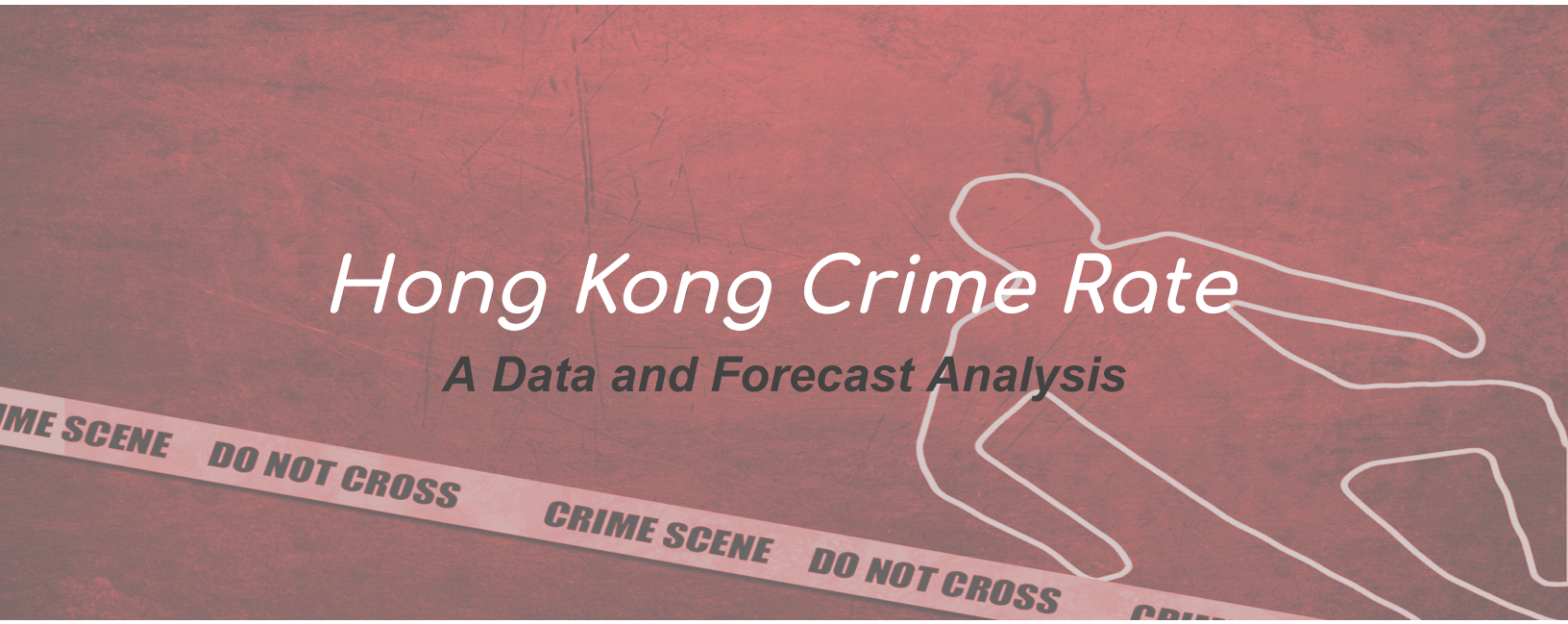


Hong Kong Crime Rate

A Data and Forecast Analysis

Group 13

Junoes Hans Nathanael 1155147304
Lui Lam Wai 1155142424
Fu Tsz Ho 1155176919
Chan Chiu Sing 1155176932
Ko Ming 1155177100
Chen Man Hin 1155177307

Table of Contents

1. Introduction

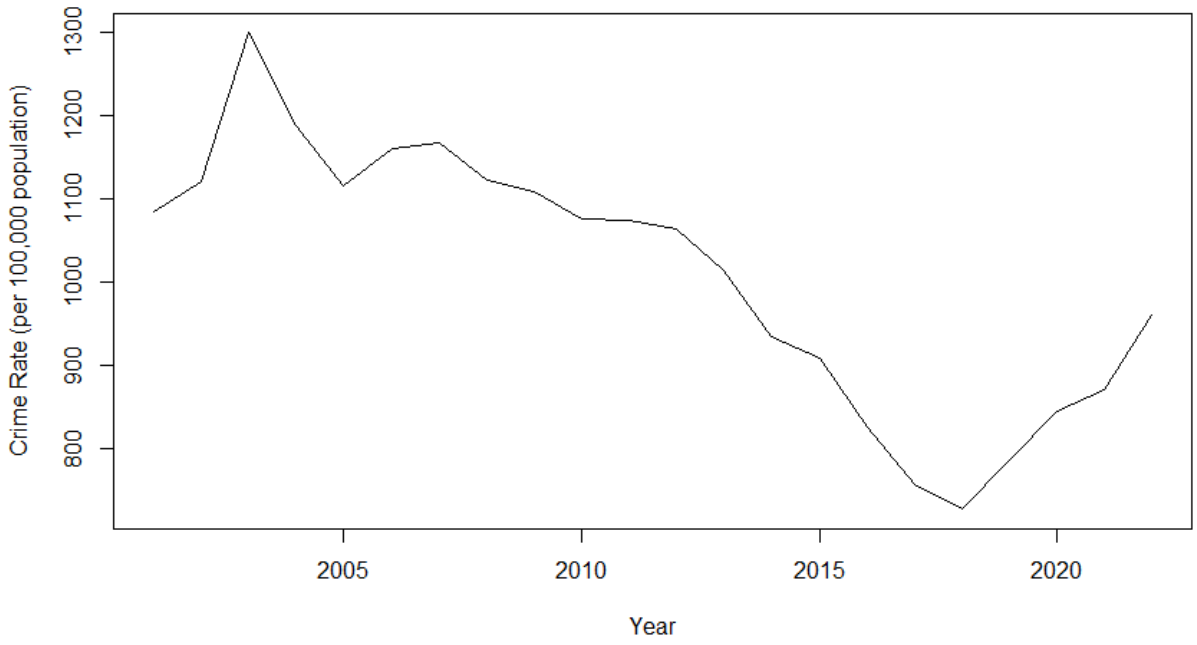
Rank	Country	Crime Index
1	Venezuela	83.6
12	Brazil	62.0
57	United States	48.2
115	China	29.4
119	Singapore	27.6
136	Hong Kong	21.9

Hong Kong is actually among the safest country in the world, ranking 7th globally for countries with sufficient data, even lower than Singapore and China. Despite this, however, one may notice the seemingly increase of crime rate as of late. This trend may be realized by residents of Hong Kong. This is one of the reasons why Group 13 has chosen to explore this topic. Viable data from the Hong Kong police force may be easily accessed, and also potentially a point of interest for those of us that are born and raised in Hong Kong, so that we can relate to what is happening in the country. The role of government and other factors in contributing to a relatively safe environment merits discussion about what can be learnt from the Hong Kong experience. Cultural factors such as utilitarian familism, Confucianism and extended kinship structures are often cited as contributing factors to the low crime rates in HK. Although boasting a relatively low crime rate, that is not to say that one should not worry about the increasing trend.

We have decided to follow the standard protocol of 3 forms of smoothing and more advanced time series modeling. The 3 forms of smoothing are Moving Average (MA), Centered Moving Average (CMA), and Exponential Smoothing. In part 4, we also used ARIMA modeling. The results of each forecast and the corresponding MSE and MAE are compared in the conclusion.

1.1 Data Collection

The data on crime rates in Hong Kong from 2001-2022 is retrieved from the website of the Hong Kong Police Force. A plot of time series is shown in figure 1 where the x-axis denotes the year and the y-axis denotes the crime rate per 100,000 population. There is a general decreasing trend but an increase in recent years.



(Figure 1: Time series on hk crime rates)

1.2 Data Preprocessing

The z-score of each data point is calculated as follows:

$$z = \frac{xi - \bar{x}}{s}$$

where $\bar{x}$ is the mean of the crime rate data, s is the standard deviation of the crime rate data.

There is no data point that has a z-score outside the range of ±3. Therefore, no outliers are removed.

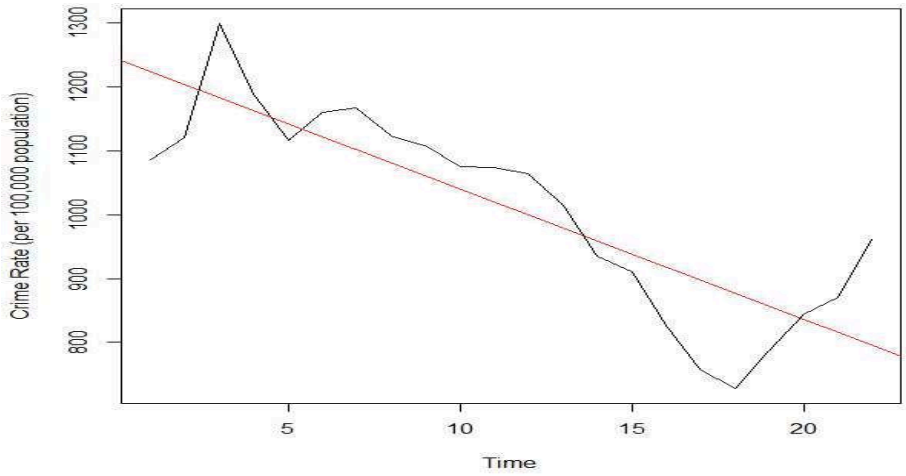
2 Trend Projection

The formula for the trend projection is:

$$T_t = b_0 + b_1 t$$

The intercept and slope of the line, representing the coefficients (b0, b1) , is calculated using the least square method, which is given by (1244.86, -20.42654). The mean absolute error (hereinafter called the "MAE") and mean squared error (hereinafter called the "MSE"), which measure the

forecast accuracy of the trend line, is computed. The results can be seen in the figure 2:



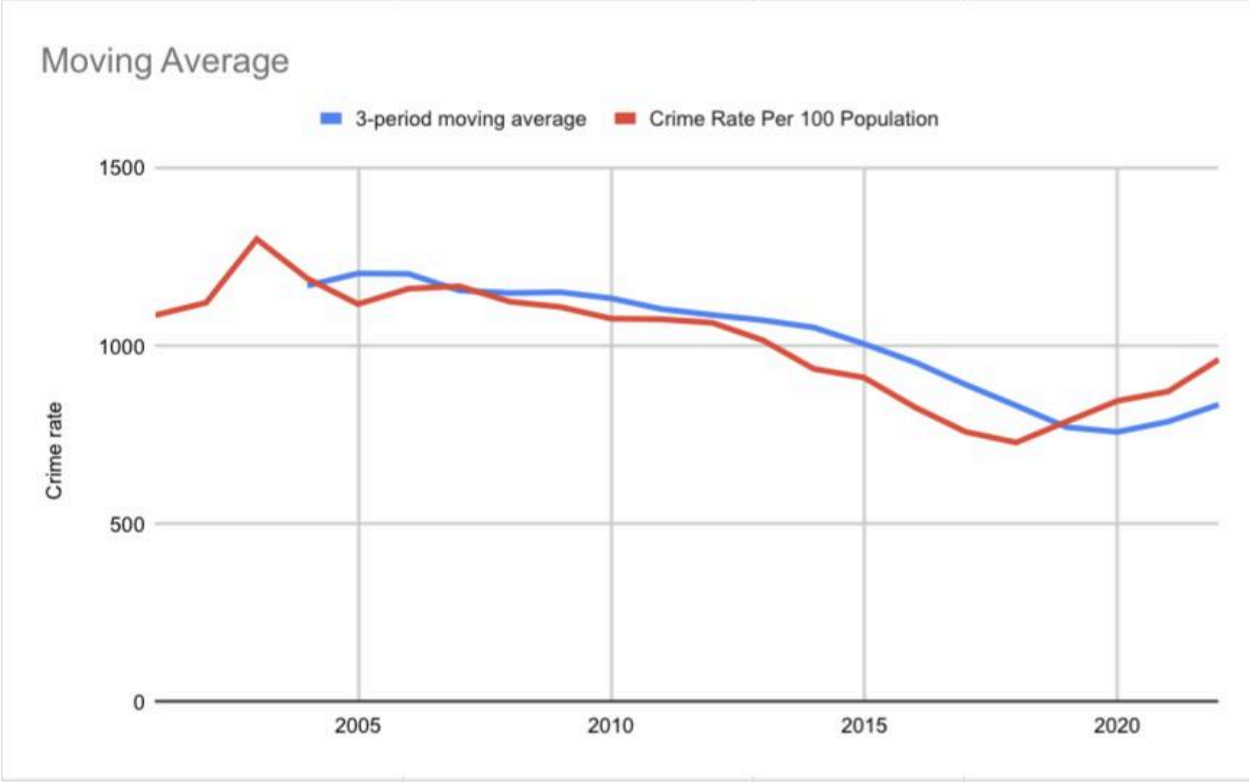
(Figure 2: The trend projection on the hk crime rates)

The performance of the trend line is acceptable since the forecast error,measured by MSE and MAE are 6676.635 and 68.24527 respectively.

3 Smoothing

3.1 Moving Average

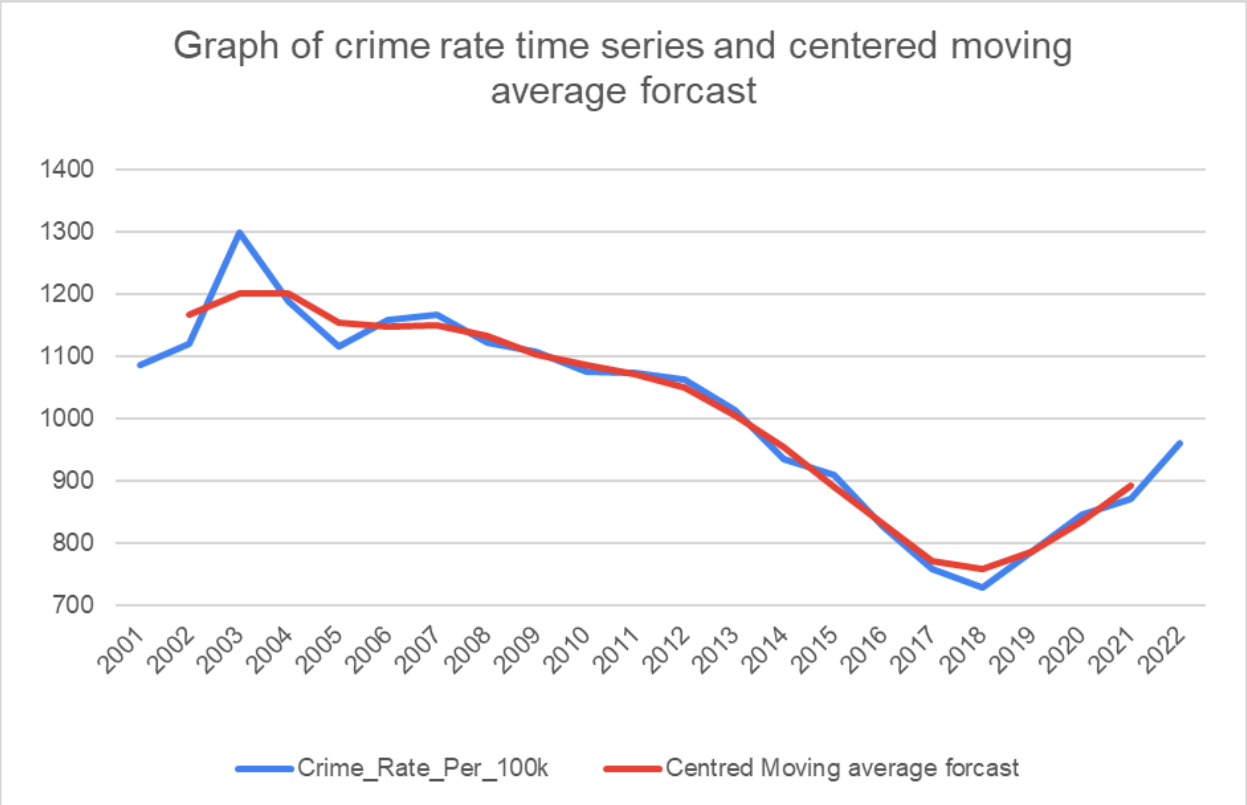
A simple moving average (SMA), is calculated by taking the arithmetic mean of a given set of values over a specified period. A set of numbers, or prices of stocks, are added together and then divided by the number of prices in the set. In this case we choose 3-Period to forecast the crime rate. For example, the sum of the crime rate from 2001 to 2003 is 3505, so we predict that in 2004 the crime rate will be 3505/3=1168.333. The squared forecast error is 406.6957886. The others will shown below :



The total MSE is 14346.44 , and the forecast of crime rate per 100 population for 2023 would be 892.333 .

3.2 Center Moving Average

The centered moving average method consists of computing an average of n periods' data and associating it with the midpoint of the periods. For this case, we choose 3 time period to make the predictions of the data the most applicable and have the smallest error. For example, in 2011,2012,2013, the crime rate per 100k is 1074,1064,1015, so the forecast crime rate of 2012 is (11074+1064+1015)/3=1051, forecast error is 1064-1051=13. Total MSE (crime rate, from 2001 to 2022) is 827, MAE is 319.7. The predicted value and the actual value data line are shown in the figure below.

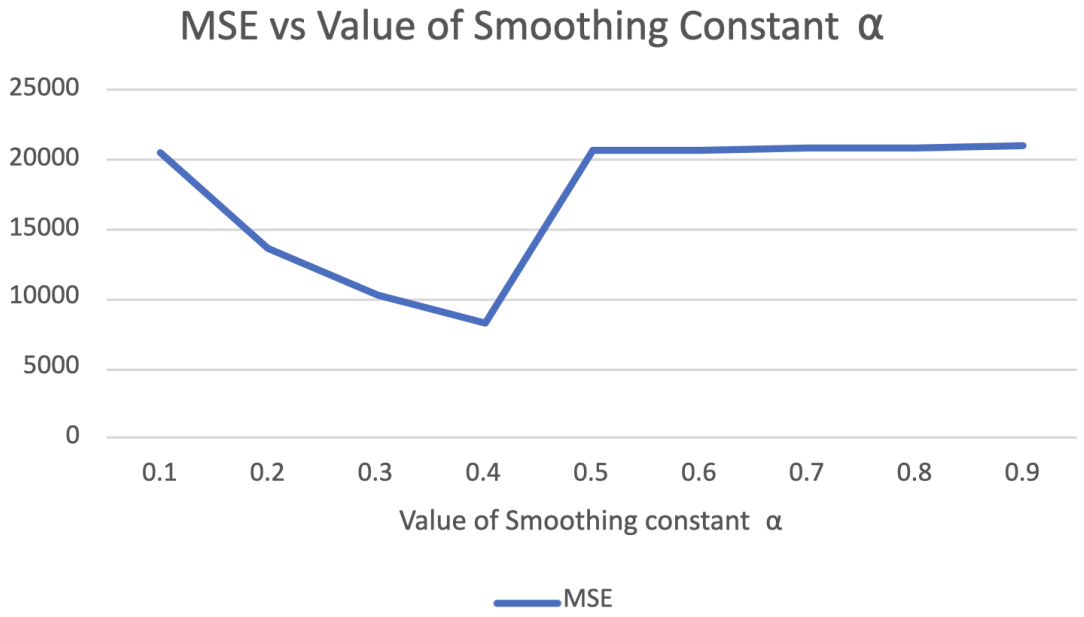


3.3 Exponential Smoothing

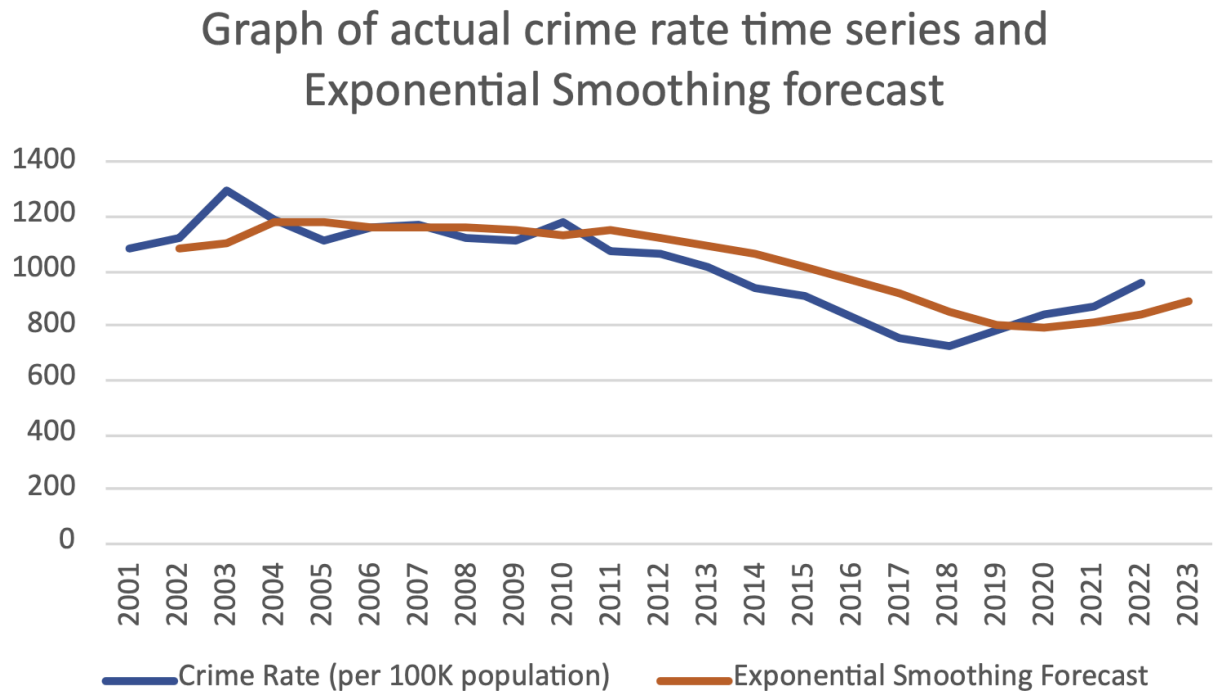
Exponential smoothing is a technique used in time series analysis to forecast future values based on past data. In this case, we calculated the mean squared error (MSE) by using different smoothing constants α, ranging from 0.1 to 0.9. A smaller MSE indicates better forecasts.

The equation is:

$$\text{Forecast for next period } (t+1) = \alpha \times \text{actual value for current period}(t) + (1-\alpha) \times \text{forecast for current period } t$$



The graph illustrates the relationship between the value of the smoothing constant α and the mean squared error (MSE). The line plot indicates that the MSE decreases initially as the value of α increases, reaching a minimum at α=0.4 with a value of 8296.8. Therefore, the optimum value of α that results in the smallest MSE is 0.4. This indicates that α=0.4 better forecasts compared to other values of α.



In the graph, the Exponential Smoothing forecast is computed using a smoothing constant of α=0.4.
The total MSE and MAE are 8296.8 and 73.9 respectively , and the forecast of crime rate per 100K population for 2023 would be 887.1 .

4 Auto-Regressive Integrated Moving Average (ARIMA) Modeling

Method Introduction

As can be surmised from the plot given in part 1 and 2, the graph clearly represents a non-stationary time series. A weakly-stationary time series is characterized by the presence of a constant mean and variance across all time points, whose autocovariance depends only on the time lag h, or its duration. In equations, it can be summarized into the following :

$$E(X_t) = \mu$$

$$Cov(X_t , X_{t+h}) = \gamma(h)$$

for a stationary time series $\{ X_t \}$.

Since this is not the case, the most common model for a non-stationary time series is utilized here, namely the ARIMA model. It consists of two parts; the auto-regressive and moving average component. With the addition of the integrated factor, the model is complete.

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - .... - \phi_p B^p \text{ (AR component)}$$

$$\Theta(B) = 1 - \Theta_1 B - \Theta_2 B^2 - .... - \Theta_q B^p \text{(MA component)}$$

The AR component stands for auto-regressive and MA the moving average. Multiplying the AR component by Y_t and MA component by Z_t, we have the following ARMA(p,q) model :

$$\phi(B) Y_t = \Theta(B)\, Z_t$$

$$Z_t \sim WN(0,\sigma^2)$$

The left hand side of the equation consists of the lagged values of Y_t , the observed value at time t. It is then related to the noise component Z_t which follows a white noise distribution on the right hand side. This forms the ARMA(p,q) model, where p and q represent the order of the AR and MA component respectively. The ARMA model is for stationary time series, so in order to accommodate for a non-stationary time series the addition of the integrated factor is necessary. Finally, we have the following ARIMA(p,d,q) model :

$$\phi(B)(1 - B)^d Y_t = \Theta(B) Z_t$$

where $(1 - B)^d$ is the integrated factor, or differencing, applied to Yt in order for it to behave in a non-stationary way.

Modeling Process

We first opted to fit the data with the auto.arima( ) function provided by the 'forecast' package in R. As the name might suggest, this function automatically computes the best ARIMA model to fit the data via its AIC, AICc, or BIC values. It also takes into consideration a number of other factors, such as unit root tests, maximum likelihood estimation of the parameters given, and so on. We can see from the first part that the time series data is not very complex, so we decided to optimize the function by setting 'stepwise = FALSE' for a non-stepwise selection that generally offers a more thorough search in exchange for time complexity. Finally, we turned 'approximation = FALSE' for a similar reason; we can sacrifice time for a better fitting model. The resulting MSE and MAE was calculated.

```
> crime = read.csv("C:\\Users\\hansj\\OneDrive\\Desktop\\CUHK\\Y3 T2\\STAT2011\\crime_rate_data.csv")
> crime = crime$Crime_Rate_Per_100k
> crime = as.ts(crime)
> t = 2001:2022
>
> fit_arima = auto.arima(crime, stepwise = F, approximation = F)
> print(summary(fit_arima))
Series: crime
ARIMA(0,1,0)

sigma^2 = 4446:  log likelihood = -118
AIC=237.99   AICc=238.2   BIC=239.03

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -5.614291 65.14545 50.85844 -0.749729 5.112284 0.9554725 0.141059
> (MSE = mean((fit_arima$residuals)^2))
[1] 4243.929
> (MAE = mean(abs(fit_arima$residuals)))
[1] 50.85844
```

The resulting model was rather troubling. The function decided that the ARIMA(0, 1, 0) model was the best fit for the given data. This has one pressing problem : given the syntax of the ARIMA equation mentioned previously, the ARIMA(0, 1, 0) model is essentially a *random walk* model.

$$Y_t - Y_{t-1} = Z_t ,$$

$$Z_t \sim WN(0,\sigma^2)$$

As the name might imply, the issue lies in the fact that the model is mostly unsystematic : the present value Y_t has only one non-random variable; that is its previous value Y_{t-1}. In other words, it does not make for a model where forecasting is reliable. The next value has an equal chance of going up or down by the same amount. Forecasting is a gamble; perhaps alike to the random walk theory in the stock market. Although a naive prediction may be performed, the auto.arima( ) function also may not be the best answer for every time series, since some of its inner workings are complex or out of our control.

Therefore, the solution is to manually input every combination of each of the parameters in the ARIMA model up to a certain amount. By doing so, we may be able to find a more optimized ARIMA model than random walk. Note that extremely high orders usually come with its drawbacks, namely its interpretability.

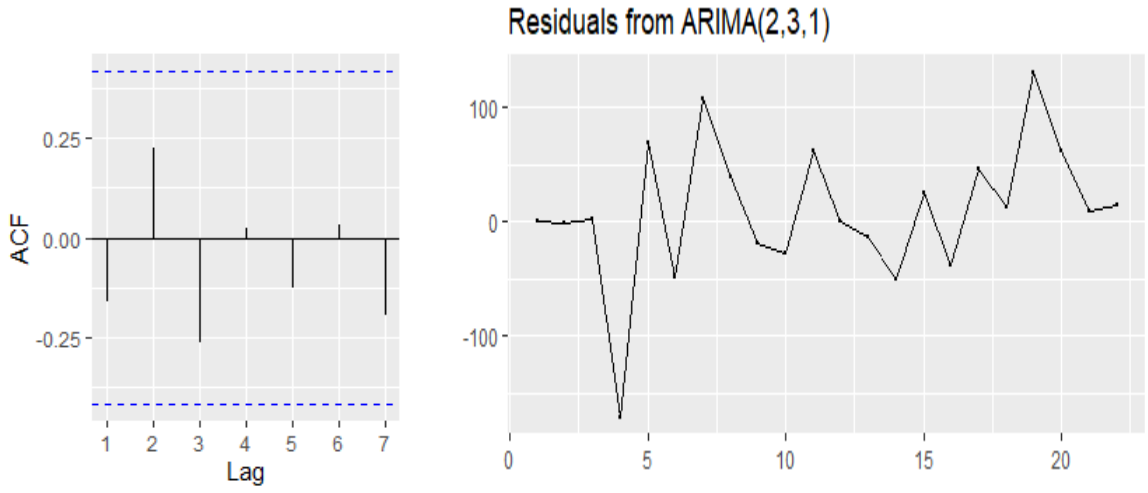
Model Selection

To start, we input every combination from 0 to 3 via loops of each of p, d, and q for the ARIMA(p, d, q) model. We then compute each resulting model's AIC and BIC values, widely used measures for the goodness of fit and parsimony(simplicity) of a statistical model. It will determine the order of the parameters of our optimized ARIMA model. By setting the initial value of the AIC and BIC value to be an arbitrarily large number, we are able to compare and update the lowest AIC and BIC value found throughout the loop.

```
> bestAIC = c(1e8, NA, NA, NA)
> bestBIC = c(1e8, NA, NA, NA)
> for(d in 0:3) for(p in 0:3) for(q in 0:3){
+       m = arima(crime, order = c(p, d, q), optim.control = list(maxit=1000))
+       if(AIC(m) < bestAIC[1]){
+         bestAIC = c(AIC(m), p, d, q)
+       }
+       if(BIC(m) < bestBIC[1]){
+         bestBIC = c(BIC(m), p, d, q)
+       }
+ }
> bestAIC
[1] 225.7596   2.0000   3.0000   1.0000
> bestBIC
[1] 229.5373   2.0000   3.0000   1.0000
```

Evidently, both the AIC and BIC measures agree that ARIMA(2, 3, 1) is our best fit model. We may then compute more detailed information from the model, along with the MSE and MAE :

```
> m = arima(crime, order = c(bestAIC[2], bestAIC[3], bestAIC[4]), optim.control = list(maxit=1000))
> print(summary(m))

Coefficients:
         ar1     ar2      ma1
      -0.5550  -0.7931  -0.8017
s.e.   0.1872   0.1844   0.1822

sigma^2 estimated as 4397:  log likelihood = -108.88,  aic = 225.76

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set 9.721786 61.62879 43.47104 1.272063 4.404514 0.8166862 -0.1615887
> checkresiduals(m)

        Ljung-Box test
data:  Residuals from ARIMA(2,3,1)
Q* = 4.467, df = 3, p-value = 0.2153

Model df: 3.   Total lags used: 6

> (MSE = mean((m$residuals)^2))
[1] 3798.107
> (MAE = mean(abs(m$residuals)))
[1] 43.47104
```



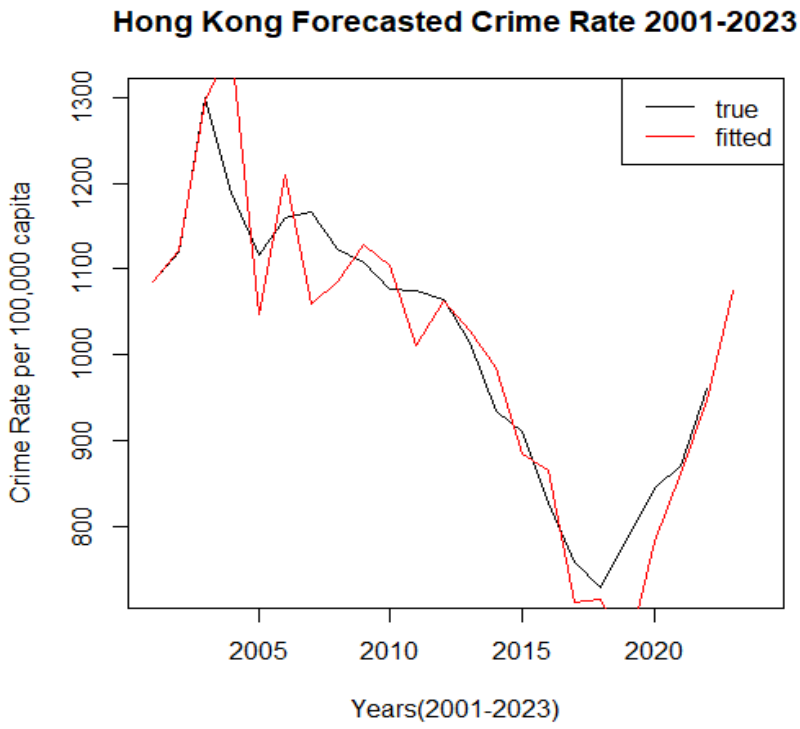
The resulting ACF plot also showed that the model stayed within the 2/√n confidence interval, which is good since we expect the model to be

stationary with no seasonal component. We note the values of the MSE and MAE, that is 3798.107 and 43.47104 respectively. Both are effectively lower than our initial random walk model (4243.929 and 50.85844).

Forecast

The last step is to forecast and visualize the data. Using the built-in predict( ) function in R and setting 'n.ahead = 1', we are able to predict Hong Kong's crime rate in 2023. Merging this forecast with the fitted model we may also visualize the adequacy of the ARIMA(2, 3, 1) model next to the dataset.

```
> (p = predict(m, n.ahead = 1)$pred)
[1] 1074.6
> (p - crime[22])
[1] 113.6005
> plot(t, crime, type = 'l', main = "Hong Kong Forecasted Crime Rate 2001-2023", xlim = c(2001,2024), xlab = "Years(2001-2023)", ylab = "Crime Rate per 100,000 capita")
> forecast_fit = c(fitted(m), p)
> lines(2001:2023, forecast_fit, col = 'red')
> legend("topright", legend = c("true", "fitted"), lty = c(1, 1), col = c("black", "red"))
```



Thus, our prediction for Hong Kong's crime rate in 2023 is 1074.6, a 113.6005 increase from the previous year as depicted above by the ARIMA(2, 3, 1) model.

5 Conclusion

Method	MSE	MAE	Forecast
Trend Projection	6676.635	68.24527	775.0494
Moving Average	14346.44	60.2192768	892.333
Center Moving Average	827	319.7	1051
Exponential Smoothing	8296.8	73.9	887.1
ARIMA (2, 3, 1)	3798.107	43.47104	1074.6

Given above are the comparisons between all the methods discussed thus far. The lowest MSE value is given by the Center Moving Average method, however it also has the highest MAE value by a very significant amount. Meanwhile, the lowest MAE value is the ARIMA(2,3,1) modeling. The highest MSE value is from the Moving Average method, much more than the other methods. Only both of Center Moving Average and ARIMA(2,3,1) have shown the continuation of the increase in trend of crime in the following year, while the rest have predicted a fall. Since 2023 is still in its early months, one can only wait to see the most reasonable and accurate method of analysis. However, due to the relatively low MSE, MAE and also the projection of the increasing trend, it is probable that the ARIMA(2,3,1) modeling may be favorable.

END OF REPORT

APPENDIX (Code)

Data Collection :
https://www.police.gov.hk/ppp_en/09_statistics/csc.html

```
id = "1_DtMXNjfrb5PHIcPp_bhXvWVMKfVgiw9"
crime_data =
read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",id))
head(crime_data)

year=crime_data[,1]
rate=crime_data[,2]

z_scores <- scale(rate)

outliers <- abs(z_scores) > 3

which(outliers)

#Plot the time series
plot(year,rate, type = "l",
    xlab = "Year", ylab = "Crime Rate (per 100,000 population)")
```

Trend projection :

```
t=year-2000
n=length(t)
tY=c()
for (i in 1:length(t)) {
  num=t[i]*rate[i]
  tY=c(tY,num)
}
b1=(n*sum(tY)-sum(t)*sum(rate))/(n*sum(t^2)-sum(t)^2)
mY=mean(rate)
mt=mean(t)
```

```
b0=mY-b1*mt
T=c()
for (i in 1:(n+1)) {
  T[i]=b0+b1*i
}
et=rate-T[1:n]
MSE=sum(et^2)/n
MAE=sum(abs(et))/n
#Plot the graph with trend projection
plot(rate, type = "l",xlab = "Time", ylab = "Crime Rate (per 100,000 population)")
abline(b0,b1,col="red")
```

ARIMA Modeling :

```
#install.packages("forecast")
#install.packages("tseries")
library(forecast)
library(tseries)

#reading file
crime = read.csv("C:\\Users\\hansj\\OneDrive\\Desktop\\CUHK\\Y3
T2\\STAT2011\\crime_rate_data.csv")
crime = crime$Crime_Rate_Per_100k
crime = as.ts(crime)
t = 2001:2022

#auto arima modeling
fit_arima = auto.arima(crime, stepwise = F, approximation = F)
print(summary(fit_arima))
(MSE = mean((fit_arima$residuals)^2))
(MAE = mean(abs(fit_arima$residuals)))

bestAIC = c(1e8, NA, NA, NA)
```

```
bestBIC = c(1e8, NA, NA, NA)
for(d in 0:3) for(p in 0:3) for(q in 0:3){
    m = arima(crime, order = c(p, d, q), optim.control = list(maxit=1000))
    if(AIC(m) < bestAIC[1]){
      bestAIC = c(AIC(m), p, d, q)
    }
    if(BIC(m) < bestBIC[1]){
      bestBIC = c(BIC(m), p, d, q)
    }
}
bestAIC
bestBIC

m = arima(crime, order = c(bestAIC[2], bestAIC[3], bestAIC[4]), optim.control =
list(maxit=1000))
print(summary(m))
checkresiduals(m)
(MSE = mean((m$residuals)^2))
(MAE = mean(abs(m$residuals)))


plot(t, crime, type = 'l', main = "Hong Kong Crime Rate 2001-2022")
lines(t, fitted(m), col = 'blue')

(p = predict(m, n.ahead = 1)$pred)
(p - crime[22])

plot(t, crime, type = 'l', main = "Hong Kong Forecasted Crime Rate 2001-2023",
xlim = c(2001,2024), xlab = "Years(2001-2023)", ylab = "Crime Rate per 100,000
capita")
forecast_fit = c(fitted(m), p)
lines(2001:2023, forecast_fit, col = 'red')
legend("topright", legend = c("true", "fitted"), lty = c(1, 1), col = c("black", "red"))
```