Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

Stat 443 Project Report

Goal: For our project, the main goal is to perform time series analysis on the previous 10 years monthly total crime of San Diego to predict the future number of crime using different models. The last 12 months of the crime data will be used as the holdout set. Also, we are to determine which of the forecasting rules (persistence, average of all past, exponential smoothing, ARIMA, ARIMAX with explanatory variables.) would provide us with the most accurate prediction of the total amount of crimes in the future. We used data from San Diego, USA to execute this forecast.

Conclusions: We start by analysing the previous 10 years of monthly total crime of San Diego to predict the future number of crime. First, we used simple rules such as persistence and average of all past.

Then, we used exponential smoothing, Arima and ARIMAX. We have selected a couple of relevant explanatory variables(the unemployment rate, the average hourly wage) to see if they can assist in increasing our prediction accuracy. We saw that the unemployment rate has a positive effect on the amount of crimes and the average hourly wage has a negative effect on the crime rate. After performing the analysis, we found out that ARIMA(2,1,1) had the lowest RMSE followed by ARIMAX , Exponential Smoothing , Persistence and Average of the previous observation. For ARIMAX we selected a few models based on the residuals acf ,pacf graph and selected some candidates models to fit the data. Among the selected we choose the one with the smallest in-sample mse to make prediction on the holdout set. Similarly for ARIMA we considered two models to make predictions namely ARIMA(2,1,1) and ARIMA(1,1,1). We thought ARIMAX will result in the smallest RMSE but the holdout predictions seems to indicate otherwise. This could be because we were unable to find other strongly correlated variable with total crime or potential outliers within explanatory variables , thus decreasing its predictions accuracy.

Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

Variables

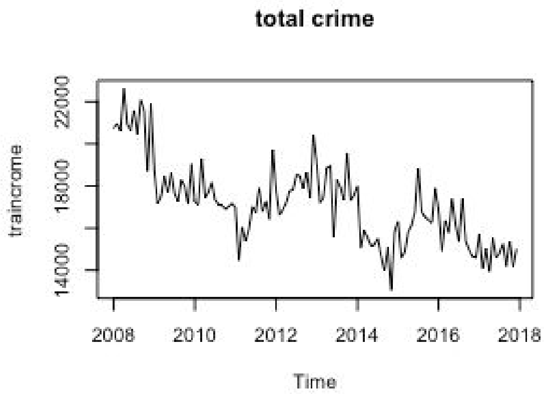
1st Variable : Total Number of Crime in San Diego over the last 10 years.

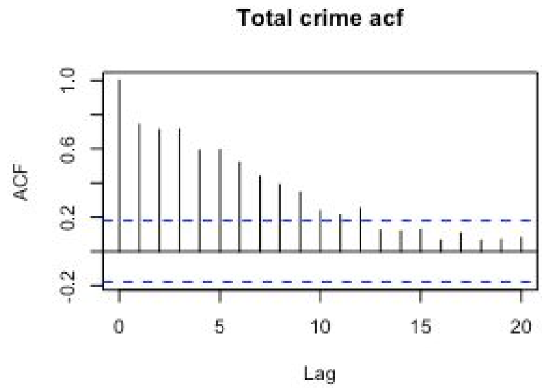
2nd Variable: Average Hourly Income (USD/hour) of San Diego over the last 10 years.

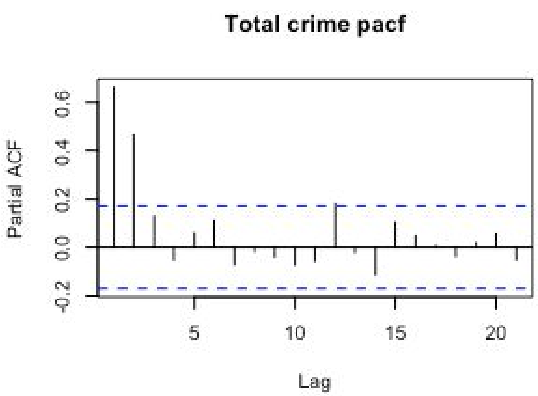
3rd Variable: Percentage Unemploymed of San Diego over the last 10 years.

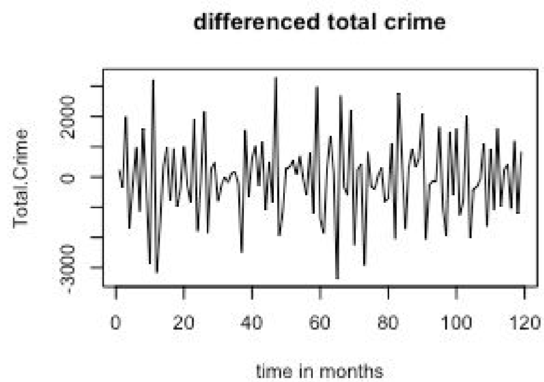
Total # of Crime Summary					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7464	15340	16930	16810	17910	22620
Average Hourly Income					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
24.8	25.67	27.1	27.15	28.5	29.5
Unemployment Rate Summary					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.9	4.8	6.8	7.0	9.63	11.1

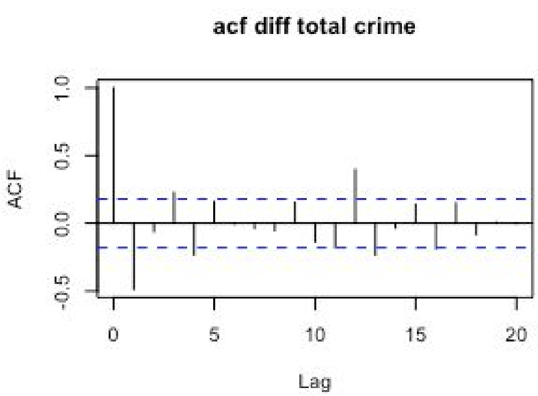
Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

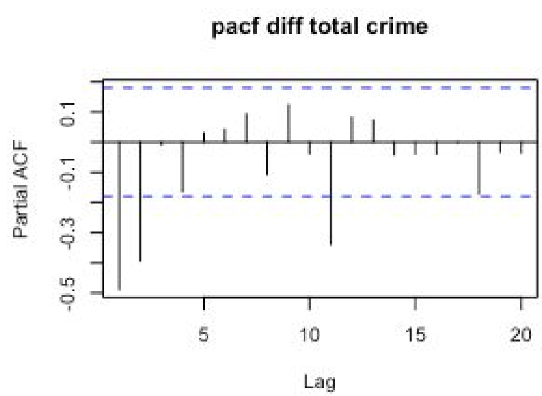






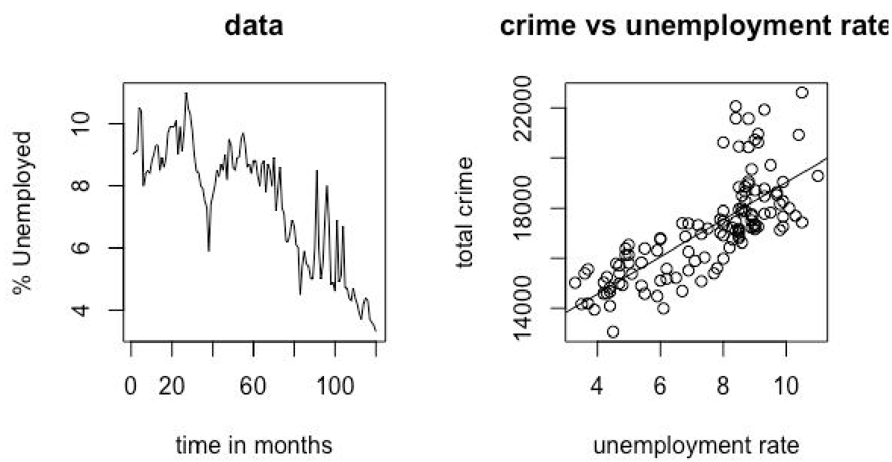






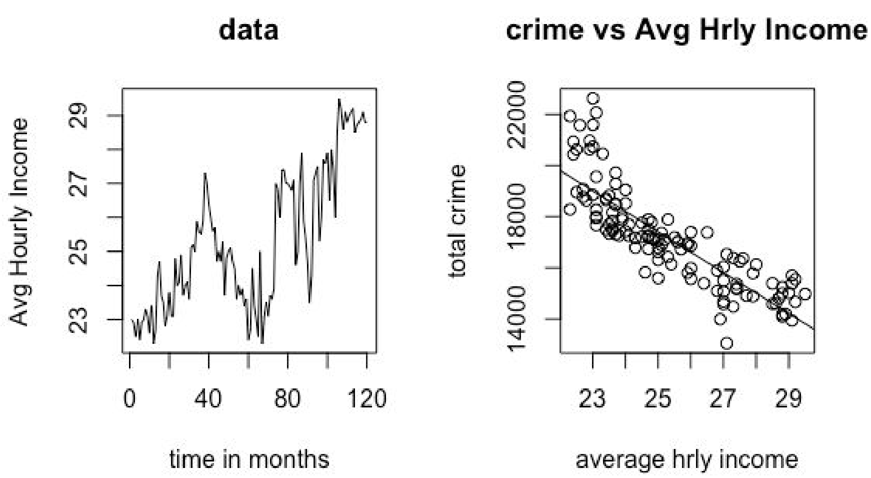
Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

Explanatory variable 1: unemployment rate



We get R^2 = 0.5625788

Explanatory variable 2: average Income



We get R^2 = 0.7146671

Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

Analyses

Persistence Rule

We used the last element of our training set to do our forecast. We get our forecast error by subtracting each elements in our holdout set to our forecast element, squaring the difference and summing them up forming our mean square error. The resulting rmse is 3488.662.

Average of All Past

We first get the mean of the training set to get the first forecast, then we add the first observation to the training set and get their mean to get the second forecast. Repeatedly, we get 12 forecast and the resulting rmse compared to the holdout set is 4444.028.

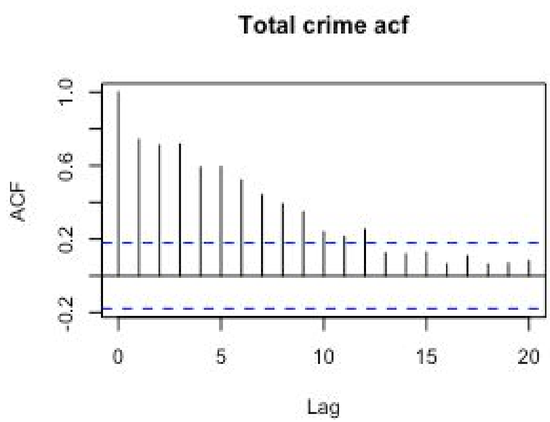
Holt-Winters Additive Exponential Smoothing

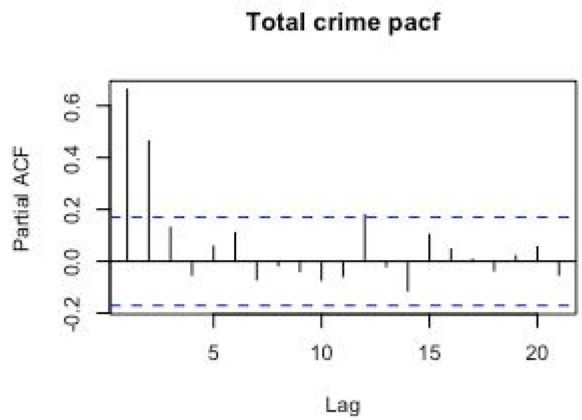
Since our data are from 2008 to 2018 month by month, we make the time series that have the frequency 12 and the Holt-Winters Exponential Smoothing should be seasonal. From this model we get a rmse of 2863.947.

```
HoltWinters(x = zcrime, seasonal = "additive")

Smoothing parameters:
 alpha: 0.4737818
 beta : 0.03262502
 gamma: 0.6782157

Coefficients:
            [,1]
a   14127.417006
b     -57.888134
```

ARIMA without the explanatory variables

We use the model Arima(2,1,1) to fit the data since we see that the acf of the total crime is exponential decreasing and its pacf cuts off after lag 2. We needed 1 differencing since the series presented a trend and was not stationary so, the ARIMA(2,1,1) model was considered. The ARIMA(2,1,1) model gave us a rmse of 2324.793.





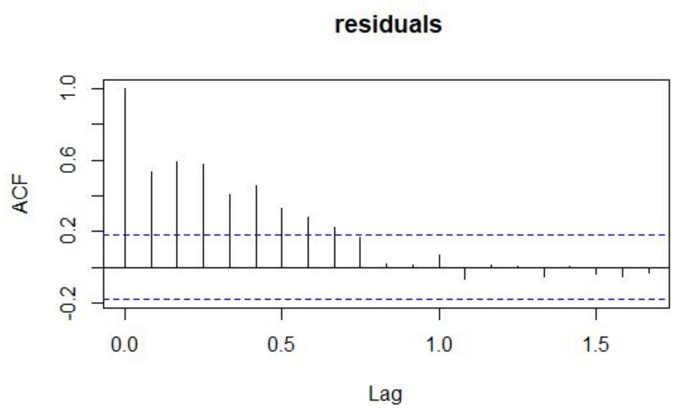

```
arima(x = zcrime, order = c(2, 1, 1), method = "CSS")

Coefficients:
          ar1      ar2     ma1
      -0.7926  -0.4435  0.1407
s.e.   0.4631   0.2170  0.5548

sigma^2 estimated as 1206537:  part log likelihood = -1002.05
```

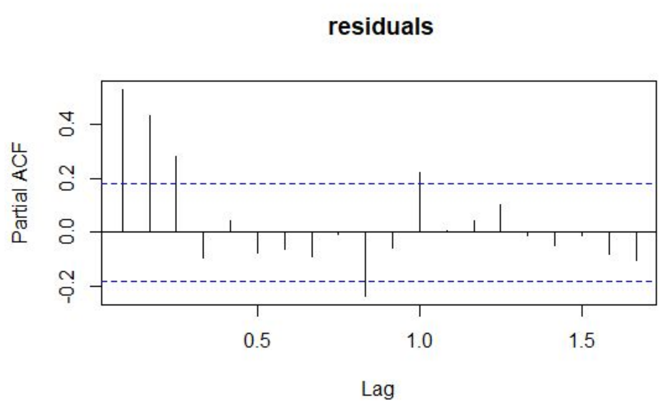
We also tried fitting the data with the model ARIMA(1,1,1), and it gave us a rmse of 4699.231. Which is bigger than that of ARIMA(2,1,1).

```
arima(x = zcrime, order = c(1, 1, 1), method = "CSS")

Coefficients:
          ar1      ma1
      -0.1633  -0.5294
s.e.   0.1280   0.1009

sigma^2 estimated as 1236536:  part log likelihood = -1003.51
```

Naman Bansal
Qiwen Li
Xiaojia Li
Leo Liang

ARIMAX





For ARIMAX we found 2 explanatory variable, Percentage of Population Unemployed and Average Hourly Income, to be correlated with the response variable, Total Crime for the same time period. After running the regression we found the residuals to follow the AR(3) model. Other model such as AR(2), AR(4) were also considered. Based on the in-sample RMSE, AR(4) was chosen to be the most appropriate since AR(5) only showed marginal gain.

	ARIMAX(2)	ARIMAX(3)	ARIMAX(4)
In sample rmse	733.42	723.61	717.64

DATE / forecast rules	HOLDOUT	Persistence	Average	Exponential Smoothing	ARIMA(2,1,1)	ARIMA(1,1,1)	ARIMAX (4,1,0)
2018-01-01	16345	15007	17143.16	15995.04	14911.52	15015.57	15377.26
2018-02-01	13663	16345	17136.56	14044.51	15371.02	16830.3	14516.34
2018-03-01	14236	13663	17108.09	14562.05	15131	12424.2	14876.89
2018-04-01	13439	14236	17084.74	14215.21	14054.64	15101.59	14355.25
2018-05-01	15581	13439	15055.34	14411.08	14021.96	12688.97	13952.93
2018-06-01	13714	15581	17043.54	14098.05	14287.14	16762.25	14584.89
2018-07-01	14769	13714	17017.12	14419.02	14693	12405.13	14170.46
2018-08-01	15116	14769	16999.42	14866.69	14313.56	15848.14	14665.63
2018-09-01	14612	15116	16984.7	13798.11	14815.59	14671.73	14249.46
2018-10-01	8053	14612	16966.31	14755.6	14426.68	14662.67	14716.22
2018-11-01	14214	8053	16897.75	11141.42	12269.63	5624.92	12745.61
2018-12-01	7464	14214	16877.26	13845.27	10757.8	17754.96	11653.29
rmse		**3488.66**	**4444.03**	**2863.95**	**2324.79**	**4699.23**	**2434.50**

Contributions

1) are a team of friends
2) alphabetical by surname
3) major contributions:
    - Naman : topic, coding, statistical analysis, criticism, organizer of discussion meetings
    - Qiwen : topic, organizer of discussion meetings, model analysis
    - Xiaojia : topic, model analysis, coding
    - Leo : topic, statistical analysis, writing

Overall we feel the work was divided equally and everyone did their part.