Los Angeles City Crime Rate Time Series

Modeling and Forecasting

Table of Contents

Executive Summary

Crime presents threats to public safety, and its influencing factors could be complex. This report aims at identifying the influencing factors and patterns of the occurrences of crimes in Los Angeles. Understanding the impact of economic and other external factors such as weather conditions on crime is critical for ensuring government's agility to deal with recurrent crises. Specifically, we measure the impact of GDP, homelessness, weather conditions and unemployment rate on crime per capita in Los Angeles from January 2010 to August 2020. The research reveals that crime shows a weak positive correlation with GDP, rent CPI, temperature and a slight negative correlation with the unemployment rate. Based on the overall robustness, simplicity, accuracy and interpretability, an ADL model with seasonal components, crime rate lagged by 2, as well as GDP lagged by 12 and its quadratic terms and cubic term is chosen to provide a one-step forward crime rate forecast in LA.

Introduction & Motivation

Recently, frequent crime-alerts from the USC Department of Public Safety have been worrisome for USC students. In addition, it has been noticed that leading up to the 2020 presidential election, the Republican party and President Trump's campaign have been claiming that cities and states governed by the Democratic party have experienced more constant security threats, and the city of Los Angeles is one of them.

Going into 2020, in the midst of the COVID-19 pandemic, the national economic distress, and the recent nation-wide protests stimulated by police brutality and racial injustice, more cases of violent crimes in Los Angeles have been reported in the media than in ordinary times. It is easy for one to conclude that the crime rate has been rising in LA. However, public perceptions of crime may be misled by bias in political campaigns and media.

To get a clearer picture on this issue, we want to investigate Los Angeles crime rate, its history, its trend, potential influencing factors, and how we might predict future crime rate changes. The report aims to explore various components of the Los Angeles crime time series data, understand the relationship between the crime rate time series and potential influencing factors, namely economy, homelessness, and weather, and to recommend a model for one-step forward forecasting of the Los Angeles crime occurrences.

Data Description

1. Crime Rate Data

The raw crime data contains daily Los Angeles crime incidents form Open Data Portal. The raw dataset consists of over 2 million rows of data at incident level with reported date, time of occurrence, location, victim description and nature of crime. We retrieved data from January 2010 to August 2020, and aggregated it into a monthly level for time series analysis. Furthermore, we normalize the raw crime incident count into crime rate by dividing it by the yearly Los Angeles population.

2. External Data

- Monthly US Gross Domestic Product (GDP) data: The period of data is the same as the Los Angeles crime data, from January 2010 to August 2020. The data contains the monthly US GDP in trillion dollars, retrieved from ycharts.com. Note that the Bureau of Economic Analysis only publishes quarterly GDP data. To get a more accurate monthly measurement, we utilized data produced by the economic research institution YCharts, which is extrapolated based on a set of assumptions about the quarterly data. Since Los Angeles is one of the biggest economic powerhouse cities in the US, the US GDP and the Los Angeles GDP should have a similar trend. Therefore, the monthly US GDP data is used as a proxy for Los Angeles' economic development.
- Monthly Los Angeles City Unemployment Rate data: The period of this dataset is from January 2010 to August 2020, retrieved from the Bureau of Labor Statistics.
- Monthly Los Angeles Rent CPI: We use this rent data as a proxy for homelessness, since no monthly homelessness data is available for analysis. Based on previous research, there is a significant positive relationship between a community's housing market conditions and the size of its homeless population[1]. The period of this monthly Los Angeles Rent CPI dataset is from January 2010 to August 2020, retrieved from the Bureau of Labor Statistics.
- Monthly Los Angeles Temperature data: This dataset includes the average monthly temperature of Los Angeles City from January 2010 to August 2020, retrieved from National Centers for Environmental Information.
- Yearly Population Data: This dataset is from January 2010 to August 2020, including the yearly population of Los Angeles City in thousands, retrieved from populationstat.com. Note that the official population census data is only updated every ten years. The yearly population data is extrapolated from the census data with a set of assumptions.

[1] Research can be found at: https://www.huduser.gov/portal/sites/default/files/pdf/Market-Predictors-of-Homelessness.pdf

Methodology

We split all the time serieses into training, testing and validation sets. The data from January 2010 to December 2018 is set to be the training set. Time series models are trained with training data to forecast next month's crime rate. To test the model performance, the data from January 2019 to December 2019 is set to be the testing set, which is used to calculate the accuracy of the predicted crime rate generated by the priorly fitted model. Finally, the models are retrained with data in the training and testing sets, and are used to predict crime rate in the validation period and compared with the validation set, which includes the data from January 2020 to August 2020. The reason why 2020 is excluded from model training is that GDP and unemployment rate have been severely impacted by COVID-19 and its shock to the economy. Using data with rare structural breaks to train models will cause problems.

To build models for forecasting, six statistical approaches are used to build time series forecasting models, and MAPE is calculated for model evaluation and comparison.

1. **Linear regression**: A linear approach to modeling the relationship between a dependent variable and one or more independent variables.

   The advantage of linear regression is interpretability. The limitation of linear regression is that the strong assumptions about residuals need to be carefully examined to assess the usability of the approach. We tried including trend, season, quadratic trend, and cubic trend as the independent variables in linear models, and also tried adding one or more of the external variables such as GDP, unemployment rate in the models.

2. **Moving average (MA)**, a data-driven approach that calculates the averages of moving windows as the smoothed time series, and uses the last data point in the series as the one-step forward forecast. To forecast a series at time t+1, we use a trailing MA that ends at time t:

$$F_{t+k} = \frac{y_{t-W+1} + y_{t-W+2} + \cdots + y_{t-1} + y_t}{W}$$

   The upside of MA is that it does not have any assumptions about data so it is widely applicable. The limitation of this approach is that cyclical patterns will not be captured by moving averages. If the crime rate is bouncing up and down a lot, a moving average model is not likely to capture this pattern.

3. **Exponential smoothing**, a technique for smoothing time series data using the exponential window function. It is a data-driven, adaptive algorithm that adjusts the most recent forecast based on the actual data:

$$F_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \ldots$$

when α = the smoothing constant ($0<\alpha\leq 1$).

The upside of exponential smoothing techniques is that it requires less data to be saved and is computationally efficient. The downside is that it requires frequent updating to capture the recent changes in data. The "AAdA" model, which is exponential smoothing with additive seasonal components and dampened trend, is selected as the best exponential smoothing model.

4. **Seasonal autoregressive integrated moving average (SARIMA)**, an ARIMA model with additional seasonal terms, which learns from the recent data points and recent prediction errors.

While the algorithm is adept at modelling seasonality and trends, it tends to overfit if the terms are not properly selected. Four SARIMA models are attempted for this project.

5. **Decision tree**, a data-driven algorithm that cuts up data space with decision boundaries along independent variables that minimize the variability inside a carved space, and use the average of the target values as the predicted output.

Decision trees are easy to interpret, but tend to overfit and are not very robust to change in data. A tree model that uses GDP and temperature data was built in this report.

6. **Autoregressive distributed lag model (ADL)**, a regression approach for stationary time series to learn from the past data points and external variables that have a leading effect on the series. An ADL(p,q) model assumes that a time series Y_t can be represented by a linear function of p of its lagged values and q lags of another time series X_t:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_\rho Y_{t-\rho} +$$

$$\delta_1 X_{t-1} + \delta_2 X_{t-2} + \ldots + \delta_q X_{t-q} + u_t$$

is an ADL model with p lags of Y_t and q lags of X_t where

$E(u_t|Y_{t-1},Y_{t-2},\ldots,X_{t-1},X_{t-2},\ldots)=0$.

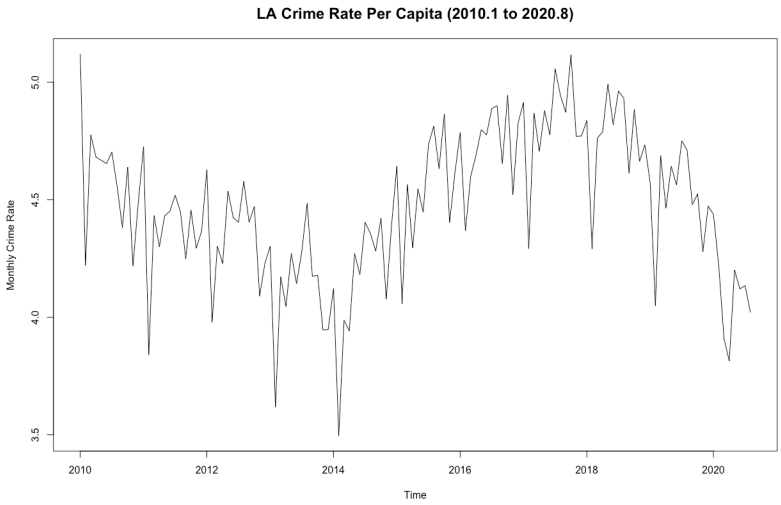
The upside of the ADL model is to be able to show the leading effect of external variables. The downside is that the independent terms are likely to be highly correlated, which in turn causes problems in coefficient estimates. Several ADL models with lagged crime time series, lagged GDP time series, lagged rent CPI time series, lagged temperature time series and lagged unemployment rate time series are built in this report.

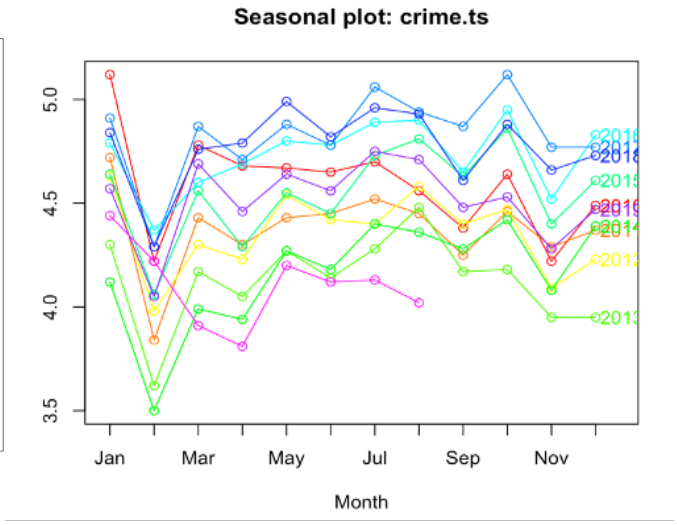
Analysis and Results

1. Data Exploration

Our response variable, Los Angeles monthly crime rate per capita, demonstrates a quadratic trend from 2010 to 2020, with a declining trend from the beginning to 2014, a rising trend from 2014 to 2017, and a declining trend again until recently. The time series also has seasonality, with crime rate higher in summer months and lower in winter months.

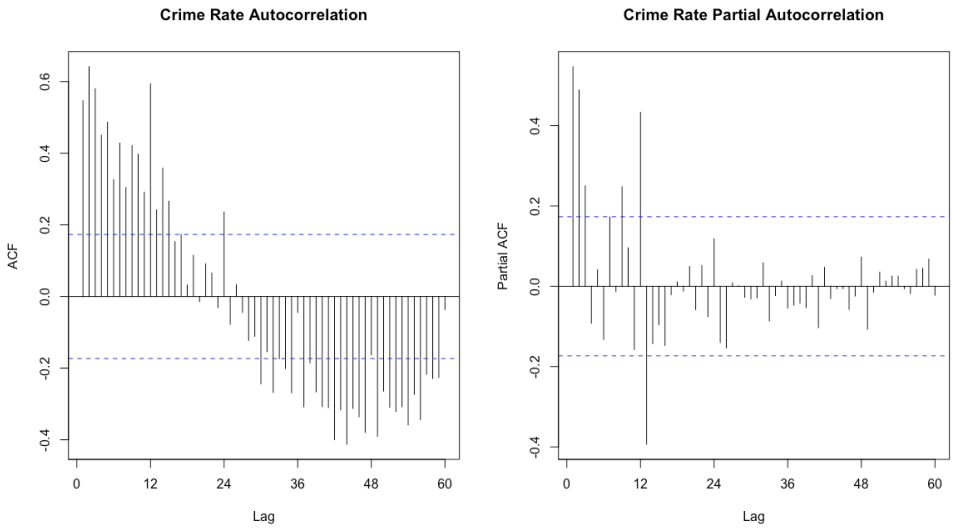
To identify autocorrelation in crime rate itself, we inspected the ACF and PACF plots, and found that crime rate's seasonal autocorrelation tails off and seasonal partial autocorrelation cuts off



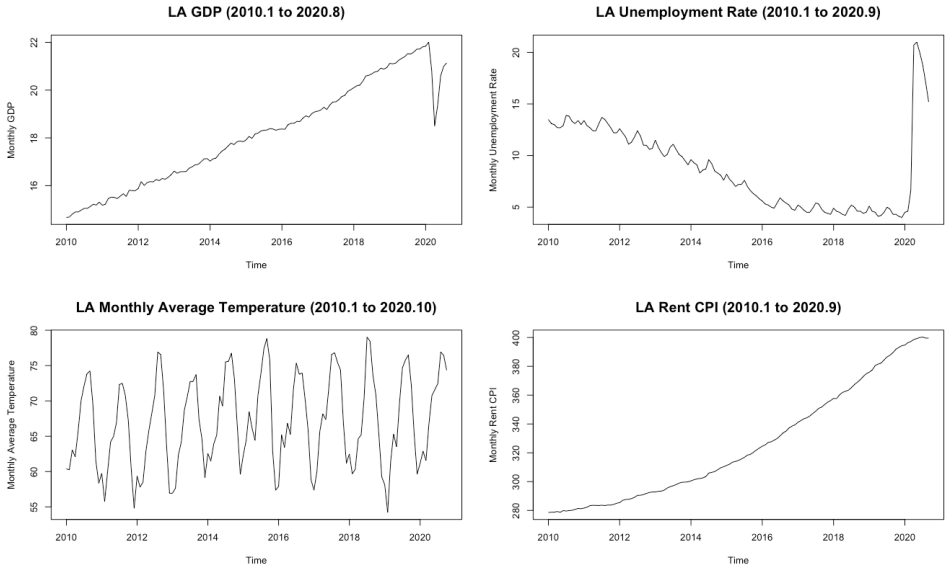


after one season, and its within-season autocorrelation tails off and partial autocorrelation cuts off after two months, indicating an AR(2) and seasonal structure.

Regarding external variables, GDP of Los Angeles demonstrates a continuous growing trend,



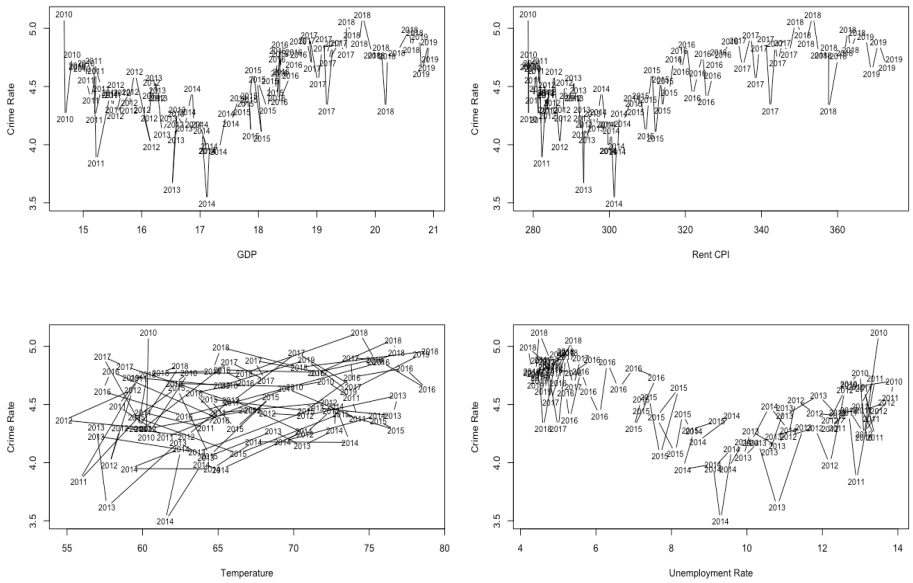
with a visible structural break in April 2020 due to the impact of COVID-19. Unemployment rate in Los Angeles shows a steady declining trend with seasonality, until a sudden rise in 2020 due to COVID-19. Los Angeles monthly average temperature shows a stable seasonality pattern but no obvious trend over the last 10 years. Los Angeles monthly rent CPI has been monotonically increasing over the past 10 years, indicating a steady increase in housing cost, which is positively associated with homelessness.



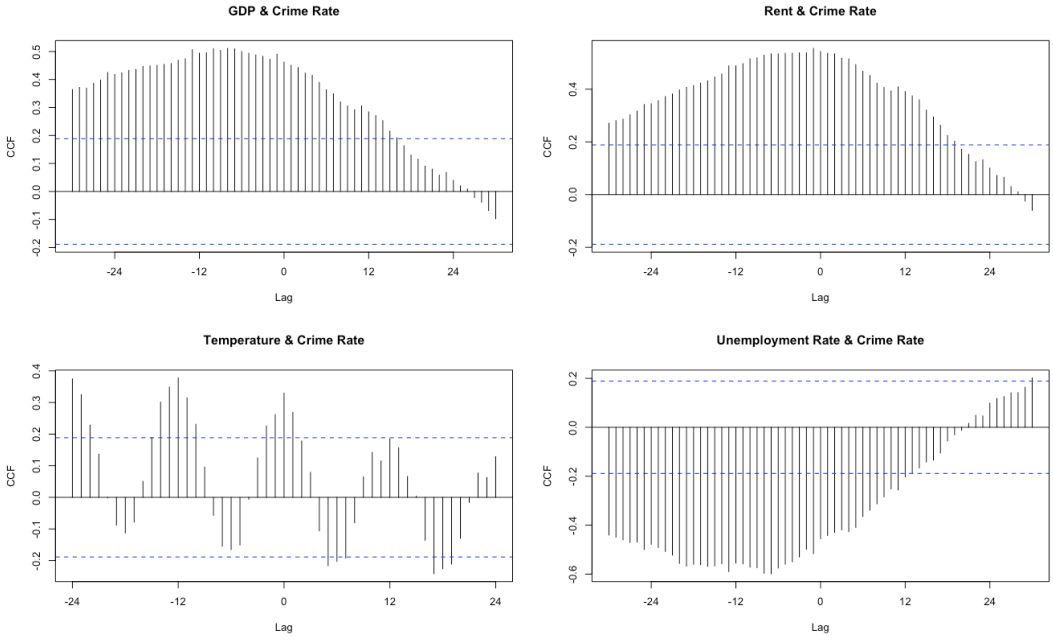
We checked the stationarity of all the time series using Augmented Dickey Fuller test and found that unemployment rate and rent CPI are not stationary, with p values 0.2821 and 0.9656 respectively (the alternative hypothesis is that the time series is stationary). Therefore, when we create ADL models, we use the detrended unemployment rate and rent CPI series.

We calculated the overall correlation between crime rate and external variables to inspect their linear relationships. Crime rate shows a positive though weak relationship with GDP, rent CPI and temperature, and a slight negative relationship with unemployment rate which may be counterintuitive at first sight. However, if we look closer at the scatterplot of unemployment rate and crime rate, we will notice that the two variables have a quadratic relationship, with a negative correlation in lower unemployment rate range and a positive correlation in higher unemployment rate range.

```
    crime        gdp       rent       temp         ur
1.0000000  0.2568808  0.2009418  0.2890742 -0.4572438
```



To identify potential leading effects of the external variables, we examined the cross-correlation plots between them, and identified the time gaps that show the most significant leading relationship, which are: GDP: lag 12; rent CPI: lag 2; temperature: lag 12; unemployment rate: lag 4.



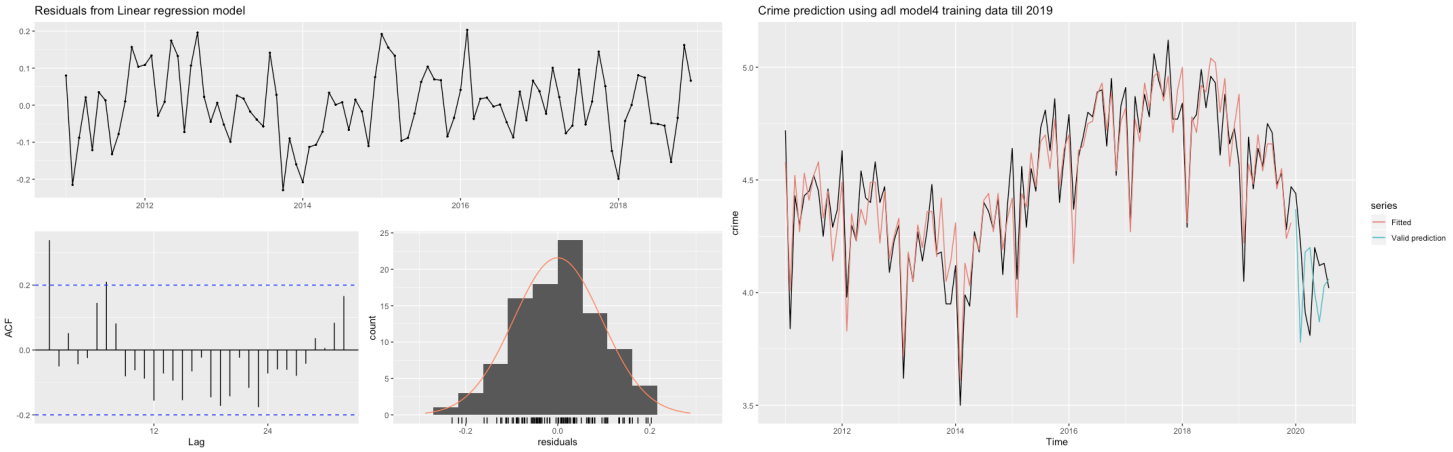
2. Model Performances

We tried a range of model algorithms, including linear regression model with trend and seasonality, linear regression with external variables, ADL model, moving average model, exponential

smoothing model, SARIMA model, and decision tree model. After comparing performances of all the fitted models, the final model we chose is an ADL model with the response variables seasonal components, crime rate lagged by 2, GDP lagged by 12 and its quadratic terms and cubic term. The model parameters are shown in the output:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     125.262303  29.185153   4.292 4.92e-05 ***
season2          -0.705346   0.054594 -12.920  < 2e-16 ***
season3          -0.308372   0.061968  -4.976 3.64e-06 ***
season4          -0.100285   0.059309  -1.691  0.09475 .
season5          -0.108350   0.054965  -1.971  0.05215 .
season6          -0.159373   0.053010  -3.006  0.00353 **
season7          -0.110726   0.059844  -1.850  0.06797 .
season8          -0.044919   0.056069  -0.801  0.42543
season9          -0.313618   0.063033  -4.975 3.66e-06 ***
season10         -0.139367   0.064849  -2.149  0.03465 *
season11         -0.367166   0.055992  -6.558 4.96e-09 ***
season12         -0.311524   0.064625  -4.820 6.71e-06 ***
crime.lead2       0.475237   0.096870   4.906 4.81e-06 ***
gdp.lead12      -21.361496   5.016362  -4.258 5.57e-05 ***
I(gdp.lead12^2)   1.229693   0.288684   4.260 5.54e-05 ***
I(gdp.lead12^3)  -0.023406   0.005504  -4.252 5.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1045 on 80 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.8989
F-statistic: 57.31 on 15 and 80 DF,  p-value: < 2.2e-16
```



From the residual checking plot above, we can see that the residuals of the model are relatively normally distributed. The variance of residuals is fairly constant. There is no obvious autocorrelation left in the residuals. The linear model assumptions are met.

As can be seen from the model performance table below, the chosen model, No. 25, has one of the lowest testing errors and is one of the most robust models. Model No. 24's testing error is also one of the lowest and has a better 2020 validation performance, but it contains nonsignificant variables. However, one thing to note is that Model No.25 assumes that the cubic relationship between GDP lagged by 12 and crime rate will be consistent in the future, which needs to be verified once we have

more future data points. Based on model performance, we choose Model No.25 as our final recommendations.

We conduct a t-test of the 2020 prediction using Model No.25 and training data until the end of 2019 and the actual 2020 data, and found that there is not a significant difference in means (p-value = 0.655), indicating that there might not be a structural break in 2020 crime rate that our model does not capture.

Model No.	Model Type	Model Components	Training MAPE (2010.1-2018.12)	Testing MAPE (2019.1-2019.12)	2020 MAPE using training data till 2019
1	Linear Regression	trend	5.11	5.91	15.16
2	Linear Regression	trend, trend^2	4.3	16.15	19.71
3	Linear Regression	trend, trend^2, trend^3	3.76	5.09	4.07
4	Linear Regression	trend, trend^2, trend^3, season	2.2	4.16	5.33
5	Naive	Last Value	-	5	9.07
6	SNaive	Last Seasonal Value	-	5.7	12.16
7	Trailing Moving Average - Same Prediction Value	Moving average with same prediction value	3.69	5.86	10.19
8	Trailing Moving Average - Rolling forward with 1 step	Moving average rolling forward one step	3.69	4.04	7.84
9	Exponential Smoothing	AAdA	1.71	5.11	8.74
10	SARIMA	SARIMA (2,1,0)X(1,1,0)12	2.95	3.48	8.24
11	SARIMA	SARIMA (2,1,0)X(1,0,0)12	3.79	3.94	8.69
12	SARIMA	SARIMA (1,1,0)X(1,0,0)12	2.93	3.57	7.85
13	SARIMA	SARIMA (1,0,0)X(1,0,0)12	1.68	4.9	9.7
14	Linear Regression	GDP	4.94	7.43	-
15	Linear Regression	Rent CPI	5.24	4.42	-
16	Linear Regression	Temperature	5.59	3.13	-
17	Linear Regression	Unemployment Rate	5.64	2.74	-
18	Linear Regression	Unemployment Rate, GDP	4.71	7.47	-
19	Linear Regression	Rent CPI, GDP	4.93	6.94	-
20	Linear Regression	Temperature, GDP	4.71	6.41	-
21	Linear Regression	Rent, Unemployment Rate, GDP	4.68	7.14	-

22	ADL	ADL Model, trend+season+Crime Rate(t-2)+GDP(t-12)+ GDP(t-12)^2 + GDP(t-12)^3+Rent CPI(t-2)+Temperature (t-12)	1.74	4.71	5.59
23	ADL	season+Crime Rate(t-2)+GDP(t-12)+ GDP(t-12)^2 + GDP(t-12)^3+Rent CPI(t-2)+Temperature (t-12)	1.74	4.36	5.58
24	ADL	season+Crime Rate(t-2)+GDP(t-12)+ GDP(t-12)^2 + GDP(t-12)^3+Rent CPI(t-2)	1.71	4.48	5.43
25	**ADL**	**season+Crime Rate(t-2)+GDP(t-12)+ GDP(t-12)^2 + GDP(t-12)^3**	**1.73**	**4.54**	**5.42**
26	Decision Tree	GDP, temp	3.54	5.6	12.49

Recommendations

In this section, we propose the following recommendations based on the findings of our analysis and selected model, and propose suggestions on further research.

Our best model includes GDP as the influencing factors. In addition, our analysis showed statistical association between crime rate and rent CPI, unemployment rate, and temperature. Therefore, we make the following recommendations for all three economic factors:

- GDP: Economic wellbeing has a leading effect on crime rate. While systematic risks are hard to inhibit, a healthy and sustainable economic growth is always accompanied with ample consumptions and investments, especially in research and development. Policymakers should encourage innovations and consumptions at a reasonable level to keep economic growth.
- Rent CPI: A higher rent level and lower median income are associated with a higher homelessness rate and higher crime rate. In 2019, the median rent of $2,095 is 45% of the $55,820 average salary in Los Angeles, suggesting many people from the low income group might be going through housing distress. To remedy this issue, the government can incentivize real estate developers to provide more affordable housing by tax abatement or pose rent limits in certain areas.
- Unemployment rate: In the higher range of unemployment rate, crime is positively associated with it. To address the negative impact of a high unemployment rate in the short term, the government should provide unemployment benefits to citizens in need. In the long haul, the government should develop core industries resilient to economic downturns in Los Angeles, such as the healthcare industry, and provide retraining programs to help people re-enter the workforce.

While our model achieves decent accuracy and robustness, we recommend the following improvements in future work:

- Predictor variables:
    - More demographic (age, level of education etc.) and societal variables can be examined with a longer time span, especially variables that increase interpretability of crime rate change, and those that can provide direction for future regulation or stimulation. Since our analysis only covers the recent 10 years, it is safe to assume that the demographic attributes of Los Angeles are relatively stable.
    - Include stronger proxies for homelessness. Although rent CPI is a recommended proxy for homelessness according to the previous literature, other variables such as participation in social safety net programs and demographic variables can together give a better representation of the homelessness issue.

- Response variable: While our analysis includes all kinds of crime incidents, it might be worthwhile to differentiate the type of crimes in future analysis. For instance, violent crimes might be more positively correlated with a higher temperature than pilferage, and a higher homeownership rate might be associated with a higher burglary rate, but negatively correlated with other types of crime.

Conclusions

In this report, we first outlined the motivations of examining Los Angeles crime rate, and then described why we chose the selected variables and their sources, namely monthly Los Angeles crime rate per capita, US GDP, Los Angeles unemployment rate, Los Angeles temperature, and Los Angeles rent CPI. Next, we laid out our methodology for model fitting. Then we analyzed the crime rate and presented our best model for Los Angeles crime rate forecasting. Finally, we proposed policy recommendations based on our findings and ideas for future work.

Appendix

Data Source:

Monthly LA City Crime Data: https://data.lacity.org/A-Safe-City/Number-of-crimes-2010-today/rvrw-58iu

Monthly US GDP Data: https://ycharts.com/indicators/us_monthly_gdp#:~:text=US%20Monthly%20GDP%20is%20at,1.71%25%20from%20one%20year%20ago.

Monthly Rent CPI data: https://beta.bls.gov/dataViewer/view/timeseries/CUURS49ASEHA

Yearly Population Data: https://populationstat.com/united-states/los-angeles

Monthly LA City Temperature Data: https://www.ncdc.noaa.gov/cdo-web/datatools/lcd