**FINAL REPORT**

**CAPSTONE PROJECT**

**TIME SERIES ANALYSIS**

**OF**

**POPULATION MOVEMENT IS COOK COUNTY, ILLINOIS USA**

**By:**

Nurul Hakimah Mohd Zaki

Date: 23 June 2024

**PROBLEM STATEMENT**

Traffic congestion in the US has gotten worse over the past year. According to Inrix, the average driver in the US spent time 29% more than in 2021, equivalent to $134 more spent on fuel in a year. This problem is expected as US population has steadily increased by 300 million people every year. By modelling the population movements, the pattern can be analyzed to determine the needs for significant changes that require attention by the town ship planner. Modelling and forecasting were done for Cook County, IL, where Chicago as the second most congested city in the world is located.

**ANALYSIS**

This dataset is mostly suitable to be analyzed using time series analysis.

**DATA**

Data was obtained from Kaggle.com, and Bureau of Transportation Statistics by the Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland.

The dataset has 22 columns of total no of trips and different frequency of trips taken in different counties in different states from 01 January 2019 until 13 May 2023. This data is recorded daily.

Upon checking null values, they are 2% of null value, that are actually comes from a summary rows. The data has combined total daily data into a state level, and leaving county details as null. Hence, it is safe to drop null values as they are redundant.

Number of Trips was chosen as the main dataset because I want to capture the whole population movements, regardless of the distance.

After loading the time series data, it is apparent visually that the series is not stationary.

This condition determines the approaches need to be taken.

**METRICS**

The stationarity test is using p-value.

The accuracy is measured by RMSE.

**STATISTICAL ANALYSIS**

In the beginning, a routine data cleaning, and visualization is done to understand the whole data.

Then, Number of Trips column is chosen to be the main dataset and analyzed. Time series data is plotted.

Dataset decomposition is done to check thoroughly on the trend, seasonality and residual components. It is apparent that all three components exist and need to be transformed, if it is not stationary.

Due to high magnitude in the dataset and high residual, log is performed before any transformation. This was confirmed in the first round of modelling and forecasting, all RMSEs are extremely high. The dataset was reevaluated and found that log method works best.

ADF test confirmed that data is not stationary, with p-value = 0.158.

Hence, 9 differencing methods was performed & found time shift period of 1 of log dataset gives the lowest p-value in ADF test.

|  |  |
| --- | --- |
| **Method** | **ADF p-value** |
| Initial Data | 1.94E-01 |
| Differencing d= 1 | 1.94E-18 |
| Differencing d = 7 | 8.00E-12 |
| Log | 1.58E-01 |
| SQRT | 1.78E-01 |
| CBRT | 1.72E-01 |
| Log & SQRT | 1.57E-01 |
| Log & CBRT | 1.57E-01 |
| **Log & Time Shift Period = 1** | **5.67E-18** |

Table 1: ADF p-value with corresponding transformation method.

ARIMA method is chosen because it is the most common model and needs differencing component.

As a start, train and test data is defined with 75:25 ratio.

For loop was performed to find the best hyperparameter combination with the lowest RMSE. The best parameters for ARIMA is p=5, d=1, q=5.

The first step is to use training data and model the prediction on test data.

The RSME for ARIMA Prediction on test data is 0.0635. This shows that ARIMA is a good model to make prediction on top of test data.

Next is to perform forecasting for 1 year ahead. The error is consider good.

Due to seasonality component, SARIMAX model is introduced to improve existing model.   
Same train : test ratio, p, d, and q parameters were used, in addition to s=28 to account for 28 days cycle.

The RSMEs for SARIMAX prediction and forecast also showing 0.625, showing good model.

**PROPOSED SOLUTION**

Necessary development such as increasing public transport coverage and frequencies, adding more highways or smart traffic lights shall be evaluated for future needs.

Notice that there is significant upward trend before a decrease in trends end of 2019, that was due to COVID-19 spread. It seems like the population will move and reach the peak. Keep in mind that the number of population is also increasing, so the peak could be higher and could be faster.   
Local and federal authority should make a more thorough development plan to avoid congestion problem.