Time-Series Analysis of a Historical Crime Data

Roshan Koirala
roshankoirala77@gmail.com

***Abstract*—Historical crime data from the county of Carry, NC have been analyzed from a time-series perspective. The data exhibits strong evidence for the annual additive seasonality and additional trend. The data is fitted on a seasonal auto-regressive integrated moving average model and forecasting have been provided. The model performs well on the data with a maximum of the p-value of Ljung-Box statistics being 0.2 of the 12^{th} lag suing the residue of the prediction by a $SARIMA$ model.**

***Index Terms*—Crime Data, Stochastic Process, Time-Series Analysis, SARIMA Model, Holt-Winter Forecasting**

I. Introduction

Crime is defined as an unlawful act that is punishable by the authorities that are responsible for maintaining law and order in society. As our society evolves the types and definition of crime changes too. What is considered a crime in one society or at one point of time in history may not be a crime in a different society or from a different point of time in history. However, there can be a general consensus that we can define some activity which affects the right of other individuals to peacefully live in society. Society needs to discourage such activities. One obvious response by the authority of the criminal activity is to punish the culprit. However, they also need to take the actions that prevent the crime in the society. The activity of crime can have psychological to the social and economic dimensions. Understanding them can help us prevent crime and focus on the root of the problem. One of the approaches to get an insight into the crime is to study the historical crime data and get insight from there.

In this work, we study the historical crime data provided by the county of Carry, North Carolina [1] based on time-series analysis. The geographical location and the Geo-spatial distribution of the crime incident are presented in FIG 1. The data contains more than 1 million records of crime mostly from the year 2000 CE to 2020 CE. There are few records of crime before 2000 CE but their occurrences are far below the average of the total occurrence of crimes after 2000 CE. So, we consider data before 2000 CE as outliers and discard them. The detail of each incident in the data set is presented in 37 columns and many columns provide redundant information. Especially we are interested in the date and number of crimes. We also studied in detail the crime by the time of the day, their types, and the geographical location of the incident. The detail of the exploratory data analysis can is presented in the companion notebook related to this work [2].

The data provides information about the various classification of crimes. First, they are classified as domestic vs non-domestic crimes. Only about 9% of the crimes are non-domestic crimes. Nearly above 31% of the reported crime are larceny. The next major category is vandalism which accounts for slightly above 13% of the total number of crimes. North Carolina State Bureau of Investigation documents crimes classifying them into two major classes [3]. Part 1 crimes include murder and non-negligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny, motor vehicle theft, and arson. All other offenses are classified as part 2 offenses. Nearly 42% of the reported crimes are part 1 crime. See FIG. 2 for the detail. Although only a subset of part 1 crime is used to calculate the crime index by the North Carolina State Bureau of Investigation, we include all crimes in our study.

There are many works done in this direction in the past. Chamlin, M. B. (1988) studied the univariate and bi-variate ARIMA analyses to study the lagged crime-arrest relationship using 13 year's monthly crime and arrest data from Oklahoma City and Tulsa, Oklahoma [4]. In contrast to their work, we find the optimal model for our uni-variate time series data being SARIMA with annual seasonality. The evidence of the seasonality in crime data was studied in Landau S. F. & Friedman D. (1993), who investigated the annual and other seasonality of homicide and robbery in Israel and used SARIMA to model the 8 years long crime data [5].

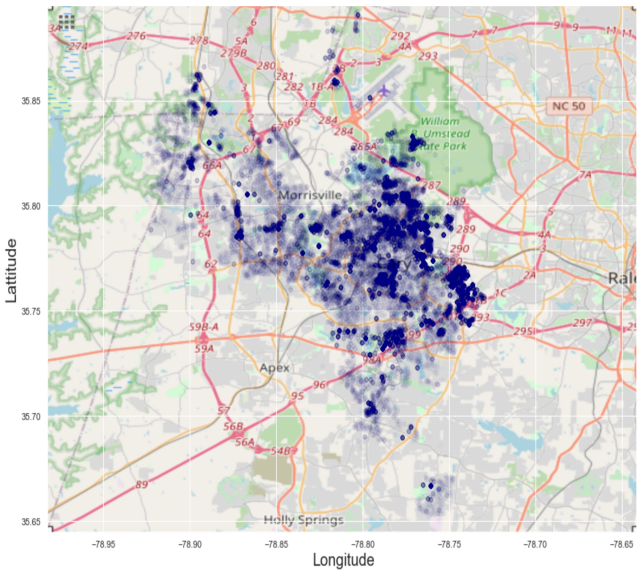

Fig. 1. Historical incident of crimes in county of Carry NC USA. The navy blue open circle indicate one incident of crime. Darker the color more occurrence of crime in that are.

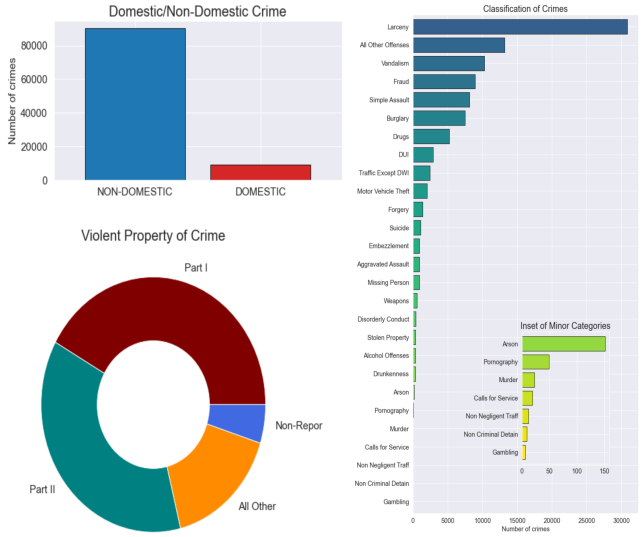

Fig. 2. Classification of crimes. See text for the definition of part I and part II crime.

Corman, Hope & H. Naci Mocan. (2000) studied monthly crime data spanning 30 years from New York City. Their time series analysis model is more specific to the crime data and depends on many other variables like police activity, poverty, etc [6]. On the other hand in this work, we work on a more general approach to the time series analysis and do not consider the further assumption regarding the effect of other variables in the model. Greenberg, D. F. (2001) studied the effect of unemployment on crime rates by analyzing nationally aggregated data finding a negative result [7]. P. Chen, H. Yuan, and X. Shu (2008) studied weekly data spanning 50 weeks from a city of China for time series analysis based on the ARIMA model [8]. In contrast to them our study contains the data from a longer period of time, 20 years to be precise, of time and we investigate monthly, weekly and daily data in detail to find the optimal model. Also, we see a strong sign of annual seasonality which their study lacks simply because the span of data is shorter than a year.

Recently a deep learning method based on Recurrent Neural Network and its variants is a serious contender of the traditional SARIMA model in time-series forecasting. Wang, B., Yin, P., Bertozzi, A.L. (2019) studied six months of hourly crime data from Los Angeles (LA) along with the weather and historical holiday data [9]. But their model is based on Convolutional Neural Network rather than RNN. Stec, Alexander, and Diego Klabjan (2018) have used RNN, especially Long Short Term Memory (LSTM), to model crime data from Chicago and Portland with additional data related to weather, census data, and public transportation [10]. However, this work does refrain from using a deep learning-based model partly because the performance of the SARIMA model is satisfactory for a small data set that is available for criminal cases and for its simplicity. Nonetheless, this can be a natural future extension of the present work by considering additional data.

The organization of this work is the following. In section II we review the basics of stochastic processes. Especially we motivate our discussion to $SARIMA$ model by discussing its components in detail. We also provide a recipe for parameter tuning and model evaluation. We present our model in section III. We show that a $SARIMA$ model fits good to the monthly crime data and presents the evaluation of the model. Finally we conclude in section IV.

II. Stochastic Processes

A. Stationary process

For a detailed review of the stochastic process discussed in this section see [11].

In statistics, a stochastic process is a process governed by random variables at each stamp of time. A random variable is defined as a map f from sample space S to a real number $\mathbb{R}$ as given by $f : S \to \mathbb{R}$. A stochastic process $X_t : t \in \mathbb{Z}$ is called stationary if there is translational symmetry in the process such that $X_{t+\tau} = X_t, \forall t \in \mathbb{Z}$ for a given τ. This condition of stationarity is called a strong stationarity condition. This condition is too restrictive to realize in most of the real world stochastic process. So we introduce a weak stationarity condition. A process is weak stationary if the mean and variance of the process remain constant over time. See FIG 3 for an illustration of a simulated stationary process.

White noise is the simplest kind of time-series data. White noise is a series of random numbers spanned over time. In white noise, the mean and variance of the time-series data are constant over time. In many cases, white noise is normally distributed random numbers and we can estimate the mean and variance of the noise using the data. In many time-series data, there are trends and seasonality in the background which are discussed in the subsequent sections. Nevertheless, the noise is usually present in the time series data, and estimation of the white noise is an important part of the time-series analysis. Mathematically white noise is given by

$$X_t = Z_t, \quad Z_t \sim N(\mu, \sigma^2) \tag{1}$$

where X_t is the value of the signal at time t and Z_t is the randomly distributed number. Here we assume Z_t is normally distributed with mean μ and variance σ^2. White noise trivially satisfies the weak stationarity condition.

B. Random Walk

Random walk can be thought of as the cumulative sum of the white noise. In the context of time-series data we consider discrete one-dimensional random walk. Mathematically

$$X_t = X_{t-1} + Z_t = \sum_{i=1}^{t} Z_i, \quad Z_t \sim N(\mu, \sigma^2) \tag{2}$$

where X_t is the value of the signal at time t and Z_t is normally distributed with mean μ and variance σ^2. The second equality follows from repeatedly applying the first equality on X_{t-1} and then X_{t-2} and so on. From this definition we can see that $X_t - X_{t-1} = Z_t$, which means that the difference of

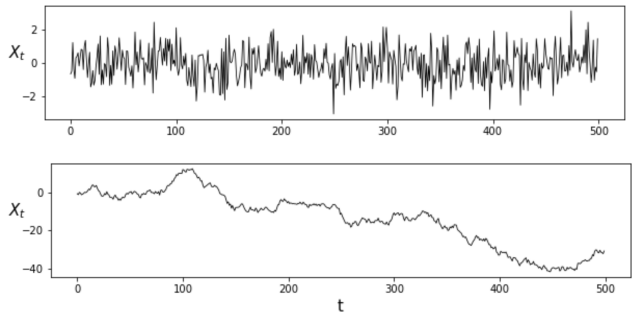

Fig. 3. (Upper) White noise. We can see that the mean and variance of the process is constant over time. (Lower) Random walk process. We clearly see the mean itself is changing over time for a random walk process.

the time-series data is a white noise. Writing this in more sophisticated language, let B be a backward shift operator such that $X_{t-1} = BX_t, X_{t-2} = B^2X_t, \cdots, \ X_{t-n} = B^nX_t$. Then a process X_t is called a random process if the related process $Y_t = \nabla X_t$ is stationary. Where $\nabla = 1 - B$ is called the difference operator such that $\nabla X_t = X_t - X_{t-1}$.

Mean and variance of the random walk process is given by

$$E[X_t] = E\left[\sum_{i=1}^{t} Z_i\right] = \sum_{i=1}^{t} E[Z_i] = \mu t, \tag{3}$$

$$V[X_t] = V\left[\sum_{i=1}^{t} Z_i\right] = \sum_{i=1}^{t} V[Z_i] = \sigma^2 t \tag{4}$$

From these relations, it is seen that the random walk does not satisfy the weak stationarity condition. See FIG 3 for an illustration of a simulated random walk process.

C. Auto-Regressive Model

Here we motivate the auto-regressive model in two different ways. Firstly, the auto-regressive model is a generalization of the random walk process. Mathematically we can write

$$X_t \ = \ Z_t + \sum_{i=1}^{p} \theta_i X_{t-i}, \quad Z_t \sim N(\mu, \sigma^2) \tag{5}$$

where $\theta_i, i = 1, 2, \cdots p$ are parameters of the model usually determined through the data, p is known as the order of the model and the model itself is denoted by $AR(p)$ model. It is important to note that when $p = 1, \theta_1 = 1$ then the model reduces to the random walk model.

Although the model still contains the random noise part Z_t, the auto-regressive part has certain degree of correlation going on in the data. Clearly there is no overall correlation in the entire range of the time-series data. However for each point in the data there is some significant correlation with data points in certain interval. Which gives the concept of the auto-correlation function which is the origin of the name of the model. The auto-correlation function of time lag k is defined as

$$\rho(k) = \frac{\sum_{i=1}^{N-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \tag{6}$$

The range of is given by $-1 \leqslant \rho(k) \ \leqslant 1$. Basically auto-correlation function is the Karl Pearson correlation of X_t with itself shifted by a time lag X_{t+k}. When $k = 1$ the value of the auto-correlation is one because the correlation of a data with itself is always one. For a white noise we expect $\rho(k > 1) = 0$.

We can work out a more practical way to calculate the auto-correlation of the process based on Yule Walker equation. Multiplying both side of equation (5) by X_{t-k} and taking the expectation

$$\begin{aligned} E[X_t X_{t-k}] &= \sum_{i=1}^{p} \theta_i E[X_{t-i} X_{t-k}] + E[Z_t X_{t-k}] \\ \gamma_{-k} &= \sum_{i=1}^{p} \theta_i \gamma_{i-k} + \sigma_Z^2 \delta_{k,0} \\ \rho_k &= \sum_{i=1}^{p} \theta_i \ \rho_{|i-k|}, \quad \text{for } k \geqslant 1 \end{aligned} \tag{7}$$

In the last line we use the property $\gamma_{-k} = \gamma_k$ and divide by γ_0 where $\gamma_0 = \sigma_Z^2$ followed by setting $k = 0$. These equations are called Yule-Walker equations. We can use this equation to find the auto-correlation of the auto-regressive processes.

The effect of a random noise Z_t cumulatively summed up in subsequent time and never dies fully though it diminishes for a convergent series. The consequence of this is that there is a high degree of col-linearity among the auto-regressive process. Hence, partial correlation is more suitable than the correlation itself for this process. Partial auto-correlation is given by

$$\rho_p(k) = \rho(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t) \tag{8}$$

Where $\hat{X}_t$ is the estimation of the X_t. For an auto-regressive process of order p we expect $\rho_p(p > 0) = 0$. Because of the random noise, there can be small leftover correlation instead

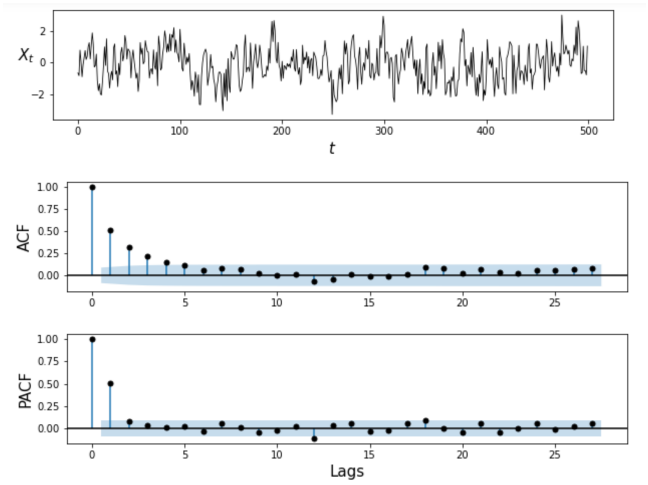

Fig. 4. (Upper) Simulated $AR(1)$ process with $\theta = 0.5$. (Middle) Auto-correlation function of $AR(1)$ is a long tailed and has no direct relation to the order of the process. (Lower) Partial auto-correlation of $AR(p)$ process has significant correlation in lag p. In our example $p = 1$.

of exact zero. The auto-correlation and partial auto-correlation of a simulated auto-regressive process is presented in FIG. 4.

AR process itself is a part of a more general $ARMA$ process. In the next section we study the second part of it called moving average (MA) process.

D. Moving Average Model

A stochastic process is called moving average process if the value at time t is given by the weighted average of the present and many previous stochastic term. Mathematically this can be formulated as

$$X_t = \sum_{i=1}^{q} \phi_i Z_{t-i}, \qquad Z_t \sim N(\mu, \sigma^2) \tag{9}$$

where $\phi_i, i = 1, 2, \cdots q$ are parameters of the model usually determined through the data, q is known as the order of the model and the model itself is denoted by $MA(q)$ model. For an moving average model of order q we expect $\rho(k \geqslant q) = 0$. We can make an auto-correlation plot as shown in FIG 5, which is a handy tool to determine the unknown order of the moving average process.

The mean and variance of the moving average process is given by

$$E[X_t] = E\left[\sum_{i=1}^{q} \phi_i Z_{t-q}\right] = \sum_{i=1}^{q} \phi_i E[Z_{t-q}] = 0, \tag{10}$$

$$V[X_t] = V\left[\sum_{i=1}^{q} \phi_i Z_{t-q}\right] = \sum_{i=1}^{q} \phi_i V[Z_{t-q}] = \sigma_Z^2 \sum_{i=1}^{q} \phi_i \tag{11}$$

From these relation we can see that the moving average process satisfy the weak stationarity condition. With a bit of algebra it can be shown that the auto-correlation function of the moving average process is given by

$$\begin{aligned}
\gamma[X_t, X_{t+k}] &= E[X_t \cdot X_{t+k}] - E[X_t]E[X_{t+k}] \\
&= E[X_t \cdot X_{t+k}] \\
&= E\left[\left(\sum_{i=1}^{q} \phi_i Z_{t-i}\right) \cdot \left(\sum_{j=1}^{q} \phi_j Z_{t+k-j}\right)\right] \\
&= \sum_{i=1}^{q}\sum_{j=1}^{q} \phi_i \phi_j E\left[Z_{t-i} Z_{t+k-j}\right] \\
&= \sum_{i=1}^{q}\sum_{j=1}^{q} \phi_i \phi_j \ \delta_{i,j-k} \sigma_Z^2 \\
&= \sigma_Z^2 \sum_{i=1}^{q-k} \phi_i \phi_{i+k}, \qquad \text{for } k \leqslant q
\end{aligned} \tag{12}$$

where in last step the contraction with Kronecker delta $\delta_{i,j} = \{1 \text{ if } i = j,\ 0 \text{ if } i \neq j\}$ have been performed. For $k = 0$ we retain the relation for the variance. Finally the auto-correlation of the moving average process is given by

$$\rho(k) = \frac{\gamma[X_t, X_{t+k}]}{V[X_t]} = \frac{\sum_{i=1}^{q-k} \phi_i \phi_{i+k}}{\sum_{i=1}^{q} \phi_i^2} \tag{13}$$

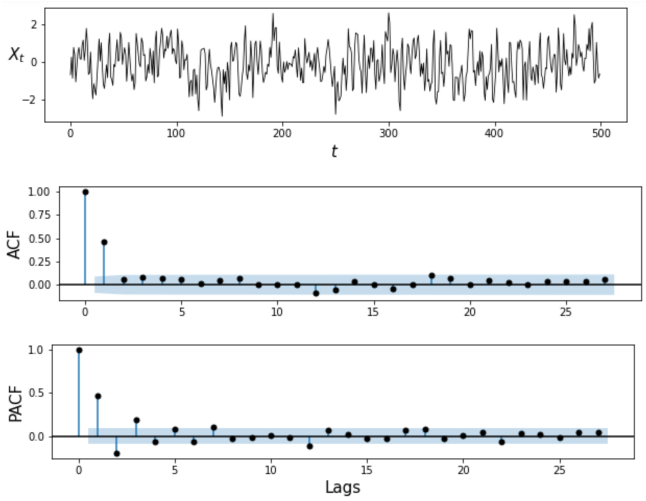

Fig. 5. (Upper) Simulated $MA(2)$ process with $\phi = [0.9, 0.6]$. (Middle) Auto-correlation function of $MA(q)$ process has significant correlation up to lag $q-1$. In our example $q = 2$. (Lower) Partial auto-correlation of $MA(q)$ process is a long tailed and has no direct relation to the order of the process.

There is a connection between AR and MA processes. We can write $AR(p)$ and $MA(q)$ processes in terms of the shift operator in the following way

$$AR(p): \quad Z_t = \alpha(B, \theta) X_t \tag{14}$$

$$MA(q): \quad X_t = \mu(B, \phi) Z_t, \tag{15}$$

where $\alpha(B, \theta) = \sum_{i=0}^{p} \theta_i B^i$ and $\mu(B, \phi) = \sum_{i=1}^{q} \phi_i B^q$. For convenience assume $q = 1$ so that $\mu(\phi) = 1 + \phi B$, then we can write

$$\begin{aligned}
MA(1): \quad Z_t &= \frac{1}{\mu(B, \phi)} X_t = \frac{1}{1 + \phi B} X_t \\
&= (1 - \phi B + \phi^2 B^2 - \phi^3 B^3 + \cdots) X_t \\
&\subset AR(\infty)
\end{aligned} \tag{16}$$

Hence we just showed that if an operator of $MA(1)$ process $\mu(\phi)$ is invertible the process has a dual formalism as $AR(\infty)$. The condition for the invertibility requires that the root of the auxiliary function of the operator $\mu(\phi)$ lies outside the unit circle of a complex plain so that the infinite sum converges. Even if we demonstrate the process for $MA(1)$ process the conclusion is more general and holds for any order such that $MA(q) \Rightarrow AR(\infty)$ and is symmetric between AR and MA process meaning $AR(p) \Rightarrow MA(\infty)$.

E. ARMA, ARIMA & SARIMA

$ARMA$ stands for the Auto-Regressive Moving Average. A process is called $ARMA(p, q)$ if

$$X_t = \sum_{i=1}^{q} \phi_i Z_{t-i} + \sum_{i=1}^{p} \theta_i X_{t-i}, \qquad Z_t \sim N(\mu, \sigma^2) \tag{17}$$

where, p and q are the orders of AR and MA process.

Similarly, $ARIMA$ stands for Auto-Regressive Integrated Moving Average. A process X_t is $ARIMA(p, d, q)$ if a

related process Y_t is $ARMA(p,q)$ where $Y_t = \nabla^d X_t = (1-B)^d X_t$ and d is the order of the difference operator. The equation for a $ARIMA$ process is given by

$$\alpha(B,\theta)\nabla^d X_t + \ \mu(B,\phi)Z_t, \tag{18}$$

For $d = 0$ this equation reduces to the equation for $ARMA$ process.

Similarly, $SARIMA$ stands for Seasonal Auto-Regressive Integrated Moving Average. In addition to the auto-correlation (or partial auto-correlation) to the immediate lags there can be auto-correlation (or partial auto-correlation) s lags after again. Which means there is significant auto-correlation between X_t and X_{t+s} while there is no significant correlation between X_t and $X_{t+j}, j < s$. In this case, the process has seasonality of order s. A process is $SARIMA(p,d,q,P,D,Q,s)$ means the process is $ARIMA(p,d,q)$ and also $ARIMA(P,D,Q)$ after lag s.

Now there are a lot of moving parts in the $SARIMA$ model. Again the ACF and PACF can give the idea about the order of the process but we have to check against all the combinations of the parameter in order to find the optimum model. We follow the parsimony principle and choose the simplest model possible. A general guideline is

$$p + d + q + P + D + Q \leqslant 10 \tag{19}$$

We need some measure to compare the performance of the models. The two most popular measure to evaluate the quality of the model is the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) given respectively by

$$AIC = 2\,\mathrm{Log}(\hat{\sigma}) + \frac{n+2k}{n} \tag{20}$$

$$BIC = \ 2\,\mathrm{Log}(\hat{\sigma}) - k\,\mathrm{Log}(n) \tag{21}$$

Where n is the number of observations and k is the number of parameters of the model. Hence AIC and BIC are different in the second term which itself is a penalty term for a model being too complicated. The likelihood function can be chosen as $\hat{\sigma}^2 = SSE/n$ where SSE stands for Sum of Squared Error.

We also need some parameter to determine the quality of model. Ljung-Box statistics can be used for this purpose.

$$Q(m) = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k} \tag{22}$$

The statistic follows χ^2 distribution. We can calculate the p-value of the statistics for the residue of the predictions. If the p values are large enough we accept the null hypothesis that there is further correlation among the residues.

F. Forecasting

One of the major motives to study the time series analysis is to make a forecast, essentially means predicting the future values for the time series. The $SARIMA$ model and its variants are also good for making the forecast. In this section, we study another systematic way to deal with the forecast which deals with both additive and multiplicative seasonality. The simplest forecasting method can be a simple exponential smoothing

$$\hat{x}_{n+1} = \alpha\, x_n \ + (1-\alpha)x_{n-1} \tag{23}$$

where $\hat{x}_{n+1}$ is the predicted value value given a time series data x_i, $i = 1, 2, \cdots, n$ and α is a weight parameter which is determined through the data. Higher the value of α higher is the emphasis on the recent data than the older data. This simple approach works when there is no trend and seasonality. If the data possess trend we can use Holt Winter with double exponential smoothing to make a prediction

$$\hat{x}_{n+h} = l_n + h\, t_n \tag{24}$$

If the data has seasonality along with trend we may use Holt Winter with triple exponential smoothing. We can classify the seasonality into two main categories, additive and multiplicative seasonality which are the cases with the data having constant variance and exponentially growing variance respectively. An exponentially growing variance transforms into a constant variance by taking the logarithm of the data. We use the following rule to forecast data having additive seasonality

$$\hat{x}_{n+h} = l_n + h\, t_n + s_{n+h-m} \tag{25}$$

And following is used for the data with multiplicative seasonality

$$\hat{x}_{n+h} = (l_n + h\, t_n)s_{n+h-m} \tag{26}$$

where level, trend and seasonal coefficients are given by

$$\begin{aligned} l_n &= \alpha\, x_n + (1-\alpha)(l_{n-1} + t_{n-1}) \\ t_n &= \beta(l_n - l_{n-1}) + (1-\beta)t_{n-1} \\ s_n &= \gamma(x_n - l_n) + (1-\gamma)s_{n-m} \end{aligned}$$

respectively and α, β, γ are calculated using the data itself. Higher the value of α, β, γ higher is the emphasis on the recent data than the older data.

III. Modeling the Crime Data

A. Selecting the Time Stamps

In the following, we will build a time-series model for the crime data. We used Python's statsmodels API to make calculations in this project. The detail of the code implementation can be found in [2]

The first decision we need to make is to select the unit of discrete-time change. The data is available for all the incidents with their date and time up to second. We look for the possibility of aggregating the data by months, weeks, and days. We studied the model performance based on monthly and weekly aggregate in detail and compared their performance. The monthly model is found to be performing better, for example in terms of BIC, and we will discuss this model further in this paper unless explicitly stated otherwise.

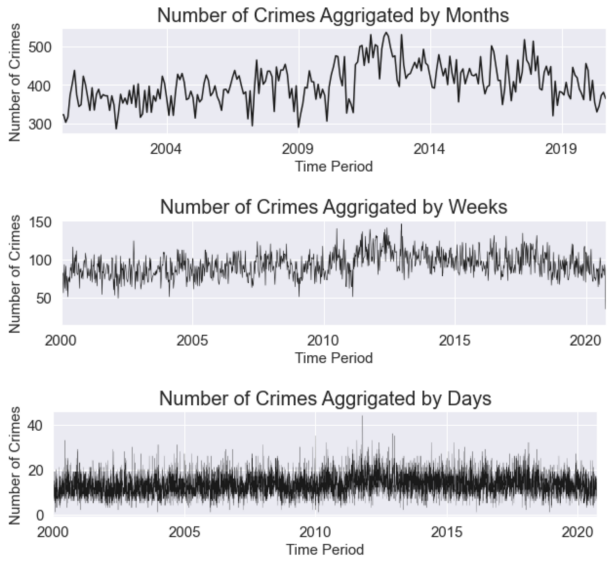

Fig. 6. Time-series of the crime data aggregated by months, weeks, and days. We can see some clear pattern on monthly and weekly data.

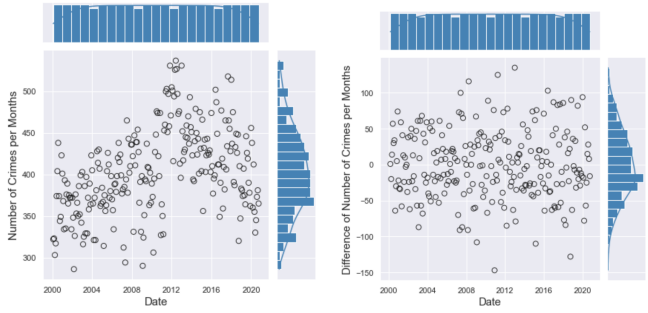

Fig. 7. Joint plot of monthly crime data. The data with trend (on the left) is transformed to data evidently without trend by first order difference operator (on the right)

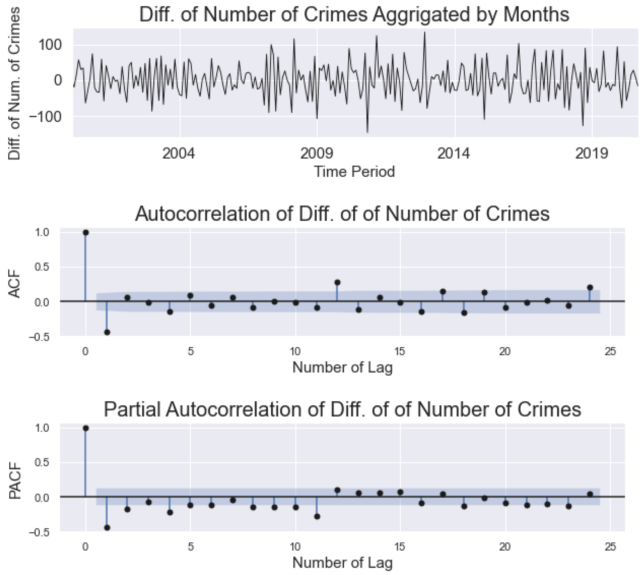

Fig. 8. (Upper) The time series of the difference of the monthly crime data. (Middle) Auto-correlation function suggesting that $q = 2$ as there is one significant lags. (Lower) Partial auto-correlation has significant value up to second lag suggesting $p = 2$. We can also see a significant ACF and PACF in 12^{th} lag indicating a annual seasonality of the crime data with $s = 12$.

B. Check for the Trend & Seasonality

Our data seems to be homoscedastic and no further transformation is needed regarding that. However, evidently, there are some trend and seasonality in the data and we need to remove those. A homoscedastic time series without any trend and seasonality is close to white noise. We take the first-order difference to remove the trend. The first-order difference seems to have performed a decent job. The differences data visually looks more like white noise than the original one. See FIG. 7. This suggests that the parameter value $d = 1$ in a $SARIMA$ model. We will discuss a more quantitative approach to evaluate our decision regarding these later. We will also come to seasonality later.

In the following, we determine the order of $AR(p)$ and $MA(q)$ processes. We calculate and plot the ACF and PACF for the difference of the monthly crime data in FIG. 8. As we discussed in the previous section the non zero lags of the ACF and PACF plot suggest to us the order of the $MA(q)$ and $AR(p)$ processes respectively. However, it is important to note that for the composite model like $ARIMA$ the suggestion for the order of the parameter by the ACF and PACF is just a guideline to follow. We need to check a few combinations of the parameters to find the optimum model.

From FIG. 8 it is seen that the time series of the difference of the monthly crime data has non zero auto-correlation function in first lag suggesting that $q = 2$. Also the partial auto-correlation has a significant value up to second lag suggesting $p = 2$. We can also see a significant ACF and PACF in 12^{th} lag indicating an annual seasonality of the crime data with the order of seasonality $s = 12$. The presence of seasonality means that the difference data is not a white noise unless we take a seasonal difference of order 12 on the top of it.

C. Model Selection & Evaluation

We studied the $ARIMA(p, d, q)$ model for both monthly and weekly data at first ignoring the seasonality. The optimal model for the weekly and monthly data was $ARIMA(1, 1, 2)$ and $ARIMA(0, 1, 1)$ respectively. We found that the performance of the monthly model is better than the weekly model based on the BIC criterion. The predicted trend was satisfactory to follow the observed trend in the data. However, the model was not satisfactory based on the p-value of the Ljung-Box statistics. Most of them were zero after a few lags suggesting that a further correlation among the residues [2]. This led us to consider the seasonal model.

We found that the optimal seasonal model for the monthly data is $SARIMA(0, 1, 1, 0, 1, 1, 12)$ by explicitly checking all possible combinations of the parameters under the constraint in equation (19). The prediction by the model is presented in

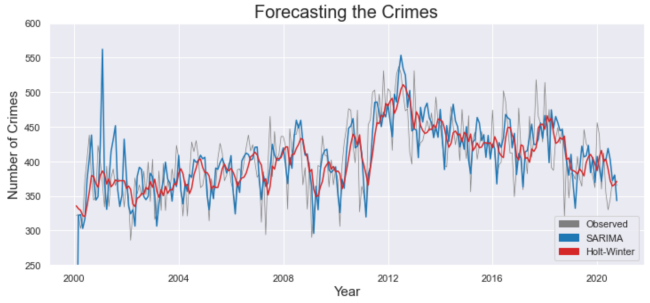

Fig. 9. Prediction by a seasonal model $SARIMA(0,1,1,0,1,1,12)$ and Holt-Winters triple smoothing along with the actual time-series.

FIG 9. Except for the overshooting of the peak on the far left the model has done a very good job capturing the ups and downs of the complicated time series data. The overshooting can be due to the boundary effect as our time series data do not go forever in the past.

Our data has additive seasonality. Hence we use Holt Winter triple exponential smoothing to make a forecast. The result of this method is presented in FIG. 9 along with the prediction by the $SARIMA$ model. From the plot we can see that the prediction by Holt-Winters is smoother than that of $SARIMA$.

So the model performance looks good visually. We can do some quantitative analysis of the performance of the model. We can define the residue of the data as $r_t = X_t - \hat{X}_t$ where $\hat{X}_t$ is the prediction of the X_t by a model. If a model is performing well we can assume the following about the residues: (a) The residues are normally distributed. (b) Time series of the residues are white noise. (c) There is no significant auto-correlation left between the residues. And (d) The p-value of the Ljung-Box statistic is large enough to reject the alternative hypothesis that there is further correlation left between the residues. Obviously, these assumptions are not independent but the different ways to confirm the same thing. We can see the model performance based on all four criteria in FIG. 10.

Overall the model performance is great. There is still a small amount of annual seasonality present in the residues. But the value of the ACF of residue in the 12^{th} lag from $SARIMA$ model is significantly smaller than that of the $ARIMA$ model. Also the p-value of the Ljung-Box statistics tends to be smaller in the 12^{th} lag indicating the same effect as in the ACF plot. Also in the $ARIMA$ model, there is equally significant ACF in 24^{th} lag as well. However, in the $SARIMA$ model ACF in the 24^{th} lag is diminished.

IV. Discussion & Conclusion

In this work, we studied the historical crime data from the town of Carry North Carolina USA spanning from 2000 to 2020 CE. We investigated the seasonality and trend in the data. We found that the $ARIMA$ model is not sufficient to describe the data as found strong evidence of the annual seasonality by analyzing residue of the prediction of the $ARIMA$ model.

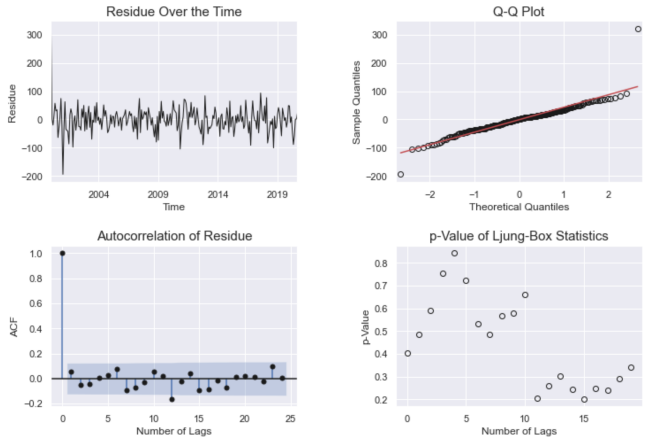

Fig. 10. (Upper Left) Residue of the prediction by the $SARIMA(0,1,1,0,1,1,12)$ model over the time. (Upper Right) Quantile-Quantile plot of the residue supporting the normality of the distribution of the residue except few outliers. (Lower Left) ACF of the residue showing the residues are almost like a white noise if we ignore a marginally significant 12^{th} lag. (Lower Right) p-value of Ljung-Box statistics also indicating no significant correlation on the residue.

Also, the seasonal model performed better in terms of BIC. With the help of the analysis of the residue by the prediction of the $SARIMA$ model, we conclude that the $SARIMA$ model with annual seasonality is a good fit for the data.

There is some room for improving this work. First, a bigger data set can be studied covering a larger geographic area and longer historical period. A cross-comparison between one region to another can be done. Like we discussed earlier there is a small residual seasonality. A further investigation can be done to remove that seasonality. It would be interesting to see the relation of crime with other historical data related to physical, psychological, social, economic area. For instance, it would be interesting to see the relation of crime with the stock market or weather. Also, there is considerable progress in analyzing time series data with the deep learning-based model especially using LSTM. This could be a natural extension of the present work.

Acknowledgment

The author likes to thank the town of Carry for providing the data set to the public.

References

[1] https://data.townofcary.org/explore/dataset/cpd-incidents.
[2] https://github.com/roshankoirala/Time_Series_Analysis-.
[3] https://www.ncsbi.gov/Services/SBI-Uniform-Crime-Reports.aspx#: :text=Part%201%20offenses%2C%20excluding%20negligent,reported%20for%20Part%202%20offenses
[4] Chamlin, M.B. Crime and arrests: An autoregressive integrated moving average (ARIMA) approach. J Quant Criminol 4, 247–258 (1988). https://doi.org/10.1007/BF01072452
[5] Landau S. F., Friedman D. The Seasonality of Violent Crime: The Case of Robbery and Homicide in Israel. Journal of Research in Crime and Delinquency. 1993;30(2):163-191. doi:10.1177/0022427893030002003
[6] Corman, Hope, and H. Naci Mocan. 2000. "A Time-Series Analysis of Crime, Deterrence, and Drug Abuse in New York City." American Economic Review, 90 (3): 584-604. DOI: 10.1257/aer.90.3.584

[7] Greenberg, D.F. Time Series Analysis of Crime Rates. Journal of Quantitative Criminology 17, 291–327 (2001). https://doi.org/10.1023/A:1012507119569
[8] P. Chen, H. Yuan and X. Shu, "Forecasting Crime Using the ARIMA Model," 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Shandong, 2008, pp. 627-630, doi: 10.1109/FSKD.2008.222.
[9] Wang, B., Yin, P., Bertozzi, A.L. et al. Deep Learning for Real-Time Crime Forecasting and Its Ternarization. Chin. Ann. Math. Ser. B 40, 949–966 (2019). https://doi.org/10.1007/s11401-019-0168-y
[10] Stec, Alexander, and Diego Klabjan. "Forecasting crime with deep learning." arXiv preprint arXiv:1806.01486 (2018).
[11] R. H. Shumway, D. S. Stoffer, "Time Series Analysis and Its Applications, Fourth Edition," Springer, 2016.