & . & x_{25k} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_{25} \end{bmatrix}$

We wish to find the vector of least square estimators (L), $\hat{\beta}$ minimizes the least squares estimator where $\hat{\beta}$ is the solution for β in the equations.

$$\frac{\partial L}{\partial \beta} = 0 \text{ and } \hat{\beta} = \left(X^T X\right)^{-1} X^T Y \quad (2.10)$$

LASSO technique is useful as it improves the quality of predictions by shrinking regression coefficients, compared to predictions based on a model fitted via unpenalized maximum likelihood.[7] Given a set of input measurements $x_1, x_2, ..., x_p$ and an outcome measurement y, the LASSO fits a linear model.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

We minimize $\sum_{j=1}^{p} |\beta_j| \leq \lambda$ subject to $\sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$

Citation: Wickrama CB, Nawarathna LS. Forecasting homicides, rapes and counterfeiting currency: A case study in Sri Lanka. *Biom Biostat Int J.* 2020;9(6):209–215. DOI: 10.15406/bbij.2020.09.00322

where the bound λ is a tuning parameter. The sum is taken over observations in the data set. When λ is large enough, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on $x_1, x_2, ..., x_p$. However when for smaller values of $\lambda(\geq 0)$ the solutions are shrunken versions of the least squares estimates. Often, some of the coefficients b_j's are zero. Choosing λ is like choosing the number of predictors to use in a regression model, and cross-validation is used for estimating the best value for λ.[7]

Feature selection is performed in order to find the importance of the variables and RKNN algorithm is run in order to classify the districts. The random Forest package[8] and rknn package in R[9] are used in this purpose. The Random Forest algorithm is used for variable selection. The relative rank (i.e. depth) of a feature used as a decision node in a tree is used to assess the relative importance of that feature with respect to the predictability of the target variable. Features used at the top of the tree are used to contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples is used as an estimate of the relative importance of the features. By averaging those expected activity rates over several randomized trees, one can reduce the variance of such an estimate and use it for feature selection.[10] After selecting the best variables, for model building, RKNN algorithm is used and RKNN constitutes of an ensemble of base *k*-nearest neighbor models, each built from a random subset of the input variables.[11] Random KNN method was introduced using some techniques used in random forest method and is similar in the method of random subspace selection used for decision forests. Random KNN uses KNN as base classifiers, with no hierarchical structure involved. Compared with decision trees, KNN is simple to implement and is stable.[12] Thus, Random KNN is stabilized with a small number of base KNN's and hence only a small number of important variables will be needed. This implies that the final model with Random KNN will be simpler than that with random forest or decision forests. Specifically, a collection of r different KNN classifiers will be generated. Each one takes a random subset of the input variables. Since KNN is stable, bootstrapping is not necessary for KNN. Each KNN classifier classifies a test point by its majority, or weighted majority class, of its *k*-nearest neighbors. The final classification in each case is determined by majority voting of r, KNN classifications. This can be viewed as a sort of voting by a majority of a majority.

Let $F = \{f_1, f_2, ..., f_p\}$ be the p input features, and X be the n original input data vectors of length p, (an n x p matrix). For a given integer $m < p$, denote $F^m = f_{j1}, f_{j2}, ..., f_{jm} \mid f_{jl} \in F, 1 \leq l \leq m$ a random subset drawn from F with equi-probability. Similarly, let X^m be the data vectors in the subspace defined by F^m, i.e., an n m matrix. Then a $\text{KNN}^{(m)}$ classifier is constructed by applying the basic KNN algorithm to the random collection of features in X^m. A collection of r such base classifiers is then combined to build the final random KNN classifier.

Results and Discussion

Figure 1 illustrates the pie chart of different crime types in percentages. This pie chart shows that the majority of crimes in 2012 is related to property crimes in which home break and theft represents 49% and robbery represents 19%. Hurt by knife is recorded as the highest number of crimes against persons which is 8% while counts of rapes account 6 %.

Further, box plots in Figure 2 are used to study the distributions of different crime rates per 100,000 population. Districts with rates of a low number of home breaks and thefts are more condensed than the districts with rates of a higher number of home breaks and thefts. Moreover, it can be observed that Colombo and Gampaha districts are outliers for many crime types. Further, Gampaha and Colombo districts are outliers for homicide and drug-related crimes respectively. Besides Gampaha district is an outlier for abduction/ kidnapping, home break and thefts and robbery. Child abuses are prevalent in Mannar and Pollonnaruwa districts.

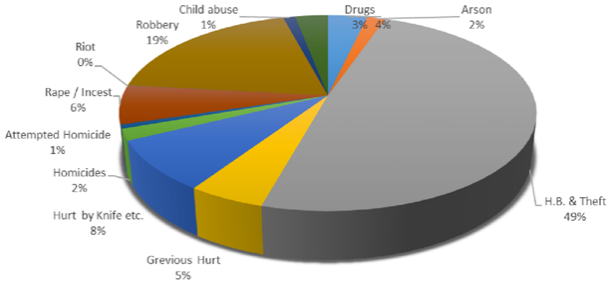

Figure 1 Percentages of crime types.

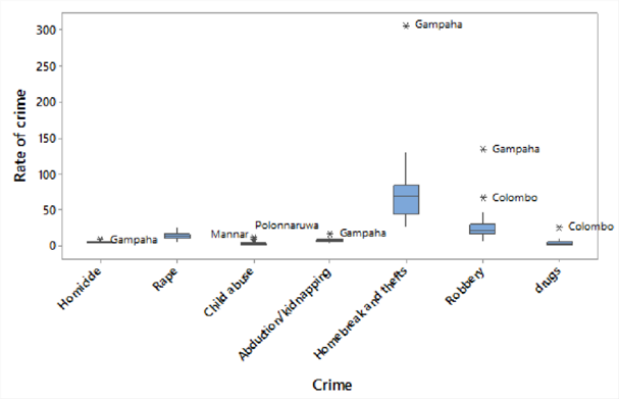

Figure 2 Box-plots of different crime types.

District ranks based on total crimes and homicides

Table 2 shows the ranks of districts based on total crimes per 100,000 population (i.e. population criteria) and per 1 square Kilometer (i.e. area criteria) basis. Total crimes of each district were used for this analysis.

Table 2 Ranks of districts based on total crimes

Rank	Population criteria (Per 100,000 people)		Area criteria (Per 1 km ^{2})	
	District	Rate	Dis trict	Rate
1	Colombo	47.40	Colombo	16.20
2	Gampaha	39.36	Gampaha	6.74
3	Killinochchi	39.03	Kalutara	2.04
4	Kegalle	31.57	Kegalle	1.57
5	Anuradhapura	31.08	Galle	1.55
6	Vavunia	30.22	Kandy	1.55
7	Polonnaruwa	26.75	Matara	1.32
8	Hambantota	26.63	Jaffna	1.04
9	Mannar	26.54	Rathnapura	0.83
10	Kalutara	26.41	Kurunegala	0.72

Citation: Wickrama CB, Nawarathna LS. Forecasting homicides, rapes and counterfeiting currency: A case study in Sri Lanka. *Biom Biostat Int J.* 2020;9(6):209–215. DOI: 10.15406/bbij.2020.09.00322

Table continued...

Rank	Population criteria (Per 100,000 people)		Area criteria (Per 1 km ^{2})	
	District	Rate	Dis trict	Rate
11	Rathnapura	24.89	Hambantota	0.64
12	Galle	23.71	Puttalam	0.52
13	Kandy	21.72	Matale	0.49
14	Monaragala	21.24	Nuwara Eliya	0.49
15	Matara	20.65	Badulla	0.49
16	Batticaloa	20.65	Batticaloa	0.42
17	Kurunegala	20.57	Anuradhapura	0.40
18	Trincomale	20.44	Killinochchi	0.37
19	Matale	19.96	Polonnaruwa	0.35
20	Puttalam	19.77	Vavunia	0.28
21	Mullativu	17.50	Ampara	0.26
22	Badulla	16.91	Trincomale	0.20
23	Ampara	16.83	Monaragala	0.17
24	Jaffna	16.59	Mannar	0.14
25	Nuwara Eliya	11.79	Mullativu	0.07

According to the results, Colombo and Gampaha have the highest crime rates based on both population and area criteria and have been ranked in first and second positions respectively, whereas Nuwara Eliya district records the lowest based on the population criteria (per 100,000 people). Based on the area criteria, Mullativu district records the lowest. It is found that a resident in Nuwara Eliya district have experienced nearly 4 times fewer crimes than a resident in Colombo district based on the population criteria and a resident in Colombo district could see 231.4 times of more crimes than a resident in Mullativu district based on area criteria. A heat map of a total crimes based on area criteria is indicated in Figure 3. It shows that crimes are more prevalent in Western Province of Sri Lanka. It also shows that Kegalle, Galle, Kandy, Matara and Jaffna districts have significant number of total crimes per area.

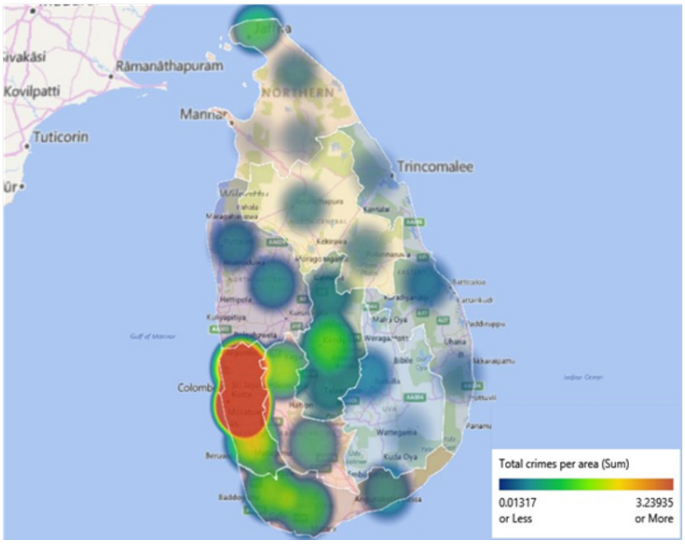

Figure 3 Heat map of total crimes based on area criteria.

Moreover, most of the crimes can be observed in Colombo and Gampaha districts and spread over to the down-south in decreasing magnitude. Further, crimes are decreasing in rate from Colombo, Gampaha, and Kegalle to Kandy. When the distance from Colombo to other adjacent districts increases, crime rates tend to be lower.

Table 3 shows the ranking of districts based on homicides per 100,000 populations and per 1 km^2 basis. Homicides of each district were used for this analysis.

Table 3 Ranks of districts based on total homicides

Rank	Population criteria (Per 100,000 people)		Area criteria (Per 1 km ^{2})	
	District	Rate	Dis trict	Rate
1	Vavunia	5.83	Colombo	0.0917
2	Monaragala	5.35	Gampaha	0.0567
3	Galle	4.72	Galle	0.0309
4	Rathnapura	4.25	Jaffna	0.0248
5	Mannar	4.04	Matara	0.0228
6	Jaffna	3.94	Kalutara	0.0222
7	Kurunegala	3.60	Kegalle	0.0161
8	Matara	3.58	Rathnapura	0.0142
9	Killinochchi	3.54	Kurunegala	0.0125
10	Hambantota	3.52	Kandy	0.0104
11	Trincomale	3.44	Nuwara Eliya	0.0100
12	Badulla	3.43	Badulla	0.0099
13	Gampaha	3.31	Hambantota	0.0084
14	Mullativu	3.30	Puttalam	0.0076
15	Kegalle	3.11	Matale	0.0072
16	Matale	2.90	Vavunia	0.0054
17	Puttalam	2.89	Trincomale	0.0051
18	Kalutara	2.87	Monaragala	0.0044
19	Anuradhapura	2.80	Batticaloa	0.0042
20	Colombo	2.68	Ampara	0.0038
21	Ampara	2.47	Anuradhapura	0.0036
22	Nuwara Eliya	2.41	Killinochchi	0.0033
23	Polonnaruwa	2.23	Polonnaruwa	0.0029
24	Batticaloa	2.09	Mannar	0.0021
25	Kandy	1.46	Mullativu	0.0012

Vavunia and Monaragala districts have the highest homicide rates per 100,000 people and have been ranked in first and second positions respectively. Based on the population criteria (per 100,000 people), Kandy district records the lowest. According to the area criteria, Mullativu district records the lowest. It is found that a resident in Kandy district has 4 times less chance of being killed compared to a resident in a Vavunia district based on the population criteria. In one square kilometer, a resident in Colombo district could see 76.4 more homicides than a resident in Mullativu district. Figure 4 shows the 3-D representation of total crimes and homicides.

Table 4 describes the status of districts as safe or unsafe based on country's total crime rate in which safe districts have its crime rate below the overall total crime rate and unsafe districts have its crime rate higher than the overall total crime rate.

According to the classification, there are 14 safe and 11 unsafe districts in Sri Lanka. It should be noted that the Central Province is a

safe province as its all districts (Kandy, Matale, and Nuwara Eliya) are safe and also Western province is an unsafe province as its crime rates of all representing districts are much higher than the overall crime rate.

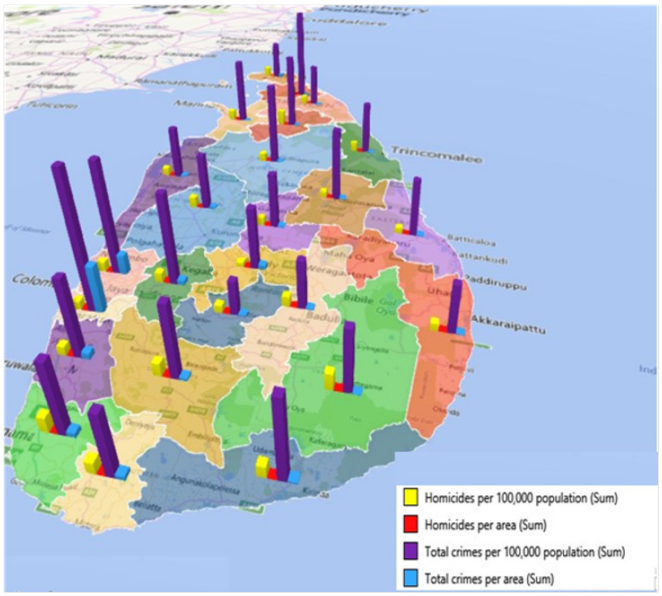

Figure 4 3-D representation of total crimes and homicides.

Table 4 Classification of districts as safe and unsafe

No	Safe districts	Unsafe districts
1	Galle	Colombo
2	Kandy	Gampaha
3	Monaragala	Killinochchi
4	Matara	Kegalle
5	Batticaloa	Anuradhapura
6	Kurunegala	Vavunia
7	Trincomale	Polonnaruwa
8	Matale	Hambantota
9	Puttalam	Mannar
10	Mullativu	Kalutara
11	Badulla	Rathnapura
12	Ampara	
13	Jaffna	
14	Nuwara Eliya	

Time series analysis for crime data

Time series analysis of homicides, rapes and counterfeiting currency was performed separately to find any underlying model. Time series analysis of homicides was done by developing ARIMA and ACP models using data from 1973 to 2012. Two outliers of homicide data were detected in 1988 and 1989 and those data points were cleaned and replaced by the linear interpolation. The Linear interpolation concerns the act of predicting or estimating extreme values based on their relationship to one or more other variables. Besides, it concerns estimation within ranges already measured. ACP models of homicides, rapes and counterfeiting currency were selected over ARIMA models as they had low AIC and BIC values. Selected ACP models for homicides, rapes and counterfeiting currency are shown in Table 5. All the coefficients of models are significant at 5% significant level. Forecasts were made using selected ACP models. Figure 5 shows the forecast of homicides, rapes and counterfeiting currency for 2013-2020. It seems that homicide counts are increasing from 2015 to 2020. The trend of increasing rape counts continues until 2020. Counterfeiting currency incidents will be stable until 2020. But a constant forecasts for counterfeiting was observed for 2013 -2015.

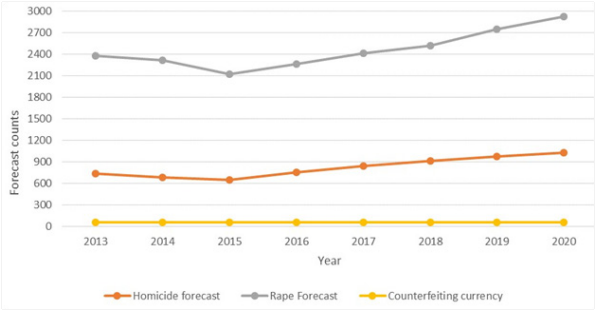

Figure 5 Forecast of homicides, rapes and counterfeiting currency for 2013-2020.

Comparison of actual and forecast values of homicides, rapes and counterfeiting currency was illustrated in Table 6. It is observed that homicides, rapes and counterfeiting currency actual values are approximately same as the predicted values.

Regression analysis for total crimes and homicides

In the regression analysis for total crimes, a model with Total migrant population, Unemployment rate, Mean Household income, Percentage of urban population and Percentage of people below poverty line are significant at a 5% significance level and the following model was selected as the best model.

$$Total\ Crimes = 6050 + 0.01458 * Total\ migrant\ population + 329.9 * Unemployment\ rate - 507.8 * Mean\ Household\ income + 40.81 * Percentage\ of\ urban\ population + 36.42 * Percentage\ of\ people\ below\ poverty\ line$$

Table 5 Coefficient estimates of ACP models for homicides, rapes and counterfeiting currency

Model	Coefficient	Estimate	Standard Error	t-value	p-value
Homicides	ω	195.80	25.24	7.7547	<0.0001
	α	0.9610	0.0233	41.1857	<0.0001
	β	-0.1064	0.0232	-4.5927	<0.0001
Rapes	ω	5.66	0.75	7.54	<0.0001
	α	1.1237	0.024	45.9972	<0.0001
	β	-0.08	0.0238	-2.5497	0.0151

Citation: Wickrama CB, Nawarathna LS. Forecasting homicides, rapes and counterfeiting currency: A case study in Sri Lanka. *Biom Biostat Int J.* 2020;9(6):209–215. DOI: 10.15406/bbij.2020.09.00322

Table continued...

Model	Coefficient	Estimate	Standard Error	t-value	p-value
Counterfeitin g currency	ω	41.86	4.34	9.63	<0.0001
	α	0.34	0.037	9.16	<0.0001
	β	-0.077	0.007	-1.0588	<0.0001

Table 6 Actual and forecasted values of homicides, rapes and counterfeiting currency

Crime	Year	Actual value	Forecast	Difference value
Homicides	2013	586	732	146
	2014	548	681	133
Rapes	2013	2181	2372	191
	2014	2008	2114	106
	2015	2033	2125	92
Counterfeiting currency	2013	59	53	6
	2014	52	58	6

Model validation was done, comparing actual values with the predicted values for the best model and results are shown in Figure 6. The predicted crimes go fairly well with the actual crimes and display almost the same pattern. This reveals the estimated model adequately utilizes the data. Total crime model has higher adjusted R-squared value of 0.9712. This means that the independent variables included in the total crime model can explain 97.12% of variation around the mean of total crimes.

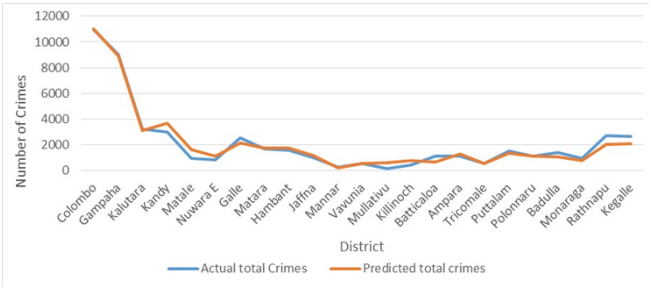

Figure 6 Actual and predicted total crimes.

Moreover, a regression analysis was conducted to find the best model for homicides. Total migrant population, Gini coefficient and percentage of the urban population are significant at a 5% significance level. The resulting model is as follows.

The actual and estimated value of crimes were compared to validate the model and the resulting plot is shown in Figure 7. The predicted crimes and actual crimes are overlapping and shows almost the same pattern. This reveals the estimated model is adequate to utilize the sample. Total homicide model has an adjusted R-squared value of 0.83. According to the model coefficients, total migrant population, Gini-coefficient, mean household income and percentage of the urban population are significant variables. Gini-index describes the income inequality of the society. This variable found to have significant at the 5% significance level. Gini coefficient is a very large factor in crime rate and finds it to have a positive coefficient. This suggests for policy makers that government should try to reduce the income inequality. They can do this by making the income distribution more even which will reduce the amount of poverty and in turn reduce the amount of crime in their districts. City planners should be concerned about their town planning, as crowded streets and sidewalks could be effective deterrents to criminal behavior. Studies done by Schuessler and Galle et al.[13,14] found positively correlated relationships between crime and population density and matched with our findings.

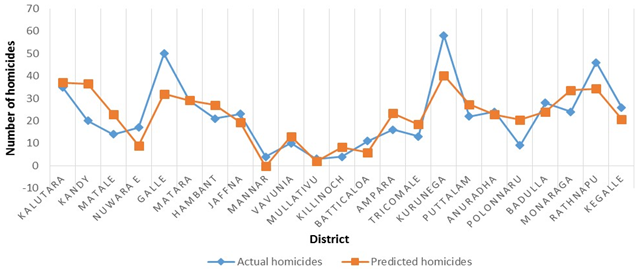

Figure 7 Actual and predicted values plot for homicides.

Forecasts of Kurunegala and Anuradhapura districts for total crimes and Colombo and Gampaha districts for homicides indicate in Table 7. This shows that all the predicted values are in the 95% prediction interval range.

Table 7 Forecasting of total crimes and homicides with OLS models

District	OLS model	Actual	Predicted	Confidence level		Difference
				Lower	Upper	
Kurunegala	Total crime	3314	3422.65	1809.47	3843.87	108.65
Anuradhapura		2662	2412.71	745.76	2765.65	249.29
Colombo	Homicide	62	97.16 81.21	31.96 24.27	167.98	35.16 5.21
Gampaha		76			134.67	

Moreover, assumptions of homoscedasticity, auto correlation, multicolinearity, normality and linear relationship are not violated in homicide and total crime OLS models. Two separate model were fitted for predicting total crimes and homicides using the Lasso regression technique as follows.

$$Total\ crimes = -181.69 + 0.00685 * Total\ migrant\ population + 0.942 * population\ density - 45.32 * Mean\ household\ income$$

$$Homicides = -2.314 + 4.79e-05 * Total\ migrant\ population + 0.0334 * Gini\ coefficient + 0.0246 * Population\ density$$

Citation: Wickrama CB, Nawarathna LS. Forecasting homicides, rapes and counterfeiting currency: A case study in Sri Lanka. *Biom Biostat Int J.* 2020;9(6):209–215. DOI: 10.15406/bbij.2020.09.00322

In comparison to the OLS homicide model, the percentage of urban population is not significant and population density is significant in LASSO homicide model.

Variable importance is done by measuring the total decrease in node impurities and the results are shown in Table 8. No schooling percentage, percentage of people below the poverty line and population density, mean household income and Gini coefficient are the most important variables in determining the safeness of districts and those variables are used to run the RKNN algorithm. Only Badulla district is wrongly categorized with the error rate is 16.6. If all 25 districts are categorized using the above selected variables, four districts as safeness results could be erroneous in general. Therefore, if a comparison is made with OLS regression and LASSO technique, Total migrant population is a common variables in OLS regression and LASSO regression for both total crimes and homicides. Population density is a key factor for total crimes in OLS and LASSO regressions, and in Safeness. Gini Coefficient is common in OLS homicide model and LASSO homicide model.

Table 8 Comparison of actual safeness and predicted safeness

District	Actual	Predicted
Anuradhapura	Unsafe	Unsafe
Polonnaruwa	Unsafe	Unsafe
Badulla	Safe	Unsafe
Monaragala	Safe	Safe
Rathnapura	Unsafe	Unsafe
Kegalle	Unsafe	Unsafe

Conclusion

Colombo district has the highest total crime rate based on per 100,000 population and per 1km^2. Vavunia district has the highest homicide rate per 100,000 population and Colombo district has the highest homicide rate per 1km^2. It is evident that all the districts in Western Province are unsafe in relation to other districts. Nuwara Eliya district has the lowest total crime rate per 100,000 population and Mullativu district has the lowest total crime rate per 1km^2. Kandy district has the lowest homicide rate per 100,000 population and Mullativu district has the lowest homicide rate per 1km^2. There are 14 safe and 11 unsafe districts in Sri Lanka. All the districts in Central Province are safe with other districts. Spearman correlation analysis suggests that minimizing of one type of crime causes to reduce another type of crime by their positive correlation. Therefore, at first, policy makers should try to reduce crimes which are easily controllable and less costly. Also, they should take actions to reduce migrations as this leads to more crimes and concentrate their efforts on stopping crimes in highly crowded districts.

Acknowledgments

Authors wish to thank the Department of Police and Department of Census & Statistics, Sri Lanka for providing the data sets used in this paper.

References

1. Louis S, Cookie WS, Louis AZ, et al. Human Response to Social Problems. The Dorsey Press, Homewood, IL, first edition, 1981.
2. Jayathunga NS. A sociological study of the homicide in sri lanka: A case study in rathnapura secretariat division. *Sabaragamuwa University Journal*. 2010;9(1):45–55.
3. Donohue JJ, Wolfers J. Uses and abuses of empirical evidence in the death penalty debate. Technical report, National Bureau of Economic Research, 2006.
4. Idele SI, Kpedekpo GMK, Arya PL. Social and economic statistics for Africa. 1987.
5. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
6. Heinen A. Modelling time series count data: An autoregressive conditional poisson model. MPRA Paper 8113, University Library of Munich, Germany, 2003.
7. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B (Methodological). 1996;58(1):267–288.
8. Liaw A, Wiener M. Classification and regression by randomforest. *R news*. 2002;2(3):18–22.
9. Shengqiao L. *rknn: Random KNN Classification and Regression*, 2015.
10. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
11. Shengqiao Li, Harner EJ, Adjeroh DA. Random knn feature selection–a fast and stable alternative to random forests. *BMC Bioinformatics*. 2011;12(1):450.
12. Dietterich TG, Lathrop RH, Lozano–Perez T. Solving the multiple instance problem with axis–parallel rectangles. *Artificial intelligence*. 1997;89(1):31–71.
13. Schuessler K. Components of variation in city crime rates. *Social Problems*. 1962;9(4):314–323.
14. Galle OR, Gove WR, McPherson JM. Population density and pathology: what are the relations for man? *Science*. 1972;176(4030):23–30.