Comparing Models for Time Series Analysis and Forecasting of London Crime Data

Aleksandra Dedinec[1], Sonja Filiposka[1] and Anastas Mishev[1]

[1] Ss. Cyril and Methodius University - Skopje, North Macedonia

(E-mail: aleksandra.kanevche@finki.ukim.mk, sonja.filiposka@finki.ukim.mk, anastas.mishev@finki.ukim.mk)

Keywords: crime forecasting; seasonal analysis; time series.

Introduction

Crime forecasting has become a major trend over the past years based on the availability of new technologies that can be used to improve prevention efforts by supporting decisions related to efficient resource allocation [1]. The availability of open data related to crime and its combination with other big data information such as demographics, geographical/location-based information or economic and social parameters provide an opportunity to gain insight into past events that can help uncover patterns and observe trends. However, the information that can be discerned from the data heavily relies on the data quality, accuracy and completeness [2].

The standard method for analysis of the available crime data is by using a hot spot crime map that correlates the available crime information with the geographical distribution of the crimes in the analyzed area [3]. The visual map display of crime activities provides an easy means for spotting hot spots (geographical crime clusters) that require possible police action. This type of visualization has become a standard type of spatial analysis of crime data. In addition to these efforts, the wide application of statistics and/or machine learning algorithms has spread to the crime data analysis providing additional means for pattern recognition based on time series [4]. The analysis is done by treating the crime data available as a time series data and using specialized methods to extrapolate time series ahead of time. Univariate or multivariate analysis can be employed aiming to uncover correlation between multiple time series and uncover new data patterns. The ability to accurately forecast the future crime trends heavily depends on the data quality and quantity, characteristics of the time series analyzed, and model and its tuned parameters used to make the forecast. In depth investigation is needed to uncover the best model that provides the most accurate forecast while the careful tuning of its parameters has a tremendous impact on the confidence interval size.

This paper aims to provide an initial analysis made by applying time series forecasting methods and models on the open data London crime dataset. The obtained results uncover different patterns in the dataset related to seasonal activities. The forecasting techniques presented are used to discuss the accuracy and expectations that can be made from the future crime forecasting based on the methodology used.

Time Series Forecasting Models

Quantitative forecasting is a technique used for a time series observed at regular intervals of time that aims to predict the future values, that is the predictor variable, of the time series in the provided intervals. The prediction can be made solely based on the past values of the predictor variable, or by using additional external variables which may affect the model behavior such as strength of economy, population, etc. The choice of model depends on the available data structure and the relationship between the forecast variable and any external variables.

Regardless of the predictive model type the crime data set analysis must be based on time series forecasting that aims to estimate the future crime values. In addition to increasing or decreasing trends, crime datasets may also display the property of seasonality, that is have periodical patterns that repeat with a constant frequency (i.e. every 12 months). The seasonality property affects the choice of predictive model in the sense that there are special versions of the prediction models that are tuned to work with time series that exhibit the property of seasonality and manage to incorporate this property into the future values predicted with the model. Statistical models typically include linear regression, exponential smoothing, autoregressive integrated moving average (ARIMA), while the AI models are mostly based on artificial neural networks or support vector regression. Since the dataset that is used in this paper has seasonality properties, in this section three seasonal prediction models are presented as representatives of three groups of prediction models: the seasonal naïve model that is used as a benchmark for the models prediction ability, the seasonal ARIMA model as a representative of statistical prediction models based on autocorrelation in the data, and the NNAR model as a representative of the later generation neural networks based prediction models.

Highly seasonal data can be forecasted in a very simple, but effective way using the seasonal naive forecasting method [5]. The method works in such a way that each new predicted value for the time series *y* is equal to the last observed value from the same season (i.e. the same month from last year):

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

where *m* is the seasonal period - 12, and *k* is the number of complete years in the forecast period. This prediction method is very easy to implement, is fast, and in many cases it provides very good first order estimation. It is also used for one of the accuracy measurements introduced later.

The non-seasonal ARIMA model [6] is based on a combination of autoregression, moving average and differencing, and can be expressed as:

$$y'_t = c + \varphi_1\, y'_{t-1} + \ldots + \varphi_p\, y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where $y'_t$is the differenced time series. The right hand side contains predictor variables that include lagged values and lagged errors. The model is presented as ARIMA(*p*, *d*, *q*) where *p* represents the order of the autoregression, *d* is the degree of involving the differencing, and *q* is the order of the moving average. The fine tuning of the values of the three parameters for the ARIMA model is not an easy task. However, for the purposes of widespread use with effectively high accuracy, there is an implementation of the ARIMA model in R, the so-called auto ARIMA, that manages to automatically set the appropriate values of the parameters, making the model suitable for use for the general public (law enforcement in this case). In the case of seasonal data, a seasonal ARIMA model can be created by using additional seasonal terms in the original model. This is effectively ARIMA(*p*, *d*, *q*)(*P*, *D*, *Q*)*m*, where the second part of the model refers to the seasonal part and *m* is the number of observations per year (12 for monthly data). The seasonal part also involves backshifts of the seasonal period.

Lately there has been a shift from using pure statistical models to artificial intelligence (AI) models for the purpose of forecasting. The main goal of this shift is to employ techniques that will enable increased accuracy performance of the forecasting model. In addition, the AI models are able to also handle nonlinear and nonstationary data, thus adapting more easily to complex data that is usually a mix of linear and nonlinear information. Neural networks [7] can be used for forecasting and are favored in cases when there is a nonlinear relationship between the forecast variable and the predictors.

When using neural networks for time series forecast, the input to the network is the lagged values of the time series, a so-called neural network autoregression (NNAR) model. NNAR(p, k) represents a feed-forward network with one hidden layer, where p is the number of lagged inputs and k is the number of nodes in the hidden layer. In the case of seasonal data, the model transforms into NNAR(p, P, k)m with k hidden neurons and P time more inputs accounting for the seasonal points. Based on this definition, NNAR(p, P, 0)m is in effect ARIMA(p, 0, 0)(P, 0, 0)m without the ARIMA restriction that the parameters must be stationary. Prediction intervals for neural networks are obtained using simulation and bootstrapped residuals.

The accuracy of the forecasts in this paper is measured using percentage errors (mean absolute percentage error - MAPE) and scaled errors (mean absolute scaled error - MASE) [8]. Since, $MASE = MAE/MAE_{in-sample,naive}$where MAE is the mean absolute error produced by the actual forecast and $MAE_{in-sample,naive}$is the mean absolute error by a naïve forecast, it must be noted that if MASE > 1, then the actual forecast is performing worse when compared to the naïve forecast and should be disregarded.

London Crime Dataset

The information about police recorded crimes in the United Kingdom are provided as open data since 2011. There are a number of limitations that must be taken into account when analyzing the dataset. The street-level information provided on the data.police.uk website has been anonymized in order to protect the privacy of the persons involved. After the first years, further privacy-driven obfuscation has taken place such as aggregating all violent and sexual offences into one category. In addition, fraud related offences are excluded from the available datasets. The available data includes references to the Lower Layer Super Output Area (LSOA) or borough where the crime has taken place, the crime type description with its major and minor category, and the frequency of the given crime type for a specific month and year.

The dataset used in this paper is a subset of the dataset available from the data.police.uk website and includes the report on crimes in the London metropolitan area in the period of 2008 to 2016. There are 33 different boroughs and 9 distinct major crime categories out of which sexual offences and fraud have been deprecated, with 32 minor crime categories mentioned.

In Fig. 1 the general statistics overview of the frequency of major crime categories per year for all available boroughs is presented. As mentioned previously, the sexual and fraud related offences appear in small numbers only in the first few years of the dataset and are thus not visible in the presented results. Based on this broad general view of the data, no obvious trends or other changes can be observed except for increasing in crime frequency for violence against the person. In Fig. 2 the crime frequency broken down into major categories for two different boroughs is presented. The results for the City of London borough clearly show the problem with data quality of the data set since no data is available before 2011 for this borough. Another observation is the rank of the frequency observed for the two boroughs where again the City of London area shows a very low frequency of crime in each category which is unexpected when compared to the rest of the boroughs.

The analysis available related to this specific dataset includes statistical analysis of the available data based on the information provided in the original, or extended data set [9]. The extended data set includes information on the outcome of the offence, but also ethnicity, and age range of the persons involved. In addition to the standard statistical analysis, in [10] an anomaly detection algorithm is applied to the dataset and additional information is extracted such as that year 2016 stands out from the overall crime distribution with the frequency of theft decreasing and harassment appearing as a new type of frequent crime. In addition, the results show that assault with injury is the highest frequency crime in the central boroughs of London. An interesting observation can be found in [11], where the authors build a spatial dependence graph model for 14 different crime categories recorded within April 2015 in the City of London. The authors results show interdependency between robbery, drugs and theft from person, and burglary, vehicle crime and other theft. This analysis shows that these types of crimes are related, while, on the other hand, occurrences such as public order, bicycle theft, criminal damage and arson are not related to any other types of crime.

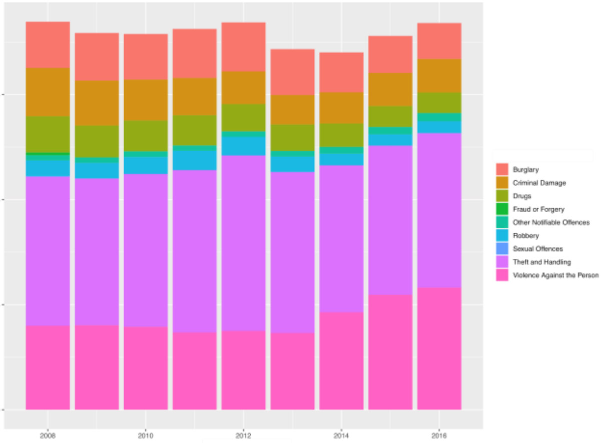

Figure 1. Major category crime occurrences over time in the London metropolitan area

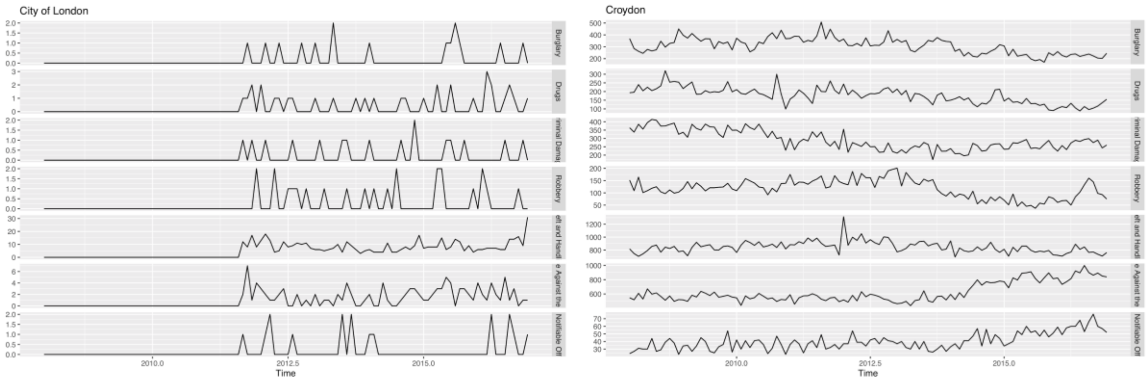

Figure 2. Crime frequency per major category in time for the City of London and Croydon boroughs

In the next section we aim to provide some initial results when applying time series forecasting models on this dataset. For these purposes the data has been divided into training and test data set roughly following the 80-20% rule, where the data available from 2008 to 2014 is used as the training set, and we use the test data for 2015 and 2016 to estimate the accuracy of the used prediction models.

London Crime Time Series Analysis and Forecasting

As previously mentioned, the London crime data set exhibits seasonality trends relating crime categories and boroughs with specific time of year. Such an example is presented in Fig. 3 where the crime rate significantly rises in the Kensington and Chelsea borough during the month of August related to the typical vacation time of the residents of one of the richest boroughs in the metropolitan area that has the highest median house price for 2015 according to the open data on London borough profiles [12]. This result shows that further in-depth investigation is needed to understand crime patterns across borough by combining and correlating the available data from the London dataset.

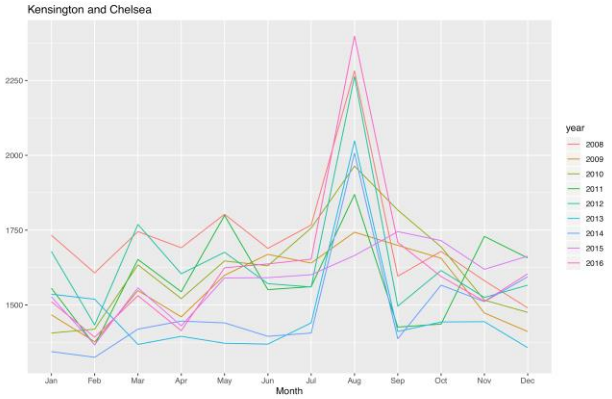

Figure 3 Seasonal analysis of Kensington and Chelsea borough

The results presented in Fig. 4 depict the forecast results (together with the 95% and 80% prediction intervals) obtained by applying the three representative forecast models to the total crime in London. The accompanying MAPE and MASE values show that the seasonal ARIMA model performs the best in the forecast with less than 4% of error compared to the test set, and slightly better results compared to the seasonal naïve method.

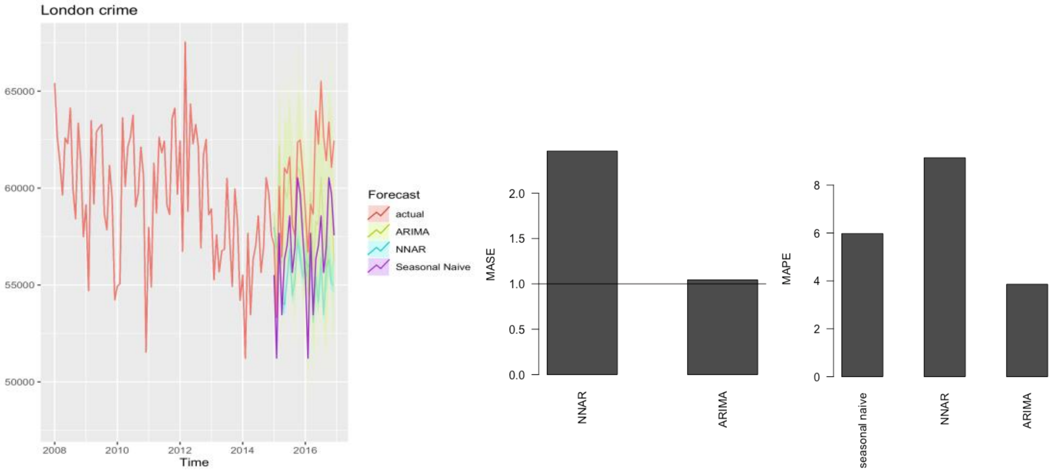

Figure 4 London crime forecast models comparison and MASE and MAPE errors

It is interesting to note that if the forecast is made on a per major crime category level, then the accuracy of the models is different for different categories, as is represented in Fig. 5. While the ARIMA model shows good results for all categories with errors ranging from 10% to 3%, the NNAR model should be disregarded in favor of the seasonal naïve in half of the cases. The extremely high error reported for violence against the person major category is due to the high positive trend for the crime category which is not observed for the other types of categories. The results show that the NNAR model copes poorly when there is a significant change in the data trend. On the other hand, in this case ARIMA also exhibits very wide prediction intervals.

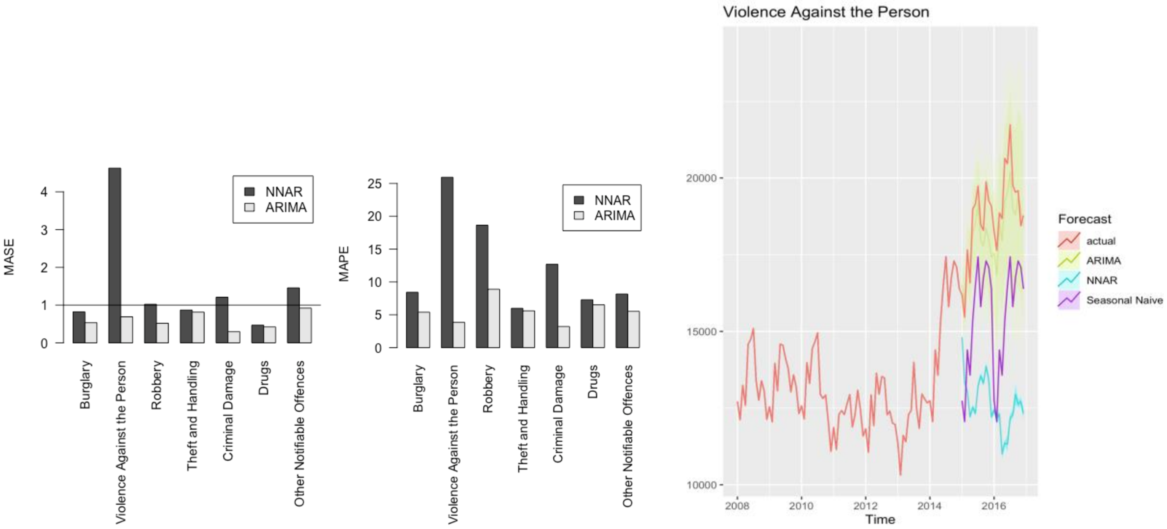

Figure 5 Crime forecast errors per major crime category, forecast for the Violence against the person category

In Fig. 6 the forecast errors are presented when forecasting the total crime per borough. Due to the small amount of data for the City of London borough, the reported error is tremendous and thus has been omitted in the figure. When comparing the models, again the ARIMA model outperforms NNAR in almost all cases (except for 2 boroughs Hillingdon and Brent both of which show future trends similar to the past). However, in this case, for ⅓ of the boroughs even the ARIMA model does not outperform the seasonal naïve model. This behavior is due to the fact that the trends observed in these boroughs (such as Lewisham) are opposite for the two forecasted years, compared to the trained data set trend, see Fig. 7. Again, the model shows this problem by providing a very wide prediction interval.

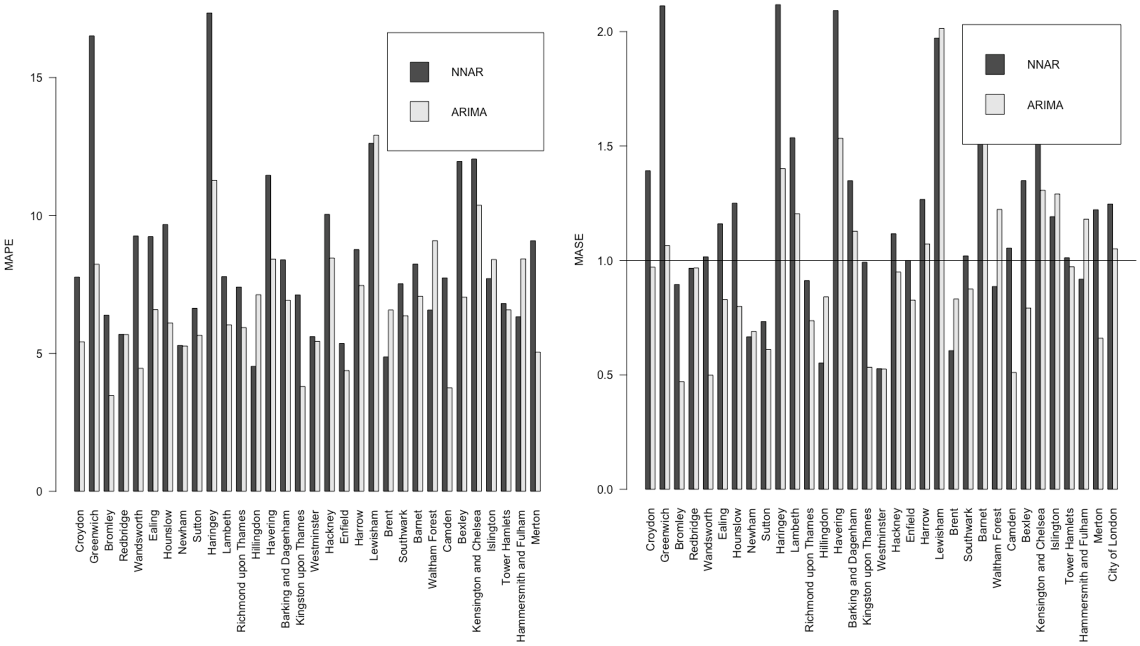

Figure 6 Crime forecast errors per borough

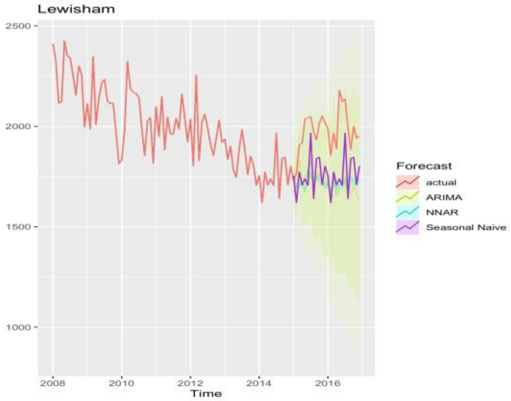

Figure 7 Crime forecast models comparison for the Lewisham borough

Conclusion

The initial dataset analysis and comparison of the ARIMA and NNAR forecast models for the London crime dataset has uncovered some interesting results that require further analysis. In addition to the typical seasonality of data, the dataset might have multiple seasonality characteristics that need to be uncovered, such as increased crime rates related to other socio-economic parameters such as wealth, holiday patterns, etc. When forecasting, the models employed have a difficulty to make a good prediction in the case of major changes in the data trend compared to the known history. This problem is accentuated for the NNAR model that is thus regarded an unfitting for the dataset observed. Additional analysis has been made with other types of neural networks such as feed-forward with multiple hidden layers, recurrent neural networks and other deep learning techniques, however the results have not improved. The main reason for this under performance is because of the small amount of data available and the big difference between the training and testing data. In order to confirm this hypothesis, further forecast attempts need to be made using the detailed London crime dataset available from the data.police.uk repository where the crime data is available on the hourly/daily level instead of the aggregated monthly report. Extending the time interval by including the data after 2016, as well as using other external variables should also contribute towards confirming this hypothesis.

References

[1] Gorr, Wilpen, and Richard Harries. "Introduction to crime forecasting." International Journal of Forecasting 19.4 (2003): 551-555.

[2] Tompson, Lisa, et al. "UK open source crime data: accuracy and possibilities for research." Cartography and geographic information science 42.2 (2015): 97-111.

[3] Eck, John, et al. "Mapping crime: Understanding hotspots." (2005): 1-71.

[4] Khairuddin, Alif Ridzuan, Razana Alwee, and Habibollah Haron. "A review on applied statistical and artificial intelligence techniques in crime forecasting." IOP Conference Series: Materials Science and Engineering. Vol. 551. No. 1. IOP Publishing, 2019.

[5] Hyndman, Rob J., and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018.

[6] Kumar, S. Vasantha, and Lelitha Vanajakshi. "Short-term traffic flow prediction using seasonal ARIMA model with limited input data." European Transport Research Review 7.3 (2015): 21.

[7] Samarasinghe, Sandhya. Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. Auerbach publications, 2016.

[8] Hyndman, Rob J., and Anne B. Koehler. "Another look at measures of forecast accuracy." International journal of forecasting 22.4 (2006): 679-688.

[9] Vaibhav Desai, “London Crime Analysis using bq_helper and pandas”, Python notebook using data from London Crime Data on Kaggle, 2019

[10] Tolpin, David. "Population anomaly detection through deep gaussianization." Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. ACM, 2019.

[11] Eckardt, Matthias, and Jorge Mateu. "A spatial dependence graph model for multivariate spatial hybrid processes." arXiv preprint arXiv:1906.07798 (2019).

[12] London datastore, London Borough Profiles and Atlas, https://data.london.gov.uk/dataset/london-borough-profiles