Forecasting Chicago Crime Cases

ADSP 31006 | TMG Team

Haoyu Zhang
Marian Xu
Yifan Han
Liuyi Pan

Meet Our Team

Haoyu Zhang

Yifan Han

Liuyi Pan

Marian Xu

Agenda

01

Introduction

02

EDA

03

Data modeling

04

Future Steps

Business Problem

Problem Statement:
The city of Chicago has experienced fluctuating crime rates over the years, posing significant challenges to public safety and resource allocation. Accurate forecasting of crime trends is crucial for policymakers and law enforcement agencies to implement effective crime prevention strategies and allocate resources efficiently.

Objective:
The primary objective of this project is to develop a predictive model for forecasting the monthly total number of crime cases in Chicago.

Data Overview

Primary:

Chicago Crime Data
Jan 2001 – Mar 2023
(https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data)

Features:
ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location

Additional:

U.S. Unemployment Rate Data
Jan 2001 – Mar 2023
(https://data.bls.gov/cgi-bin/surveymost#/)

Features:
Date, Unemployment Rate

Data Processing

- **Initial Data Cleaning (Feature Engineering)**
  - **Selection of Key Columns:** Focus on essential columns: *Case Number(ID), Date, Primary Type*.
  - **Aggregation by Date:** Group data by month and calculate the total number of distinct crime cases and also different case types to monitor trends over time.

- **Data Splitting**
  - **Train Set:** January 2001 to December 2021
  - **Test Set:** January 2022 to December 2022

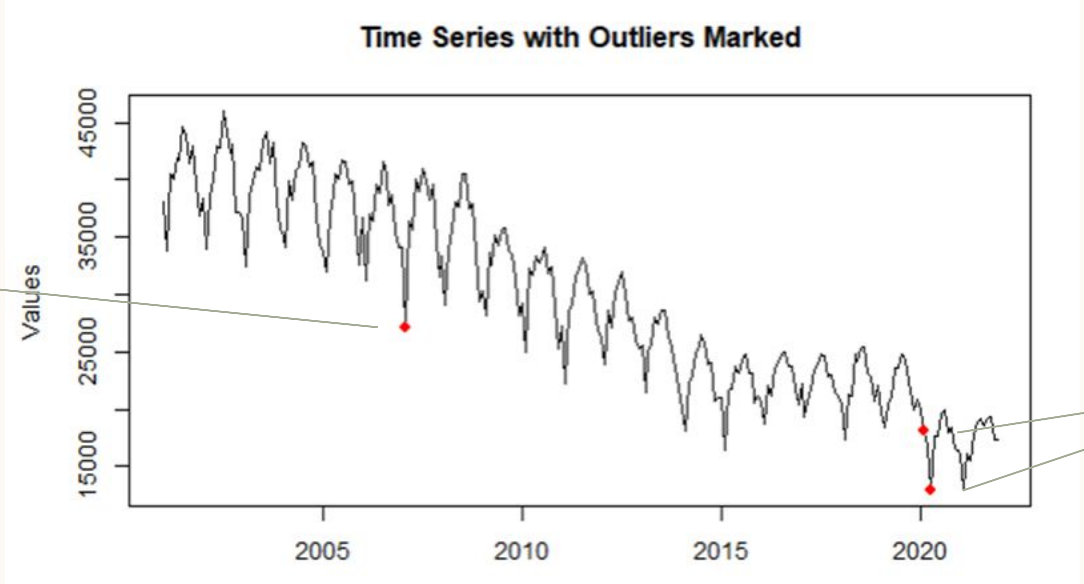
Date	Total Cases	Theft Cases	Burglary Cases	Motor vehicle Theft Cases	...(total of 33 types)
Jan-01	38119	7867	1934	2097	...
Feb-01	33784	6669	1666	1785	...
Mar-01	40565	7766	1832	2151	...
Apr-01	40088	7702	1932	2119	...
May-01	41836	8420	1997	2197	...

First Five Row of Cleaned Data

Data Processing

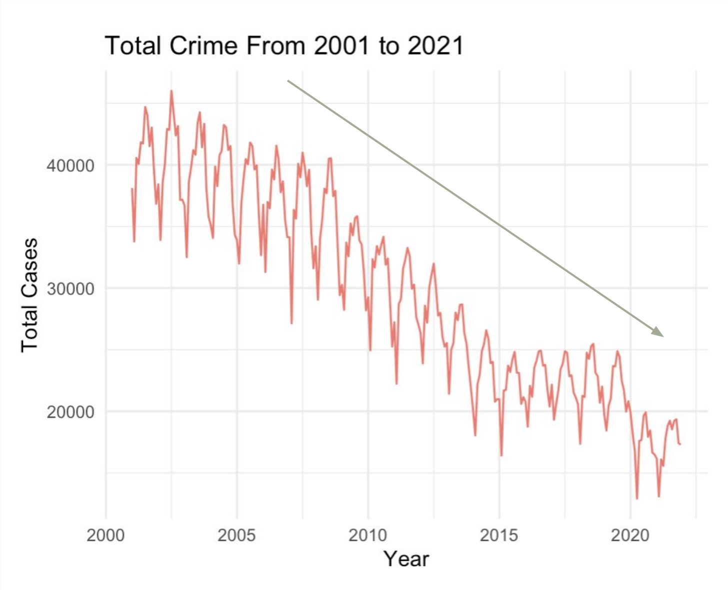
- **Anomaly Detection**
  - **Outliers:** opted to retain these outliers to preserve the complete integrity of the data.

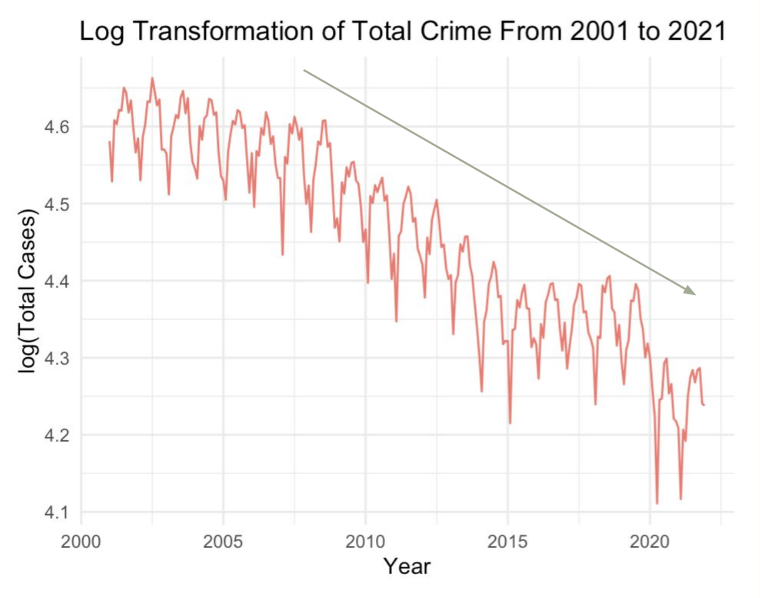
2007.2
Cold Winter Effect



2020.2, 2020.4
Cold Winter +
COVID-19 Effect

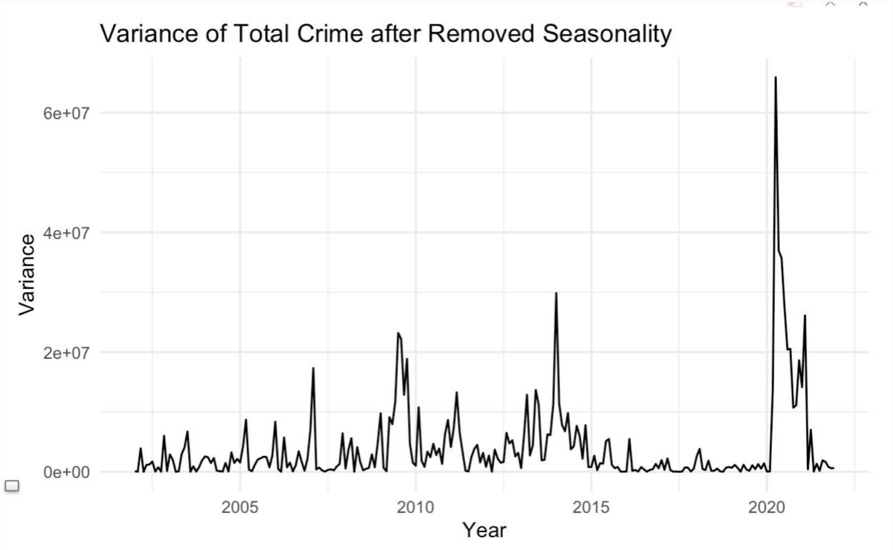
Crime Trend

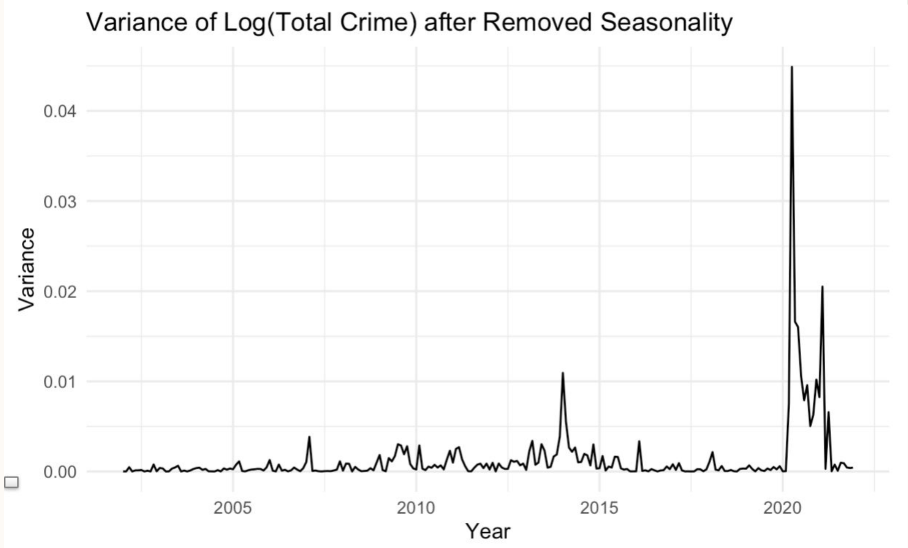




Trend and Seasonal Patterns: The time series data exhibits a clear overall downward trend from 2001 to 2021 with the presence of monthly seasonal patterns.
Variance Stabilization: The original total crime data shows unstable variance before and after 2015. Log transformation helps to stabilize it into the same level.

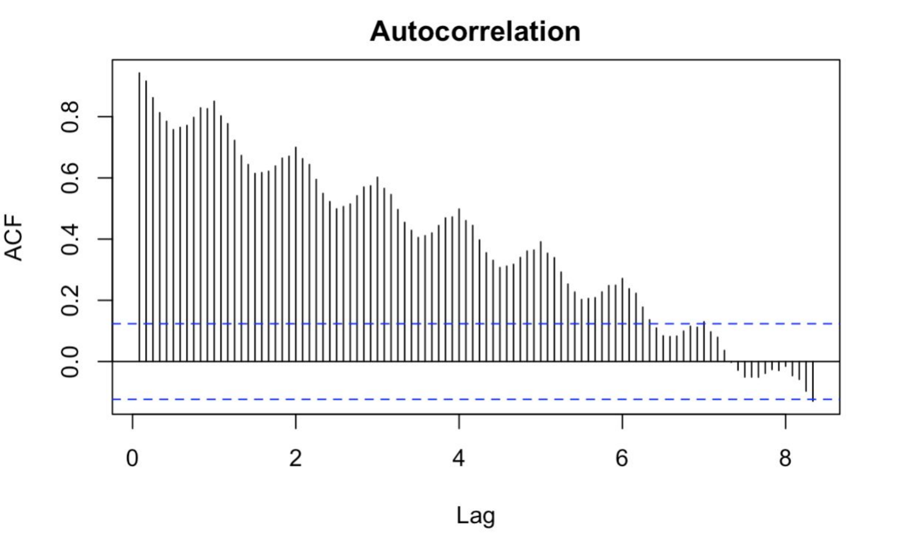
Variance of Total Crime

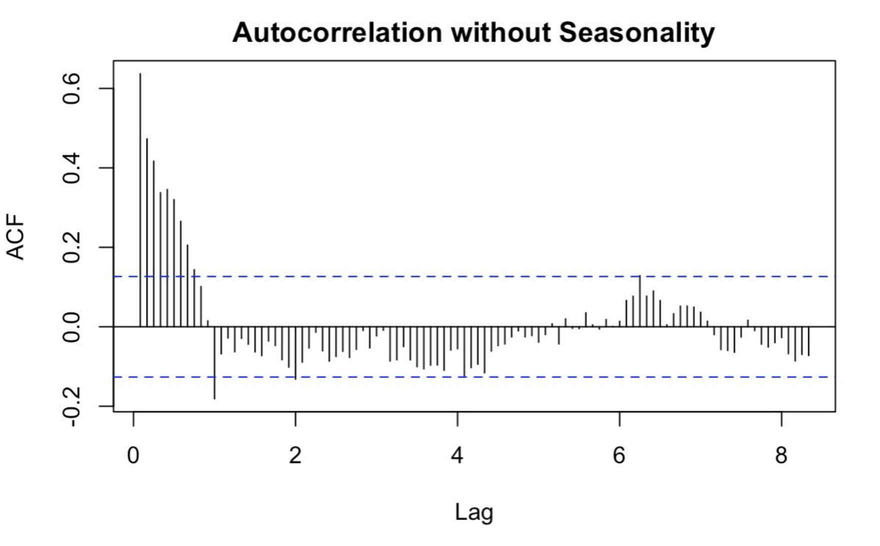




After applying a log transformation, the variance has stabilized significantly. However, some spikes still remain. The most notable spike occurred in 2020 due to the impact of COVID-19, which caused a sudden drop in the total number of crime cases.

Stationarity and Correlations





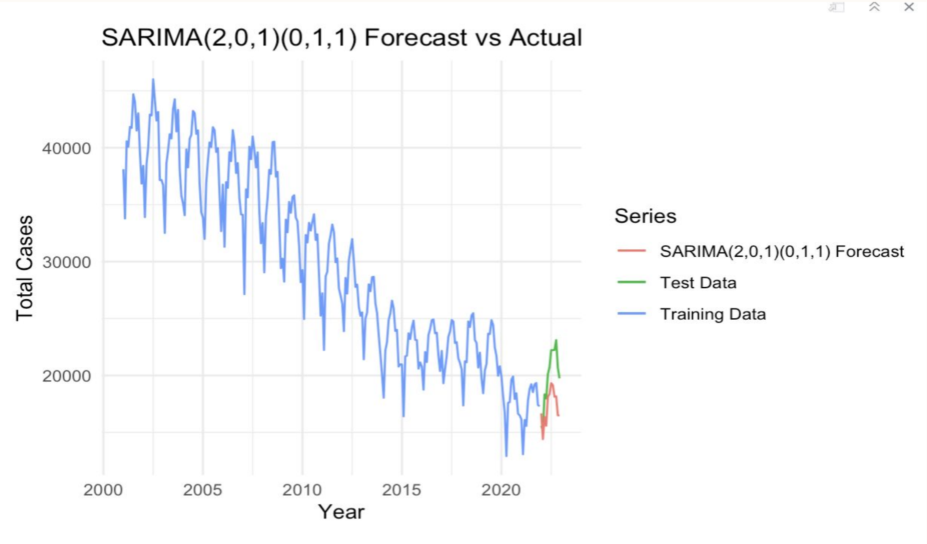
Total crime shows a clearly seasonality when look at the ACF plot.

After removed seasonality, ACF plot shows stationarity of the dataset, which also be proved by KPSS test.

Model Selection

1. Seasonal Autoregressive Integrated Moving Average (SARIMA) + Intervention
2. Exponential Smoothing (ETS)
3. Holt Winters
4. Hierarchical Time Series (HTS)
5. Bayesian Structural Time Series (BSTS)

SARIMA



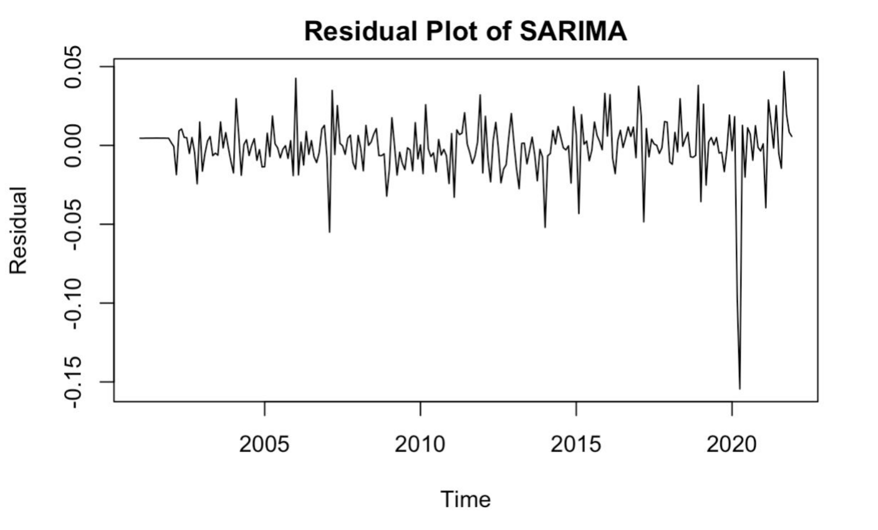
The forecasting is not good as expected. We will check the residual plot to identify any potential missing components.

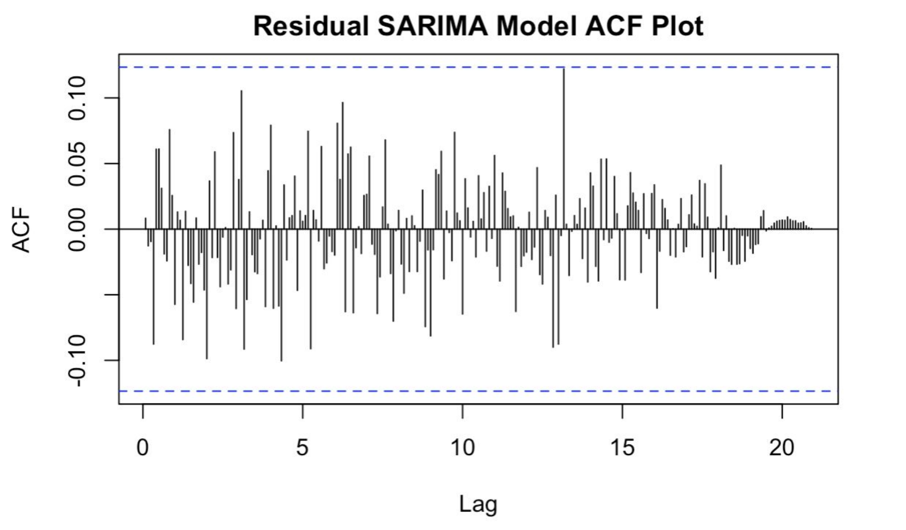
First Step: Use auto.arima as the baseline methodology to select SARIMA model, where **ARIMA(1,0,2)(0,1,1)[12] with drift** was selected.

Second Step: Iterate over p, q, P, Q parameters to select the best SARIMA model, where **ARIMA(2,0,1)(0,1,1)[12]** was selected.

RMSE	sMAPE
3004.72	0.1451

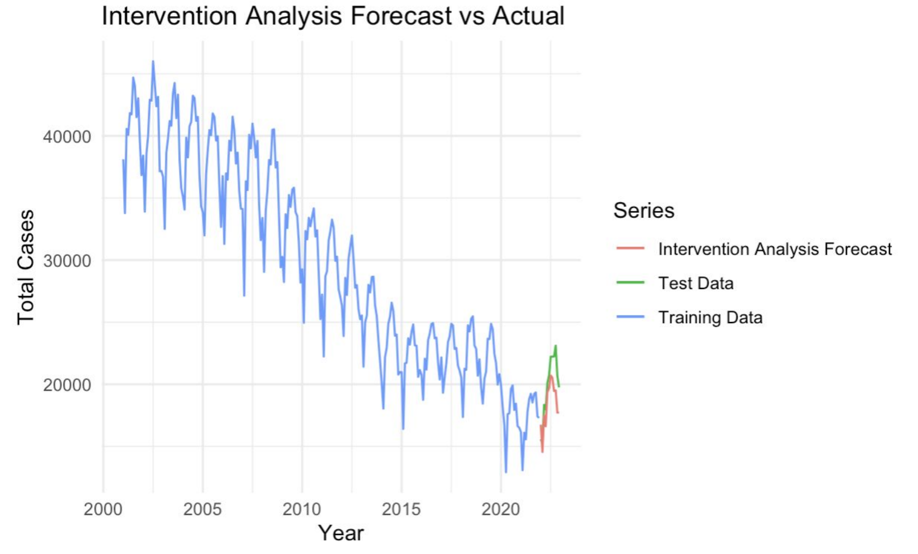
SARIMA





After checked the residual plot and its ACF plot, we can conclude that there is no autocorrelation in residuals, which is similar to white noise. However, a dramatic decline caused by COVID-19 suggests the need for an intervention analysis.

SARIMA + Intervention



- Use **ARIMA(2,0,1)(0,1,1)[12]** before and after the intervention.
- **Add pulse intervention** at the time Covid happened. **We assume Covid impact will not last forever.**

RMSE	sMAPE
1934.71	0.0875

Adding intervention term indeed helps crime predictions, which displays better performance compared to SARIMA.

Exponential Smoothing

Seasonal decomposition:
Step 1: Perform Seasonal Decomposition using Additive and Multiplicative Models
Step 2: Compute the standard deviation of the seasonal component from two decompositions
Step 3: Conclude that the type of seasonality is multiplicative

The conclusion helps decide to use ETS model with multiplicative seasonality

	Additive	**Multiplicative**
Sd.	2545.75	**0.09**

Exponential Smoothing

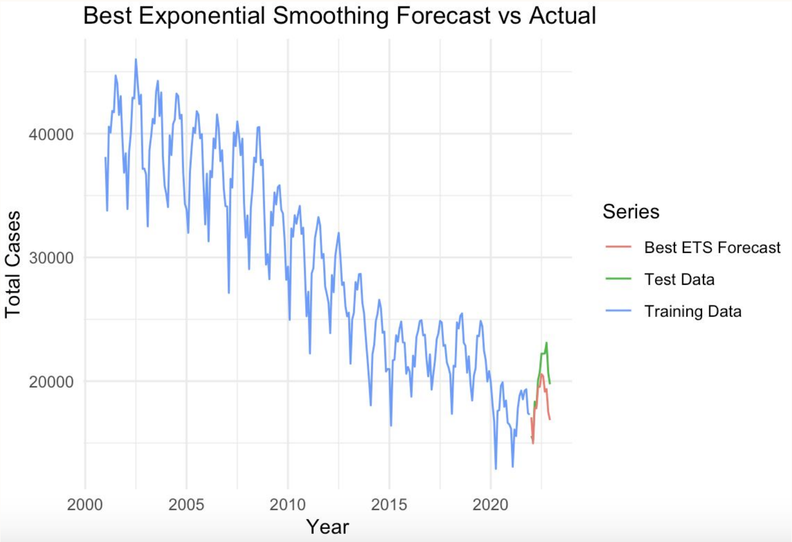
Considered models:

- **MAM:** Multiplicative errors, additive trend, and multiplicative seasonality.
- **MMM:** Multiplicative errors, Multiplicative trend, Multiplicative seasonality.

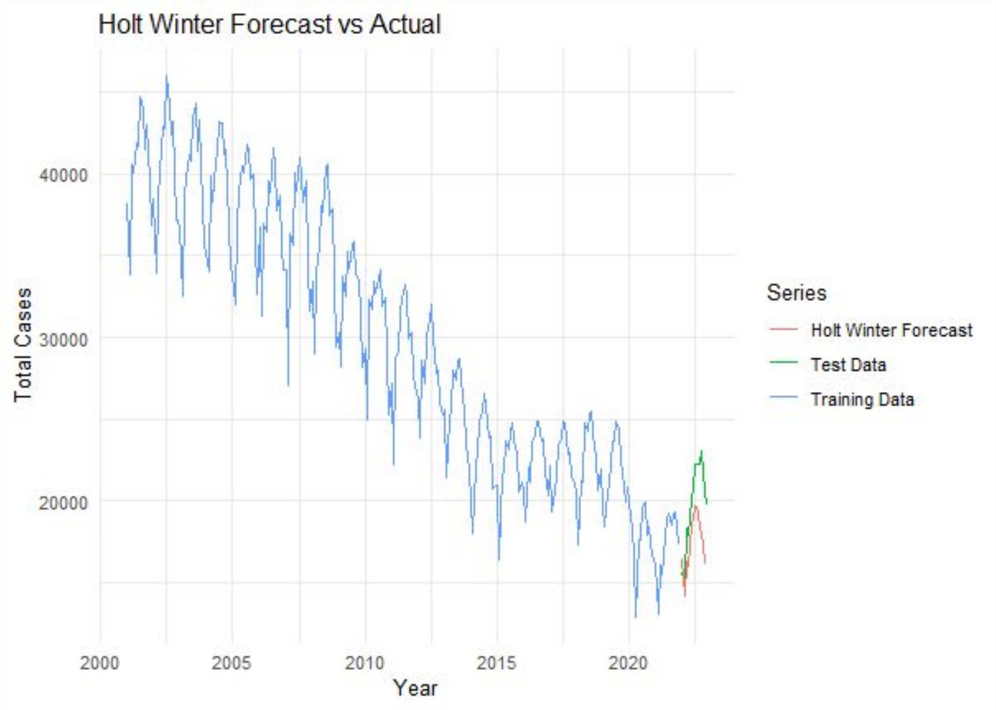
	RMSE	sMAPE
MAM	2096.47	0.0880
MMM	**2090.74**	**0.0872**

Final Model:

MMM model with lowest RMSE and MAE



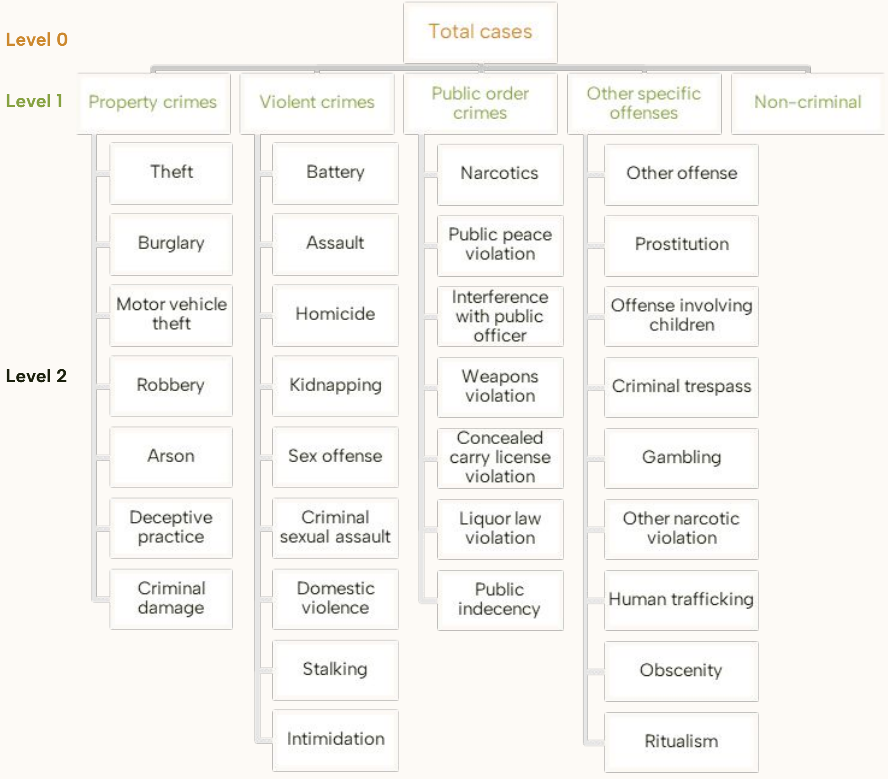
Holt Winters Multiplicative Seasonality with Trend



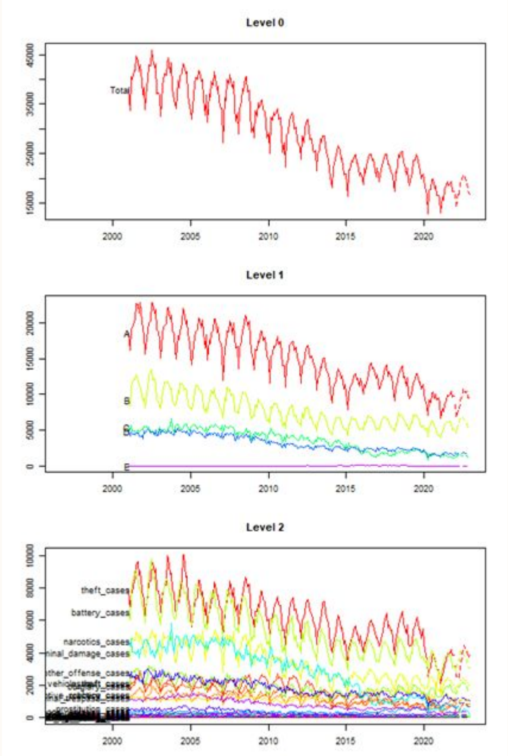
- Suitable for **non-stationary data**
- An extend of exponential smoothing to capture trend and seasonal
- The parameters **(alpha, beta, gamma) provide insights** into the level, trend, and seasonality

RMSE	sMAPE
2918.20	0.1379

Hierarchical Time Series (HTS)

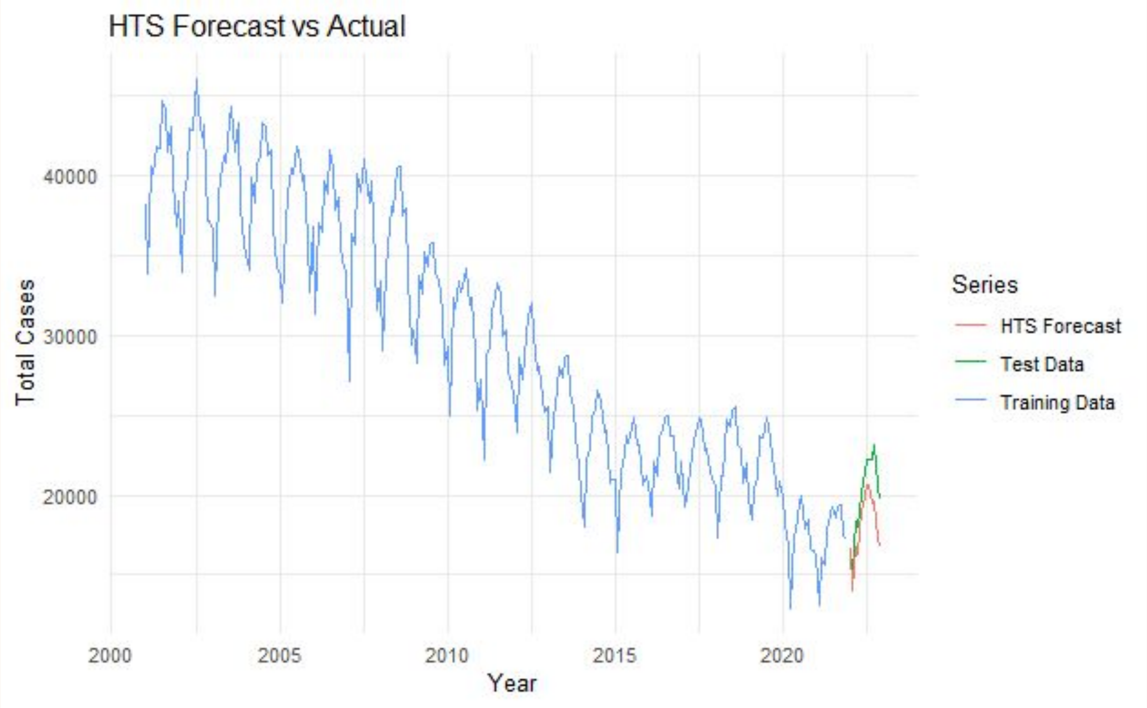


Tree structure: Hierarchy of time series data



- **Top level:** most aggregated (total cases)
- **Bottom level:** least aggregated (specific crime case types)
- Aggregate forecasts from **lower levels can refine the top-level forecasts and vice versa**

HTS – Forecasting



RMSE	sMAPE
2163.34	0.1037

* Allocate resources effectively to address crime in different districts and types of crimes.

Bayesian Structural Time Series (BSTS)

$$y_t = \underbrace{\mu_t}_{\text{trend}} + \underbrace{\gamma_t}_{\text{seasonal}} + \underbrace{\beta^T \mathbf{x}_t}_{\text{regression}} + \epsilon_t$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t$$

$$\delta_t = \delta_{t-1} + v_t$$

$$\gamma_t = -\sum_{s=1}^{S-1} \gamma_{t-s} + w_t$$

- **Flexibility and Transparency:** we can handle and visualize the underlying components, e.g. trend, seasonality, regression
- **Incorporating External Variables:** Allows the inclusion of external predictors like weather, holidays, and socioeconomic factors.
- **Handling Uncertainty:** Ability to incorporate uncertainty into our forecasts so we can quantify future risk.

BSTS – Model Comparison

State Components:

- **Model 1:** Trend + Monthly Seasonality
- **Model 2:** Trend + Monthly Seasonality + Regression (with unemployment rate)
- **Model 3:** Dynamic regression (with unemployment rate)

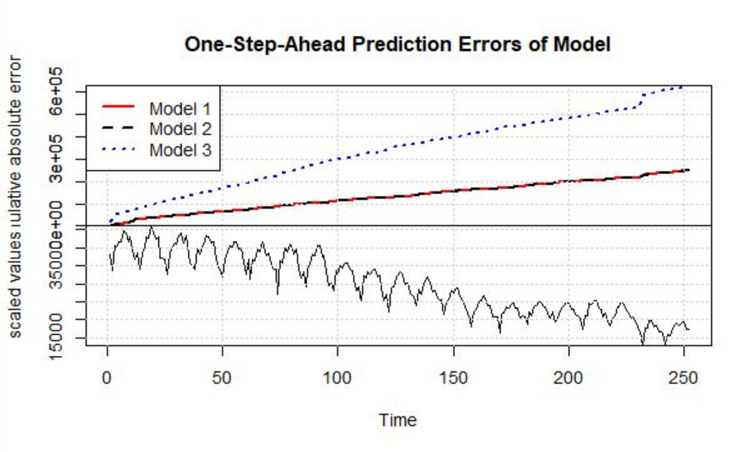


Table 1: Accuracy Comparison

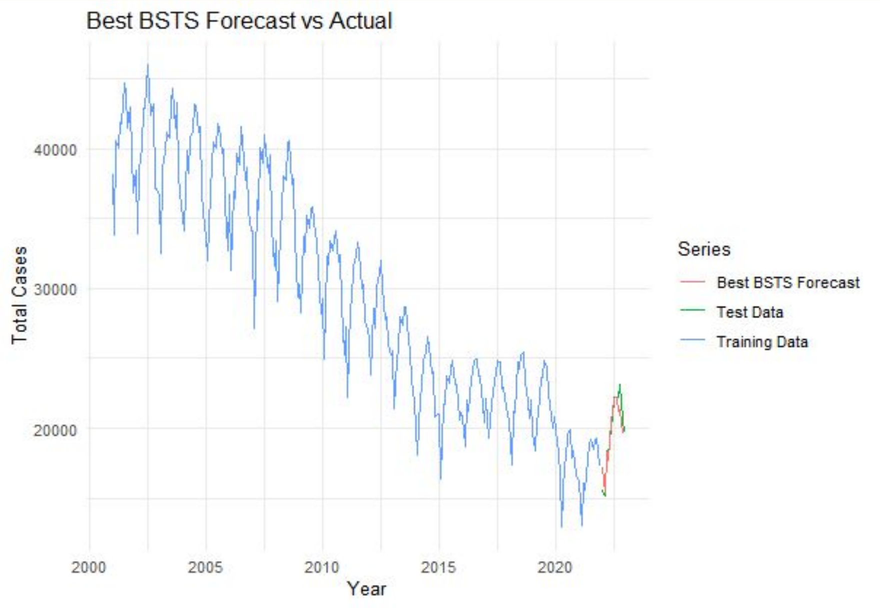
Model	RMSE	sMAPE
Model 1	1774.876	0.0732
Model 2	2188.139	0.0942
Model 3	2648.417	0.1249

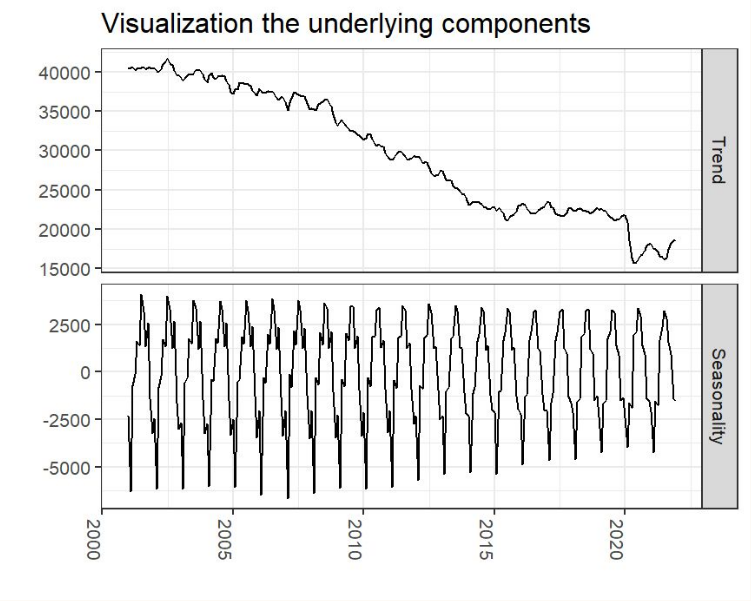
The additional predictor – the unemployment rate – used to fit Models 2 and 3 did **not** produce additional predictive accuracy.

Model 1 has the lowest RMSE and sMAPE.

BSTS – Forecasting

Final Model: BSTS with **Trend** and **Monthly Seasonality** State Specification





Model Comparison

	RMSE	sMAPE
SARIMA	3004.72	0.1451
SARIMA + Intervention	1934.71	0.0875
Exponential Smoothing	2090.74	0.0872
Holt Winters	2918.20	0.1379
HTS	2163.34	0.1037
BSTS	**1774.88**	**0.0732**

Future Steps

- Explore the inclusion of additional variables such as demographic data and other social indicators to further improve the forecasting accuracy.
- Gather additional historical data to enhance the robustness of future predictions.
- Explore and implement other forecasting models to compare performance and improve accuracy.

Thank you

Github

https://github.com/marian2216/ADSP31006-Crimes-in-Chicago.git