Implementation of Forecasting and Classification Models to Predict Crimes in Chicago

Nagendra B.V.
RACE,
REVA University
Bengaluru, India
nagendrabv.ba02@reva .edu.in

Krishna Goswami
RACE,
REVA University
Bengaluru, India
krishnag.ba02@reva.edu.in

Jay B.Simha[1,2]
[1]Abiba Systems
[2]RACE,
REVA University
Bengaluru, India
jay.b.simha@abibasystems.com

***Abstract*—** Crime is a perennial social problem to be controlled by the security forces. Over a period of time, several approaches have been proposed to understand and forecast the crime rates and are available in the literature. Once of the most widely used approach is to use a method called Compstat, developed by US police departments. However, due to its assumptions and simplistic model, Compstat will not effectively capture the patterns from the data. In this paper, a novel approach to forecasting and classification of the crime is proposed. The forecasting is based on regression using attributes derived from time series data. The forecasting is done using five different methods including average based methods (SMA, HW, ARIMA), linear models (MLR) and non linear models (GAM). In addition, the crime hotspots will be derived from the forecasted values. The proposed approach was tested on real world data set from the Chicago police department for three years. The results of the proposed systems are compared with the standard approach from Compstat. The results indicate the proposed system provides a much better fit of the data compared to the standard approach.

***Keywords*—**Crime forecasting, Time series, Regression, hotspots

I. Introduction

Crimes in Chicago are 50% higher than the national average[11,12]. American police department currently use Compstat to forecast crimes for the future period –one month. Compstat stands for 'computer statistics' dealing specifically with a crime. Compstat is used by the Department of Justice (DoJ) in the USA to track, analyze crimes, and allocate resources. The crime estimates need to be pragmatic as they would enable the police department to optimally allocate resources and thus to exercise preventive measures. This article introduces three different models namely Simple Moving Average, Multiple Linear Regression (MLR) and a non-linear model, a GAM model to forecast monthly crimes. One may question as to why a Multiple Linear Regression (MLR) model for a forecasting problem. Since the crime series found to be non-stationary, for a non-stationary kind of a series, an Attribute Oriented Regression (AOR) being employed to forecast the future period. Attribute Oriented Regression forecasting is similar piece wise linear regression whereas multiple windows of the series are created, and the response is regressed on the new features. Through this paper, an effort has been made to replicate the concept to forecast a non-stationary time series of Chicago crimes. The models developed have been compared with Compstat base model[3] ,4currently used by the US Department of Justice for performance measures.

II. Lierature Review

After the major successes of crime mapping by police in the 1990s, in 1998 the US National Institute of Justice (NIJ) awarded five grants to study crime forecasting for police use as an extension crime mapping. Instead of only mapping recent crimes and assuming that observed patterns would persist; the objective was to forecast crime one period ahead with results displayed as maps. With accurate short-term crime forecasts, police would be able to take tactical actions such as targeting patrols to hot spots and conducting surveillance for deployment of special units.

Markdy et.al [7] illustrate clustering of the indexed crime data of the province of Misamis Occidental, Philippines and prediction of its occurrence in the next five years. The study utilized the k-means clustering algorithm and Autoregressive Integrated Moving Average (ARIMA) model to cluster and forecast the indexed crime data.

Alwee et.al[1] illustrate the blend of Support Vector Regression (SVR) and Autoregressive Integrated Moving Average (ARIMA) models. Particle swarm optimization is used to optimize the parameters of SVR and ARIMA models. The proposed model is equipped with features selection that combines grey relational analysis and SVR to choose the significant economic indicators for the larceny-theft rate. The experimental results show that the proposed model has better accuracy than the linear, nonlinear, and existing hybrid models in modeling the larceny-theft rate of the United States.

Varvara et.al [9] demonstrate prediction models using linear regression, logistic regression and gradient boosting techniques. The accuracy MAE was achieved with linear regression was 17 with an R squared value of 0.90.

Bao Wang et.al [10] discuss crime forecasting at small spatial and hourly temporal scales by adapting ST-ResNetstructure.

Bowen et.al [2] demonstrates the application of a random forest model to predict crimes. The modelling was performed stepwise by first adding historic violent crime predictors lagged over a 12-month period followed by models that included historic violent crime predictors lagged over a 24-month period, historic non-violent property crime, demographic data, and finally, business density data.

Lin et.al [6] demonstrates the use of machine learning models like KNN, Ensemble and DNN models to predict the crimes and the performance was compared with 11 month MA baseline. In this case, DNN has performed better than the rest of the models. This study was conducted for one of the largest cities of Taiwan, Taoyuan.

After a review of the latest literature, it has been observed that linear regression may be suited for crime forecasting if the time series data is transformed into regression form. In this work, we propose a framework to apply the regression time series data of crime in Chicago. In addition, the hot spot classification using the predicted values is also done to provide tactical information for the planners.

III. Experimental Methodology

The methodology consists of five modules involving data collection, data preparation, exploratory data analysis, model building and diagnostics, and classification for hotspots.

A. Data Collection

The data was collected from the official Chicago Data Portal for the period 2015 to 2017. The original data comprised of 0.799 million records with about 18 features. The data was then transformed into monthly aggregates of top five crimes namely “Theft”, “Battery”, ‘Criminal Damage”, “Assault” and “Deceptive Practice” year wise across 50 wards. In this way, the new data comprise 50 rows and 37 columns.

B. Data Preparation

A seven-month rolling window of the data was created for 36 months, which lead to 30 rolling windows with 1500 records. The final data used for analysis comprising of 8 columns inclusive of ward number and 1500 rows. The data was divided into train and test with 1300 and 200 records respectively. The seventh month was used as response and the remaining six months as predictors. The rolling window is similar to piece wise linear regression, which can address non-linearity, common with time series.

C. Exploratory Data Analysis

The fig1 of EDA represent the number of crimes year wise. It can be observed that there is an increasing trend in crimes. Fig2 represent the Pareto analysis of the different crime types. The crime types “Theft”, “Battery”, ‘Criminal Damage”, “Assault” and “Battery”, ‘Criminal Damage”, “Assault” and “Deceptive Practice” are the top five crimes contributing to roughly 66% of the total crimes.

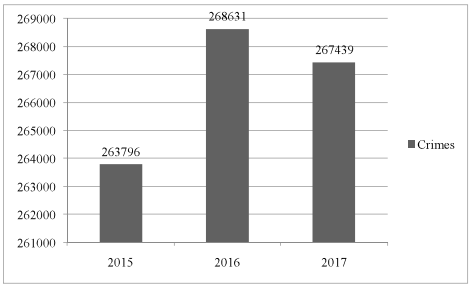

Fig1. Yearly distribution of crimes

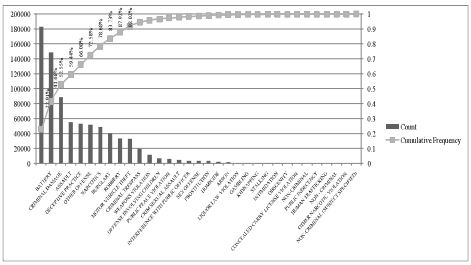

Fig2. Pareto of top5 crimes

Fig3 represent the distribution of crimes by location. The locations “Street”, “Residence” “Apartment” and “Side walk” contributed to 61% of the total of the total crimes.

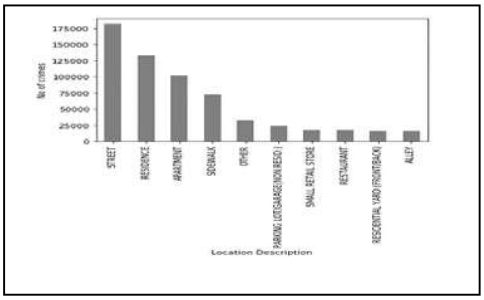

Fig3. Distribution of crimes by crime type

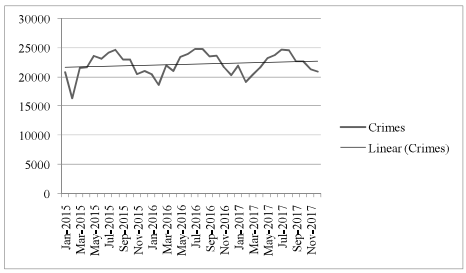

Fig4. Monthly crime trends

Fig4 represent the monthly time series plot of overall crimes for the period 2015 to 2017. There is an upward trend and the series is non-seasonal. Further, the series has a unit root present and with differencing has converged to stationarity as depicted in fig5. In fact, this reasoning has necessitated trying attribute-oriented regression forecasting model discussed subsequently. It should be noted here that the traditional models like Holts-Winters and ARIMA are unable to capture patterns properly, which is evident from figures 6 and 7 respectively.

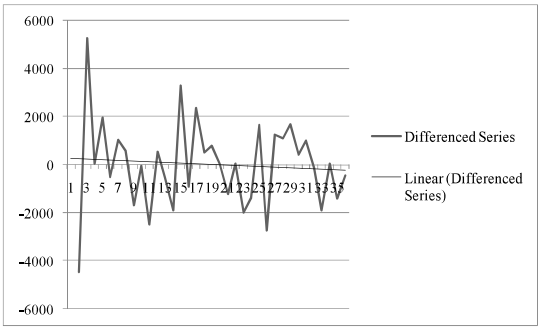

Fig5. Differenced series of monthly crimes

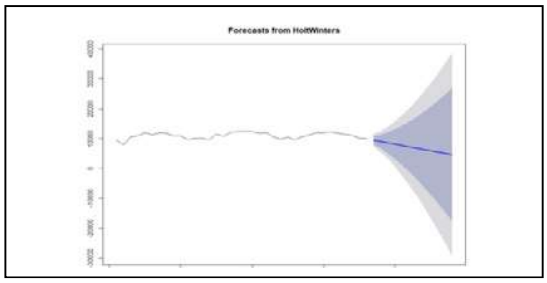

Fig6. Forecasts from HoltWinters

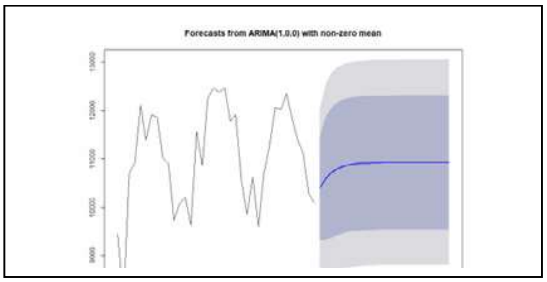

Fig7. Forecasts from ARIMA(1,0,0)

D. Model building and diagnostics

The three models developed are Simple Moving Average, Multiple Linear Regression, and the Generalized Additive Model. Simple moving average techniques involve simple arithmetic mean to compute the forecast for the next period. Multiple Regression models comprised of seventh month actual values as the independent variable and was regressed on the previous six months predictors. The Generalized Additive Models are used to model non-monotone relationships between input and output variables, which is a shortcoming with regression models. The one month ahead forecasts of MLR and GLM were compared with the observed to derive performance measures (loss functions) Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These measures were then compared with the Compstat model which is nothing but the simple moving average method of forecasting.

The SMA, MLR and GAM models have been developed with the base data of 1500 records (ward-wise 7 months rolling window), of which 1300 records are used to train the models especially MLR and GAM and the remaining 200 records to validate the results. Figures8 to 10 represent the plots of actual versus the forecasted values for the SMA, MLR, and GAM respectively. The plots of MLR and GAM reveal the best fit of the data compared to SMA. Further non-linear model is not recommended in this case as GAM has finally converged into a linear model.

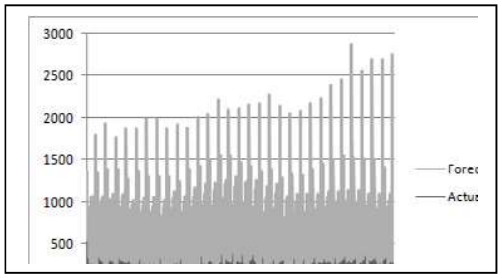

Fig8. SMA-Actual vs. predicted

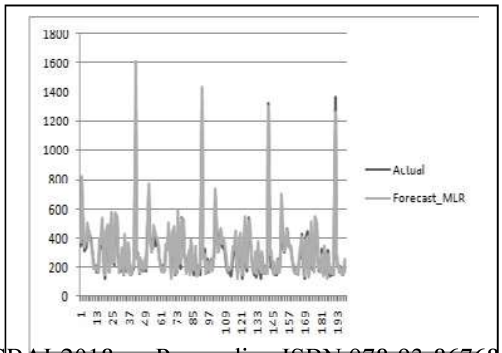

Fig9. MLR-Actual vs. predicted

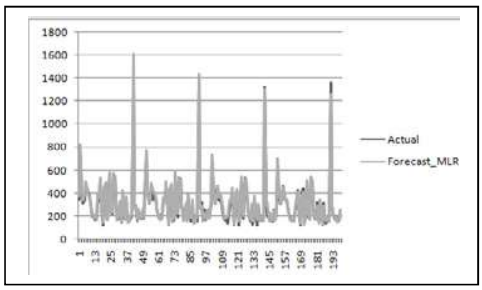

Fig10. GAM-Actual vs. predicted

The figures 11 and 12 are the line plots of the residuals. The errors of SMA are exhibiting a wave like patterns whereas the MLR and GAM errors exhibit randomness.

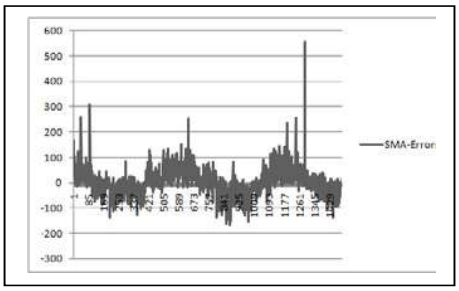

Fig11. SMA forecast residual plot

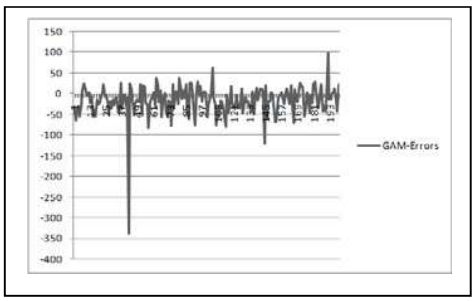

Fig12. MLR forecast residual plot

This confirms that the Compstat way of forecasting using simple moving average techniques are unable to capture the series patterns and hence not suited for one month ahead forecasts.

The following table provides the comparison of loss functions like MAE and MAPE of SMA, MLR and GAM models respectively. It can be noted that both MLR and GAM provide the same output of 9.95% MAPE.

Table 1. Errors for different models

Model	Mean Absolute Error	Mean Absolute Percentage Error	AIC
Simple Moving Average (SMA)	35.92	12.347	NA
Multiple Linear Regression (MLR)	26.51	0.0995	13130.5
Generalized Linear Model (GLM)	26.51	0.0995	13130.5

E. Classification

Post forecasting using several methods like SMA, MLR and GAM models, the crimes predicted through Compstat and MLR model are classified on the basis of a simple rule - any ward having the crime rate more than the average is classified as a hotspot. A tabular heat map of the Compstat based and MLR based approaches are shown in Fig 13.

It can be observed that a naïve approach like SMA/Compstat may be easy to implement, but not effective in really identifying the regions of interest in a hotspot analysis. On the other hand, the advanced methods like the one proposed in this paper gives much better lift in terms of decision support systems for monthly planning of the resources for monitoring the hotspots.

IV. Conclusion

Crime is a complex problem and monitoring crime on a periodic basis requires good planning, which in turn requires better decision support. In this work an alternative approach for crime forecasting and classification based on regression is proposed in place of the currently used Compstat approach.

It has been observed that the traditional forecasting models like HoltWinters and ARIMA are not able to completely capture series patterns and hence are not suited for one-month ahead forecasts. In addition, the results from the experiments on Chicago crime data indicate that the proposed regression approach better profiles the crime data. Also, the linear model is more suited than a nonlinear model. Considering these situations, it is recommended that attributed oriented regression model (MLR) is best suited for one-month ahead forecasts, which gives better results than Compstat model. Since the MLR way of predicting crimes is found to be far more superior to the Compstat way of computing, it is obvious that MLR based HotSpot classifier would outperform the Compstat based classifier. The work is under progress to include other risk factors in the models and additional types of models to improve the forecast accuracy.

References

[1] R.Alwee, S.M.Shamsuddin, and R Sallehuddin. "Swarm Optimized Grey SVR and ARIMA for Modeling of Larceny-Theft Rate with Economic indicators." International Journal of

Computational Intelligence and Applications 16, no. 02,2017: 1750008.
[2] D.A.Bowen, L. M. Mercer Kollar, D.T. Wu, D.A. Fraser, C.E. Flood, J.C. Moore, E.W. Mays, and S.A. Sumner. "Ability of crime, demographic and business data to forecast areas of increased violence." International journal of injury control and safety promotion, 2018: 1-6.
[3] J.Cohen, Department of Public Policy and Management Information Systems, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA, 3TruNorth Data Systems, Freedom, PA., 2005
[4] S.R.Flaxman, "A general approach to prediction and forecasting crime rates with Gaussian processes." Heinz College Second Paper. Pittsburg: Carnegie Mellon University, 2014
[5] W.Gorr, and R.Harries. "Introduction to crime forecasting." International Journal of Forecasting 19, no. 4, 2003: 551-555.
[6] Y.Lin, M.Yen, and L.Yu. "Grid-Based Crime Prediction Using Geographical Features." ISPRS International Journal of Geo-Information 7, no. 8, 2018: 298.
[7] O.Markdy., A.M. Sison, and A.A. Hernandez. "Mitigating vulnerabilities through forecasting and crime trend analysis." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 57-62. IEEE, 2018.
[8] McDonald, P. P. Managing police operations: Implementing the NYPD crime control model using COMPSTAT, Wads-worth, 2002.
[9] I.Varvara, and S.Ivanov. "Crime rate prediction in the urban environment using social factors." Procedia Computer Science 136,2018: 472-478.
[10] B.Wang, Y.Penghang, A.L.Bertozzi, P.J..Brantingham, S.J. Osher, and J.Xin. "Deep Learning for Real-Time Crime Forecasting and its Ternarization." arXiv preprint arXiv:1711.08833 , 2017.

Web references:

[11] https://www.areavibes.com/chicago-il/crime/) - Last accessed 27 Oct 2018
[12] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2 - Last accessed 27 Oct 2018