Crime Prediction Utilizing ARIMA Model

Yuting Lu*
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA 61820
* Corresponding author: yutingl7@illinois.edu

Abstract. With the development of society and the progress of human beings, the issue of crime has attracted more attention. This research uses one of the most common time series models named the ARIMA model as the main algorithm to predict the crime data by month in San Francisco from 2003 to 2015, obtained from Kaggle. This research is about to study all categories of crime and specific crime cases as well. The focus is on the number of cases that happened every month. By applying the ARIMA model, it can be tested whether the trend of happened crime cases fits the model and predict the future crime tendency. The verification and prediction of crime data are studied with the help of the ARIMA model. And the goal of studying this dataset is to improve the safety and security of human beings. In this research, the ARIMA model simulates the previous crime trend effectively and forecasts the future crime trend with high accuracy. This is helpful to improve the safety level and reduce the crime rates in San Francisco.

Keywords: Crime data, Time series analysis, ARIMA model, Prediction.

1. Introduction

Crime has been a tough issue for a long time. It is urgent to reduce the crime rates and accommodate rapidly when real criminal events happen. There is a universal belief that crimes like murder, rape, and stealing are regarded as illegal [1]. Therefore, some methods have been proposed to predict crime data, aiming to improve the severe situation of crime in current society.

Data Mining and machine learning methods have been widely used in the crime prediction field [2]. Classification is the most common and essential method of data mining [3]. It has been used to predict graduate students' performance, diabetes, and even heart disease [4-6]. Therefore, some researchers also utilized classification to predict crime and get the result of about 70% accuracy, which is lower compared to the accuracy of 84% accuracy of the decision tree method [4]. This suggests that applying classification to crime prediction has a large margin of error.

Therefore, some researchers did a study on predicting crime using machine learning methods such as linear regression, Additive Regression, and Decision Stumps. A method such as linear regression performs well in predicting crime data as it has the highest correlation coefficient and lowest errors in every category of crime analyzed [7]. While method such as decision stump is not effective since it has the lowest correlation coefficient and the highest error of over 40% [7]. This suggests that although machine learning methods have been widely used and are assumed better at predicting crime data, methods dealing with big data such as the decision tree can still be invalid.

On the other hand, some scientists claim that the time series analysis may be more effective to predict crime data [4]. For several decades, researchers and criminologists have been devoted to predicting crime data and reducing crime rates. And these years, crime data with a timeline attracts attention in this field, and thus the time series analysis is put forward. Among all the time series analysis methods, the ARIMA model is utilized because of its high accuracy in predicting crime data [8-9]. The ARIMA model is fitted to time series data in order to clarify the data or forecast upcoming series points [10]. More specifically, in some situations when data indicate that the mean function is non-stationary, the ARIMA models are used after performing an initial differencing step one or more times to remove the non-stationarity of the mean function [10]. Therefore, in this research, the ARIMA model will be built to test the accuracy and effectiveness of predicting the crime data. With the crime data in San Francisco from January 2003 to May 2015, the ARIMA model will be utilized to fit the data. Additionally, the ARIMA model can also forecast the future trend of crime. The focus of this research is to test the effectiveness of forecasting future crime tendencies using the ARIMA model with previous crime data. Moreover, with the crime data in a timeline, it is easier for police and the government to adopt measures to reduce crime rates and even prevent the occurrence of certain crimes.

2. Methodology

This research uses a time series model named the ARIMA model to analyze the crime data prediction in San Francisco from 2003 to 2015. By applying the following model, we will fit the data into the model to see if this model is reliable.

2.1 Autoregressive Moving Average Model (ARMA Model)

The autoregressive moving average model is an important model to study time series. It combines two models: the autoregressive model (AR model) and the moving average model (MA model). The autoregressive model describes the relationship between current values and historical values. And it can be defined by the following expression:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{1}$$

While the moving average model describes the accumulation of errors in the autoregressive part, and it can be defined as follows:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

2.2 Autoregressive Integrated Moving Average Model (ARIMA Model)

ARIMA model means autoregressive integrated moving average model. It is a general form of the autoregressive moving average model. Both of these models are utilized in predicting data, especially data involving time. In time series analysis, the ARIMA model is used to deal with data that shows no stationarity, which means the dataset is not stable during the given timeline. In the model ARIMA, there are three parameters named p, d, and q. AR in the model means autoregression, and p represents the number of autoregressive items. MA means moving average, and q represents the number of moving average items. Although d is not presented in the name of the ARIMA model, it means differential and represents the number of differences needs to make it a stationary series. The definition of the ARIMA model can be expressed by the formula below:

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t \tag{3}$$

where L is lag operator, $d \in Z, d > 0$.

2.3 White Noise

The term white noise is also often used to denote a spatially noisy signal whose autocorrelation in the correlation space is 0, so the signal is "white" in the spatial frequency domain, and the same is true for signals in the angular frequency domain, such as a signal that radiates to all angles in the night sky.

2.4 Hypothesis Test

The hypothesis test is a statistical method used to certify whether a statement can be supported by the dataset or not. The steps of performing a hypothesis test are as follows: The first step is to set up a null hypothesis that needs to be tested, and then set up the one-sided or two-sided alternative hypothesis corresponding to the null hypothesis. Next is to perform an appropriate distribution (such as Normal distribution and Chi-squared distribution) and get the value of the p-value. After that, compare the p-value obtained with a significant level (often 0.05) and determine if the null hypothesis should be rejected. And finally, make a conclusion about the result and decide whether there is adequate evidence to accept the null hypothesis. The null hypothesis mentioned above is often with zero difference. For example, when testing whether the school bus is late is due to chance or not, the hypothesis test should be set up as the school bus is late is not due to chance (there is no big difference between the on-time bus and late bus). The alternative hypothesis can be set up from two aspects. The first one is two-sided, which means the alternative hypothesis is that a late bus is not due to chance, and it has a big difference between punctual buses. The p-value is the probability of a sample observation or more extreme results obtained when the null hypothesis is true. If the p-value is small, the probability of extreme observations occurring under the null hypothesis is small.

3. Results and discussion

3.1 Data Analysis

The dataset obtained from Kaggle to conduct this research is a free public resource. It contains the crime data in San Francisco from January 2003 to May 2015, including features of date, district, category, and so on. In this research, date and category are the main variables to be studied. The main goal of this dataset is to figure out the number of all crimes, theft, assault, and robbery during this period. The first step is to make out data visualizations of all the above crime cases by month. And then fit them into the ARIMA model, however, it is important to figure out the best parameter (p, d, q) and obtain the best model for each of the features. Finally, the future trend of each crime category can be shown. Table 1 generates the dataset analyzed in this research. It shows the total number of crime cases reported from January 2003 to May 2015.

Table 1. Margins and print area specifications

	Theft	Assault	Robbery	All
Report Total	174900	76876	23000	878049

3.2 Visualization Results

The dataset used in this research is obtained from Kaggle. It contains the crime data from January 2003 to May 2015. 39 categories of crime are included in this dataset, and for this research, only 3 striking crime categories have been chosen for the striking number of cases. The chosen crime types are larceny/theft, assault, and robbery. Since the dataset contains the criminal cases with the actual happening time, crime occurrence by month will be analyzed to see if it fits the ARIMA model. The following figures are performed to visualize the crime cases per month:

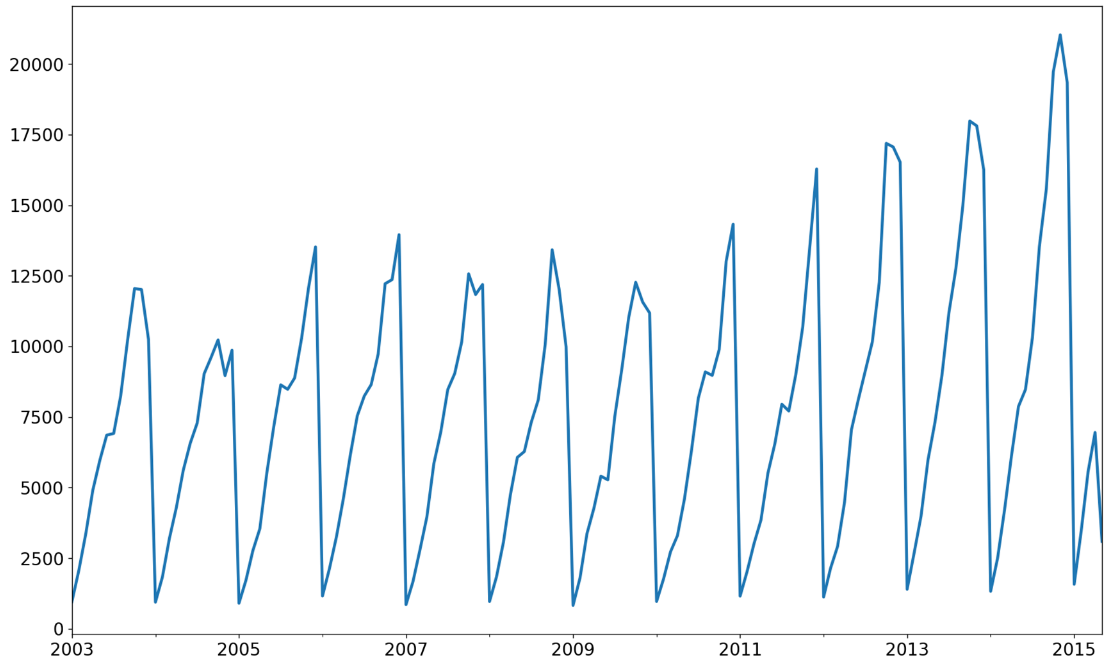

Figure 1. The Number of Thefts Every Month

Figure 1 shows the change in the number of theft cases that happened every month from 2003 to 2015. It can be seen from the figure that at the beginning of each year, the theft cases are the least. And in the middle of every year, there are most theft cases that happen. However, the overall trend of theft cases is gradually increasing, with the highest number of cases in 2013 and 2014. This suggests that theft in San Francisco had been worse from 2003 to 2015, with the number of theft cases from 12,500 to 20,000 in the same period over 10 years later.

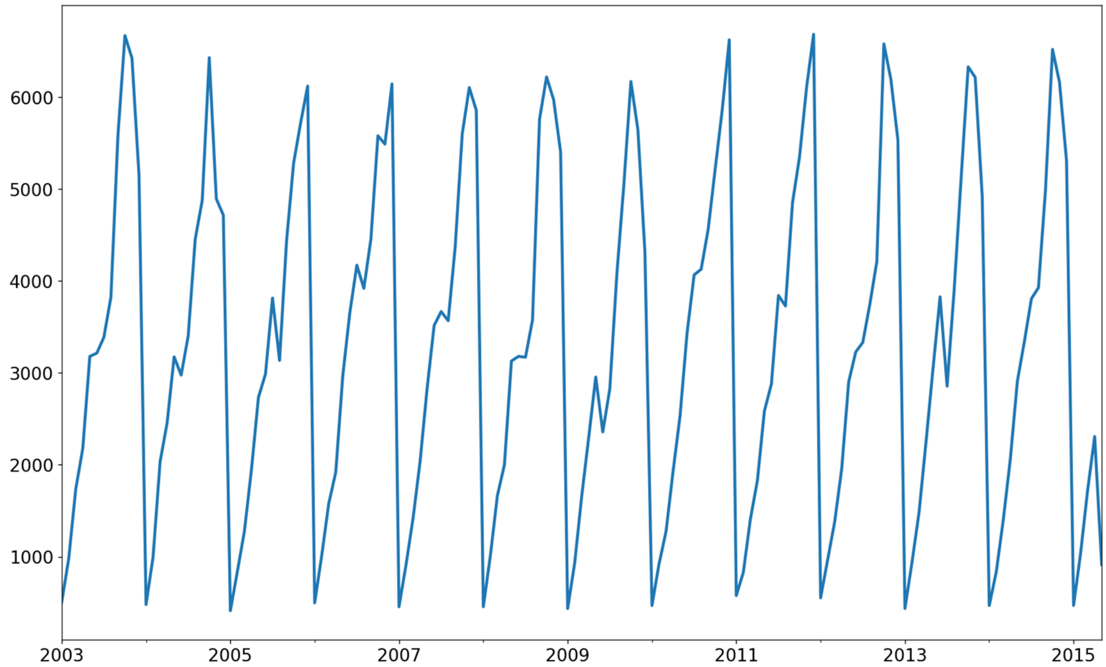

Figure 2. The Number of Assaults Every Month

The above figure 2 shows the number of assaults every month from 2003 to 2015. Similar to the situation the theft, San Francisco has the most severe assaults in the middle of every year with the number of reported cases over 6,000. And San Francisco has the least number of assault cases reported at the beginning of every year with less than 1,000 cases reported. Nevertheless, the trend of assault cases is not similar to that of theft cases. It can be seen from figure 2 that the peak of every year is stable from 2003 to 2015, with all of these years having over 6,000 assault cases reported. It suggests that the change of assault reported in San Francisco is not significant.

Figure 3 below shows the number of reported cases of robbery from 2003 to 2015. But unlike the previous two crime categories, for this category of crime, the number of reported cases reaches its peak at the end of each year, but its lowest point is still at the beginning of the year, annually. The trend of the reported robbery cases is stable in general, with slightly more cases reported in 2007 and 2012.

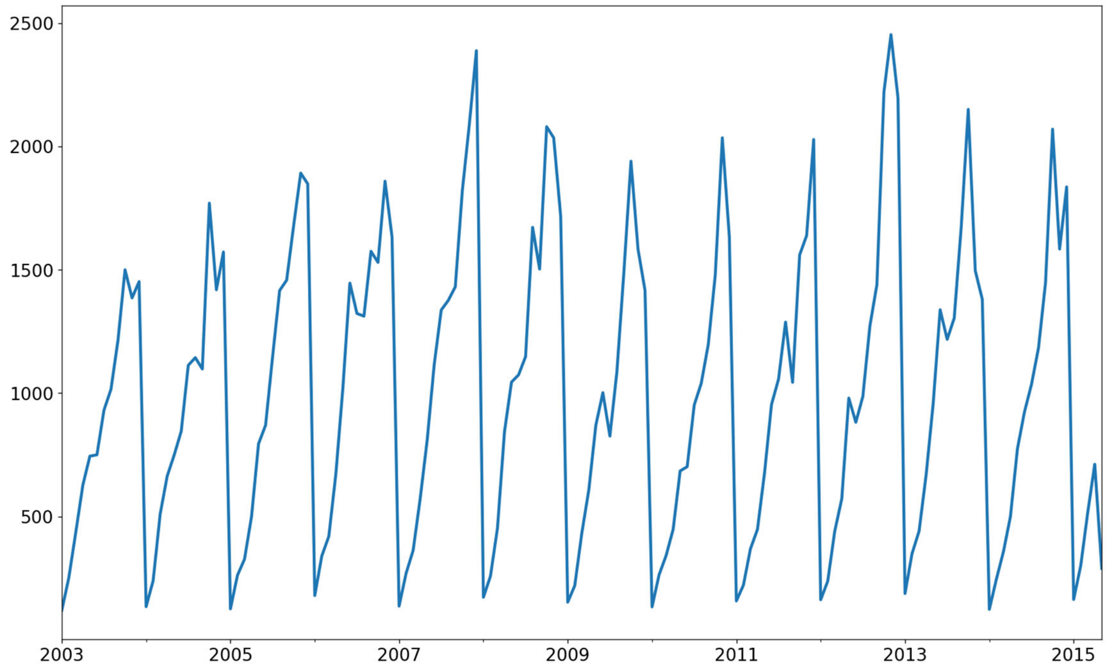

Figure 3. The Number of Robberies Every Month

The fourth figure above shows the trend of all categories of crime cases reported monthly every year from 2003 to 2015. This figure indicates that the middle of the year has the highest rate of crime, and the beginning of the year has the lowest rate of crime. In addition, from 2003 to 2015, there are around 70,000 crime cases reported annually, which suggests that safety in San Francisco has been an issue for over 12 years without improvement.

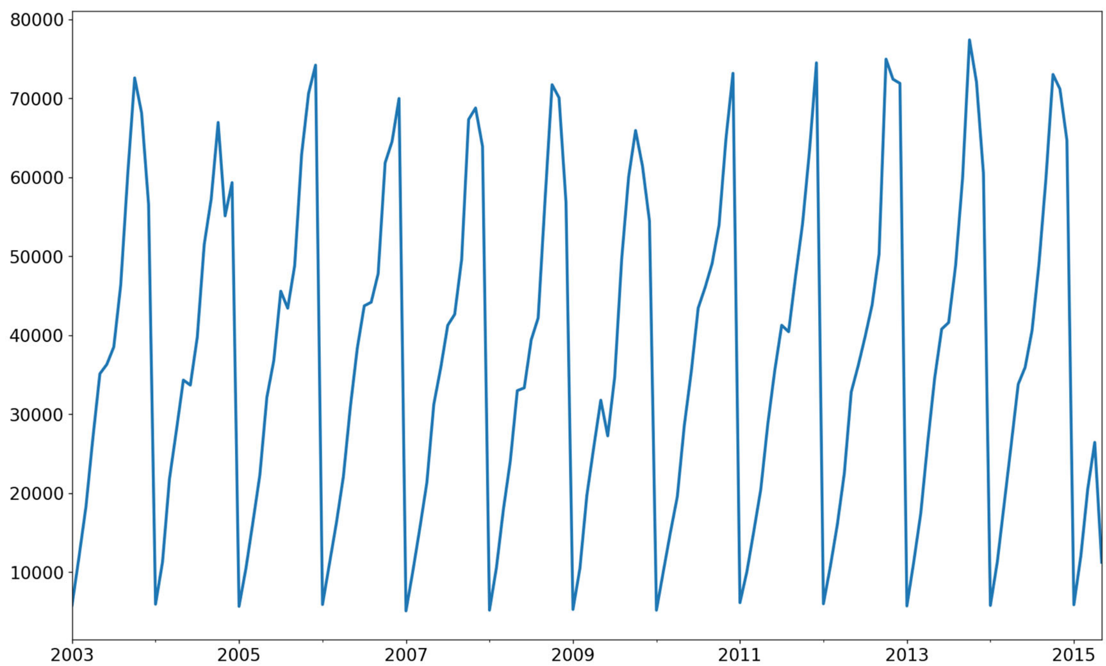

Figure 4. The Number of All Crimes Every Month

Overall, it can be seen from the above 4 figures that all of them are not stationary from year to year, and thus ARIMA model is more appropriate here. The following results are the best ARIMA model for each category of crime.

Below are the best ARIMA models for thefts, assaults, robberies, and all crime categories. Table 2 shows the ARIMA model for thefts. It can be seen from the table that the parameters of this model have p=1, d=4, and q=4, which is ARIMA (1, 4, 4). In addition, the column "P>|z|" in the following tables means the p-value of this log operator. For thefts, it can be seen that except for the first and fourth parameters, others have p-values greater than 0.05, which is usually the significant level of the null hypothesis. Therefore, for the ARIMA model for thefts, only AR(1) and MA(4) are significant.

Table 2. Best ARIMA Model for Thefts

	coef	std err	z	P>|z|	[0.025	0.975]
AR(1)	-1.0000	0.008	-127.017	0.000	-1.015	-0.985
MA(1)	-1.9910	6.086	-0.327	0.744	-13.920	9.938
MA(2)	0.0009	23.475	4e-05	1.000	-46.008	46.010
MA(3)	1.9897	6.210	0.320	0.749	-10.182	14.162
MA(4)	-0.9997	0.122	-8.193	0.000	-1.239	-0.761
σ^2	1.524*107	1.52*10-6	1*1013	0.000	1.52*107	1.52*107

Similarly, table 3 shows the best ARIMA model for assaults. By looking at the table below, only AR(1) and MA(4) have p-values less than 0.05, which means they are significant in this ARIMA model with the parameters of (3, 1, 1).

Table 3. Best ARIMA Model for Assaults

	coef	std err	z	P>|z|	[0.025	0.975]
AR(1)	0.7279	0.121	6.004	0.000	0.490	0.965
AR(2)	-0.0759	0.211	-0.359	0.719	-0.490	0.338
AR(3)	-0.2679	0.158	-1.692	0.091	-0.578	0.042
MA(1)	-1.0000	0.142	-7.047	0.000	-1.278	-0.722
σ^2	1.872*106	7.58*10-8	2.47*1013	0.000	1.87*106	1.87*106

Table 4 shows the results of the ARIMA model for robberies. By looking at the column “P>|z|”, it can be seen that AR(1) and MA(1) have p-values less than 0.05, suggesting this model with parameters (5, 1, 1) has only two significant coefficients.

Table 4. Best ARIMA Model for Robberies

	coef	std err	z	P>|z|	[0.025	0.975]
AR(1)	0.6621	0.111	5.988	0.000	0.445	0.879
AR(2)	-0.0699	0.139	-0.505	0.614	-0.341	0.202
AR(3)	-0.2445	0.149	-1.646	0.100	-0.536	0.047
AR(4)	0.0792	0.175	0.453	0.651	-0.264	0.422
AR(5)	-0.2012	0.126	-1.598	0.110	-0.448	0.046
MA(1)	-1.0000	0.133	-7.538	0.000	-1.260	-0.740
σ^2	1.897*105	7*10-7	2.71*1011	0.000	1.9*105	1.9*105

Table 5 shows the results of the best ARIMA model for all crime categories. Again, only AR(1) and MA(1) have p-values less than 0.05. This model has the same parameter as the model for robberies in table 4, which is (5, 1, 1).

Table 5. Best ARIMA Model for All Crime Categories

	coef	std err	z	P>|z|	[0.025	0.975]
AR(1)	0.6781	0.174	3.898	0.000	0.337	1.019
AR(2)	-0.1136	0.307	-0.370	0.711	-0.715	0.487
AR(3)	-0.2039	0.294	-0.694	0.488	-0.780	0.372
AR(4)	0.0212	0.310	0.068	0.945	-0.586	0.628
AR(5)	-0.1700	0.260	-0.653	0.514	-0.680	0.340

MA(1)	-0.9925	0.133	-7.455	0.000	-1.253	-0.732
σ^2	3.298*108	3.98*10-10	8.29*1017	0.000	3.3*108	3.3*108

In conclusion, the above tables show the best ARIMA models for theft, assault, robbery, and all categories, respectively. By looking at the model, the parameters are different for different crime types. For example, the best ARIMA model for theft has the parameter (1, 4, 4), and the best ARIMA model for assault has the parameter (3, 1, 1). In addition, the column “P>|z|” means the p-value of this log operator. It can be seen that nearly not all parameters have p-values less than 0.05, which is usually the significant level of the null hypothesis. To specify, the null hypothesis is the coefficient is significant. Since some of the p-values here are greater than 0.05, and by the definition of the hypothesis test, the null hypothesis should be rejected. And the conclusion is that there is no adequate evidence to prove the significance of all coefficients, instead, only certain coefficients with p-values less than 0.05 are significant.

Utilizing the above information, it is easy to fit the previous crime cases per month and predict the future trend of the crime situation in San Francisco.

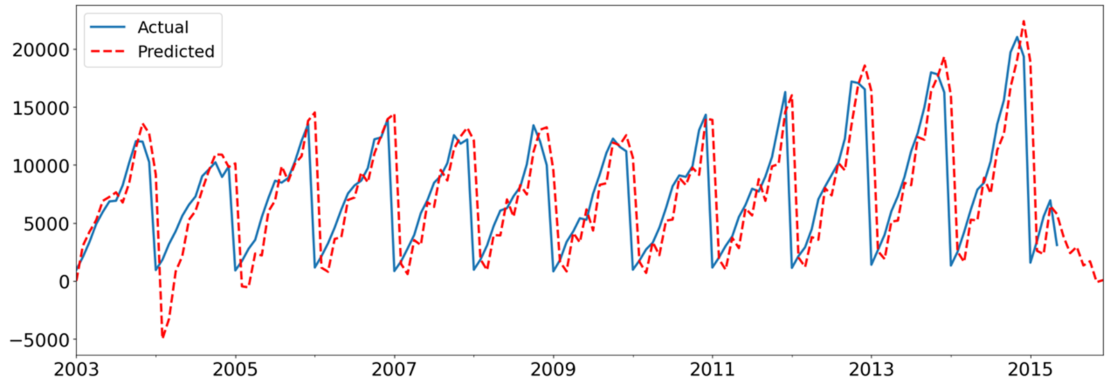

Figure 5. Prediction of Theft Cases

Figure 5 is the analysis and prediction of theft cases. The blue solid line in figure 5 is the same as that in figure 1, which is the actual cases reported. And the red dotted line in figure 5 is the prediction of reported theft crime utilizing the ARIMA model. It is obvious that the overall shape of the dotted line is similar to that of the solid line. Moreover, the peaks and low values of the solid line and dotted line almost overlap each other, which indicates the high accuracy of utilizing the ARIMA model to predict theft crime data in this research.

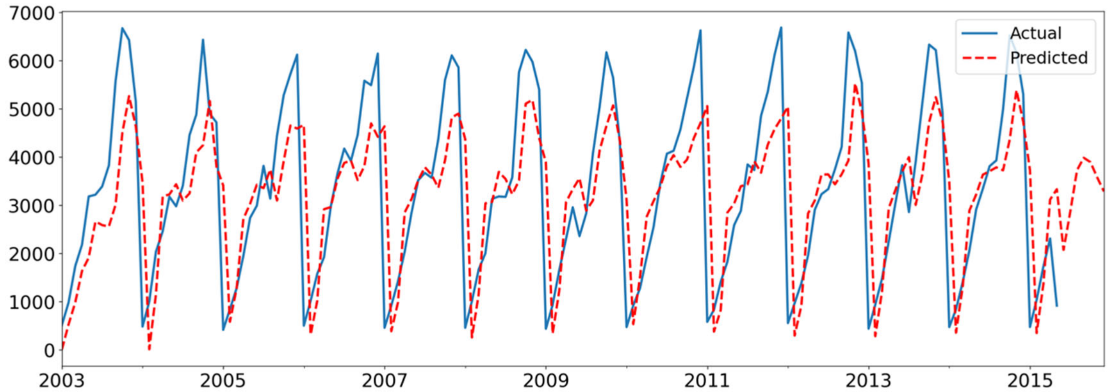

Figure 6. Prediction of Assault Cases

The figure 6 shows the analysis and prediction of monthly assault cases from 2003 to 2015. Likewise, the solid line represents the actual assault data, and the dotted line represents the prediction using the ARIMA model. Comparing to the figure 5, the similarity of figure 6 and figure 5 is the shape of the prediction and the actual crime data, proving the effectiveness of the ARIMA model when predicting crime data. However, the difference is the peaks. It can be seen from the figure 6 that the predicted annual peaks of assault cases are all lower than the actual annual peaks, which suggests the prediction may have slight errors when predicting the extreme values.

Figure 7 below and figure 6 share the exact same characteristics. Similar shape between actual data and predicted data suggests that the ARIMA model is practical, however, the residuals between the annual peaks indicates the accuracy may not high.

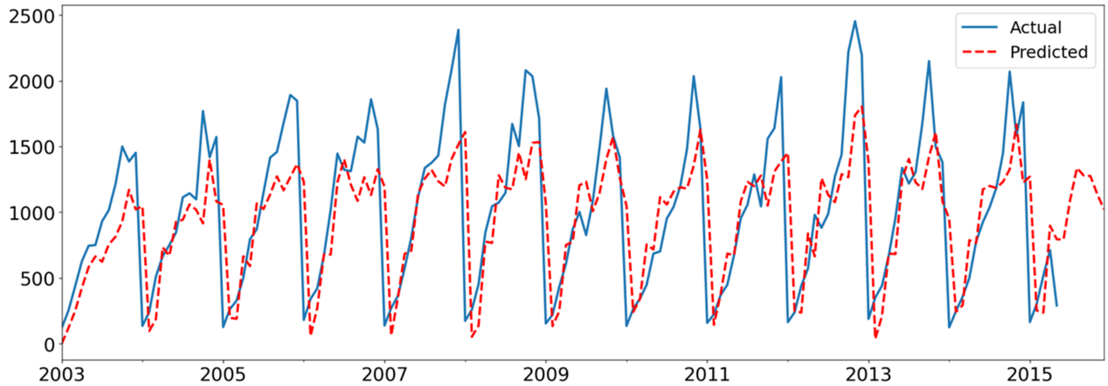

Figure 7. Prediction of Robbery Cases

It is the same situation when analyzing all crime categories together. The blue solid line in figure 8 below are the actual crime cases that happened in San Francisco at that time, while all the red dotted line is the prediction of the number of crime cases in San Francisco during the corresponding period. In figure 8, though the shape of the two lines is similar, the peaks of the predicted line are clearly lower than those of the actual line. This may indicate errors in this model.

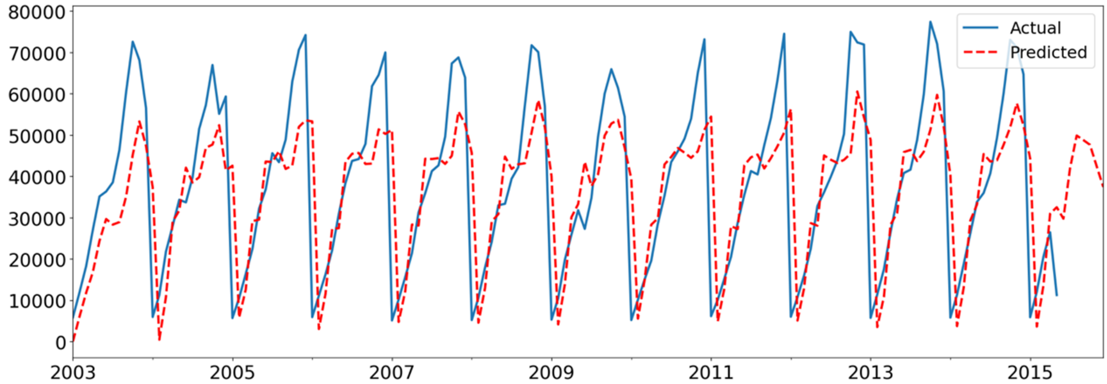

Figure 8: Prediction of All Crime Categories

To conclude, all the 4 figures above point out that the shape of the actual number of cases and the shape of the predicted number of cases are similar, especially the peak and lower points. The dataset provides the data from January 2003 to May 2015, but the red dotted line representing the prediction by the ARIMA model continues to December 2015. The similar shape of each figure indicates the accuracy of the ARIMA model, and thus the prediction is convincing. It suggests that the ARIMA model can be utilized to predict future crime data. One of the possible explanations may be that when combining all crime categories together, there are some extreme values affecting the prediction level. Overall, the ARIMA model performs well when fitting the previous crime data and predicting future crime data.

4. Conclusion

From the above results, it is obvious that the ARIMA model has high accuracy and effectiveness when predicting crime data. It gives prediction in the figures, which helps to obtain more accurate crime data prediction. On the other hand, sometimes the model may not fit the previous data perfectly, which indicates that there are also some errors or residuals when fitting the ARIMA model in time series analysis. Furthermore, time series analysis has been widely used in life at current. In this research, one of the most common time series methods has been studied. Nevertheless, there are other models such as the ARMA model that can also be used to analyze datasets with timelines. Applications of time series analysis models make our life better and easier. When utilizing the ARIMA model, drawbacks and errors should also be considered at the same time. By applying the time series model in this research, people benefit from it by learning the situation of the crime and then safety condition when choosing where to settle down. In the meanwhile, approaches to improving the security of those areas are encouraged as well to help protect people living in there.

References

[1] Crime. https://en.wikipedia.org/wiki/Crime

[2] Kim, S., Joshi, P., Kalsi, P. and Taheri, P, “Crime Analysis Through Machine Learning,” IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp:415-420 (2018).

[3] Statistical Classification. https://en.wikipedia.org/wiki/Statistical_classification

[4] Bunkar, K., Singh, U. K., Pandyam, B. and Bunkar, R, “Data mining: Prediction for performance improvement of graduate students using classification,” 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), pp:1-5 (2012).

[5] Sisodia, D. and Sisodia, D.S., “Prediction of Diabetes using Classification Algorithms,” Procedia Computer Science, vol. 132, pp:1578-1585 (2018).

[6] Masethe, H. D. and Mosima, A. M., "Prediction of heart disease using classification algorithms," Proceedings of the world Congress on Engineering and computer Science, vol. 2, pp:25-29 (2014).

[7] McClendon, L. and Meghanathan, N., “Using Machine Learning Algorithms to Analyze Crime Data,” Machine Learning and Applications: An International Journal, vol. 2, pp:1-12 (2015).

[8] Islam, K. and Akhter, R., "Forecasting crime using ARIMA model," arXiv preprint arXiv, (2020).

[9] Vijayarani, S., E. Suganya, and C. Navya, "Crime analysis and prediction using enhanced Arima model," International Journal of Research Publication and Reviews, vol. 2, pp:257-266, 2021.

[10] Autoregressive Integrated Moving Average. https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average