Predicting Traffic Congestion with Machine Learning: A Historical Analysis of Traffic Data

Project Report

April 21st, 2023

Names

Gustave Munezero Bwirayesu

Yves Marie Ishimwe Kirunga

Deo Uwimpuhwe

Abstract

This project entitled ***"Predicting Traffic Congestion with Machine Learning: A Historical Analysis of Traffic Data***" aims to propose a mobile big data analytics approach that utilizes traffic data to predict congestion. This study uses data from 29 regional roads in Chicago to train time series models that can predict traffic in 10, 20 and 30 minutes ahead. Exploratory data analysis (EDA) was carried out to understand the data, and both univariate (ARIMA) and multivariate (VAR) time series-based machine learning algorithms were used. Feature selection approaches, such as mutual information and r_regression, were employed to identify features that are moderately correlated with speed. The models were evaluated using Root Mean Squared Error (RMSE). The study finds that the VAR model, which uses bus count, hour, and speed, outperformed the Auto-ARIMA model. The study recommends using the VAR model for speed forecasting, and the rolling technique for updating the model as new data becomes available. The study recommends the implementation of the traffic congestion forecasting system in Rwanda to optimize road usage and support decision making. It also proposes future research to develop an improved model that could increase performance and forecast traffic congestion while recommending less congested routes in real time. Briefly, the project demonstrates the potential for efficiently using existing transportation infrastructure and can be implemented in Rwanda.

I. Background & Problem Statement

The population of African cities, including Rwanda, is rising significantly which poses a problem on transportation. If the transportation infrastructure cannot keep up with demand, this could lead to increased traffic congestion especially in the densely populated cities [1]. Transportation plays an important role in the economic development of any country, mainly used to support daily business flow where people can move quickly between different places. However, traffic congestion is a major challenge that hinders people from fully benefiting from transportation. In Rwanda, the number of vehicles has been increasing. In 2011, it was reported that 250 automobiles were imported monthly, with 80% of them used in Kigali [2]. In 2021, the annual vehicle growth rate was around 12%, with most vehicles used in Kigali [3]. This surge in vehicles has surpassed the transport infrastructure, leading to severe traffic congestion, particularly during rush hours [2]. Furthermore, since the government of Rwanda changed the starting working time of government employees, traffic congestion has increased in the morning when people are going to work and earlier in the evening when people are going back to their homes. To mitigate this issue, the government encouraged citizens to adopt public transportation and started expanding road infrastructure by enlarging existing main roads and constructing bypass roads. In addition, public bicycle sharing programs have been introduced in some cities to reduce carbon emissions and to provide an alternative means of transportation during heavy traffic congestion [4]. Nevertheless, despite these measures, traffic congestion continues to increase. For these reasons, a technology-based solution is needed to forecast upcoming congestion in city roads and provide insights (trends and patterns) that can support decision making. This solution can help stakeholders and decision-makers plan and act based on the insights derived from the system analysis and help road users plan accordingly. To this end, this study aims to use traffic data to predict congestion in selected road branches.

The main goal of this project is to provide insights (patterns and trends) from historical data and accurate predictions of the traffic congestion using Chicago traffic congestion dataset to suggest how it can be adopted in Kigali to reduce congestion impact on road users and to support decision making.

II. Methods

This project uses historical data of Chicago traffic congestion by region from the Chicago data portal to forecast traffic congestion [5]. [https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Re/t2qc-9pjd]. The dataset covers 29 regions with different roads' information including geographical data and road traffic statuses. The regional traffic data were collected on a ten minute interval and 593118 historical data records from 1st of January 2020 to 1st of March 2023 for 4 different regions were retrieved. In the 29 regions comprising the dataset, 4 neighboring regions (Rogers Park- West Ridge, Far Northwest, North Park Albany Lincoln Sq, and EdgeWater Uptown) were selected to conduct congestion analysis and forecasting. The dataset has 17 features and most of them were providing spatial and descriptive information whereas the few remaining can be used for quantitative analysis. Among 17 features of the dataset, seven are considered relevant for analysis and modeling while the spatial information features were used for visualization of the selected regions as shown on the map in Figure 1. The remaining features were identification numbers for the records, regions and number of sensors which were not necessary for the modeling process. The seven selected features for analysis are time, region ID, estimated speed of vehicles, number of vehicles on road (bus count), hour of the day, day of the week, month and speed. The speed was considered as the measure of traffic congestion where the slower speed on the road indicates the presence of congestion.

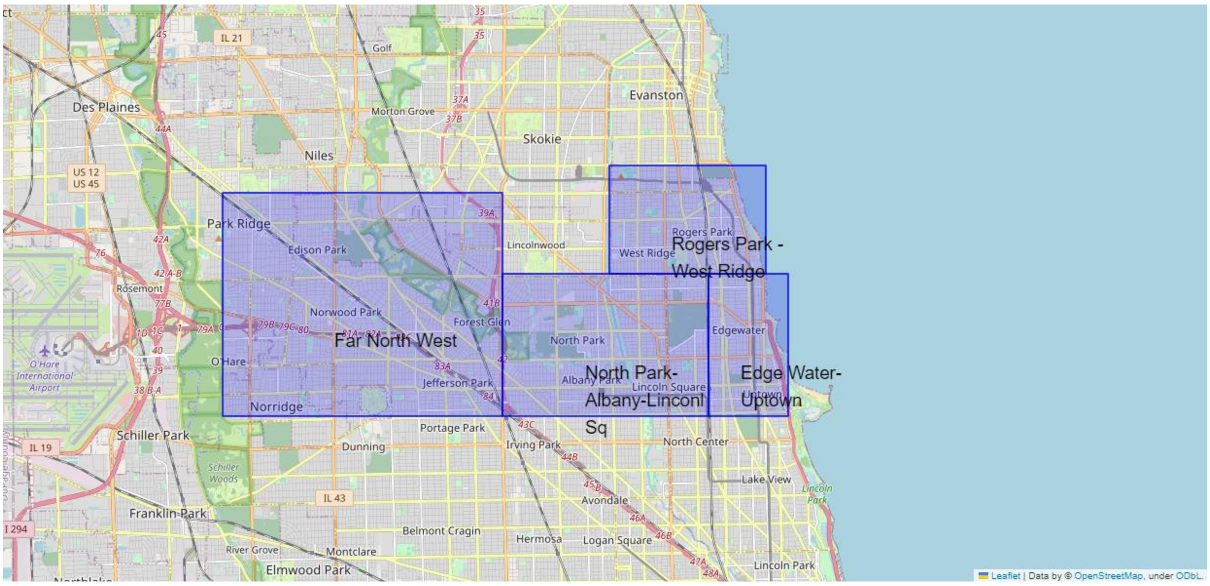


Figure 1: A map of the selected region which has to be analysed in this project

Exploratory data analysis (EDA) was conducted to understand the data and identify the patterns and trends in the data. During data exploration, various approaches such as correlogram, speed distribution, and speed stationarity among others were used to understand data. Speed distribution, summary statistics, correlation, speed autocorrelation, hourly and daily patterns were also computed for deep understanding of the data.

The correlation between variables showed that both the number of vehicles and hour of the day are more correlated with speed than others. Additionally, the speed distribution on each road was computed as shown in Figure 2, where the estimated speed for all streets were generally in the range of 15 and 40 miles per hour. It was found that the estimated speed rarely reaches 60 miles per hour as depicted in Figure 2.

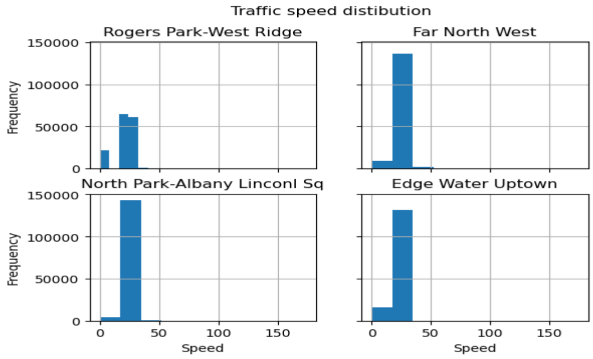


Figure 2:Traffic speed distribution

Average hourly speed for each load was also computed as shown in Figure 3, where the speed varies as time passes. It can be seen that for all roads, the speed noticeably decreases during the night from 1am to 3am. Additionally, there are signs of hourly patterns in speed which are almost similar for all roads and the peak speed occurs between 4am and 5am.

The speed autocorrelation was also computed to understand the extent at which past speed data can explain the current and future speed. As shown in Figure 4, the speed is moderately correlated with its delayed version during the first 3 to 4 lags for all roads which translates to 30 to 40 minutes because the data were recorded at 10-minute intervals.

Additionally, the stationarity of the speed time series for each region was tested using Augmented Dickey-Fuller (adfuller) test where the results showed that the speed time series for each selected region is stationary and there is mean reversion.

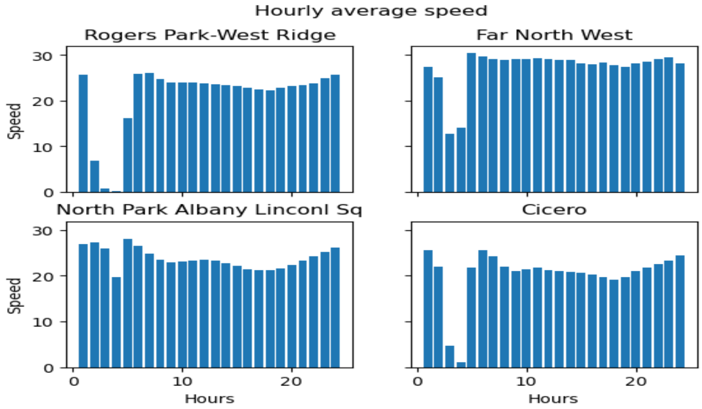


Figure 3:Daily speed profile

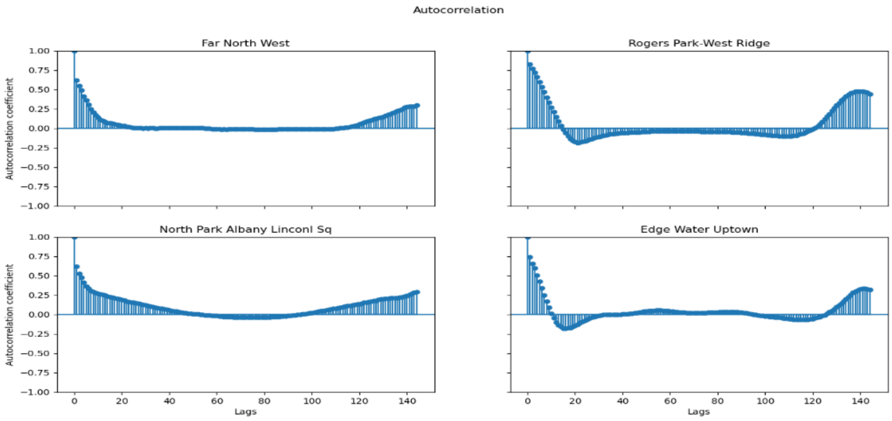
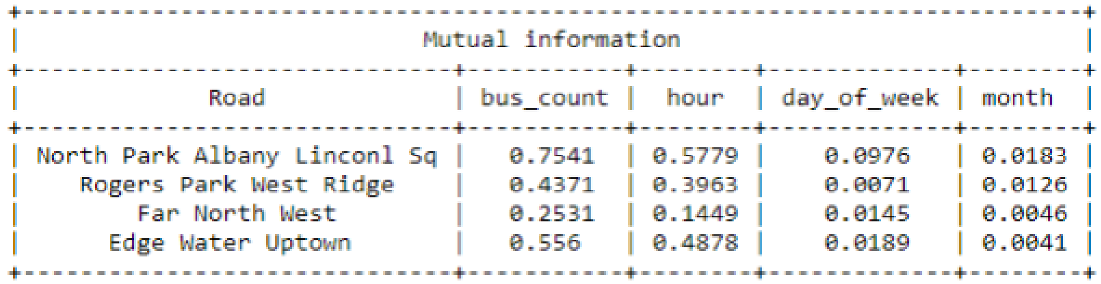


Figure 4: Speed autocorrelation

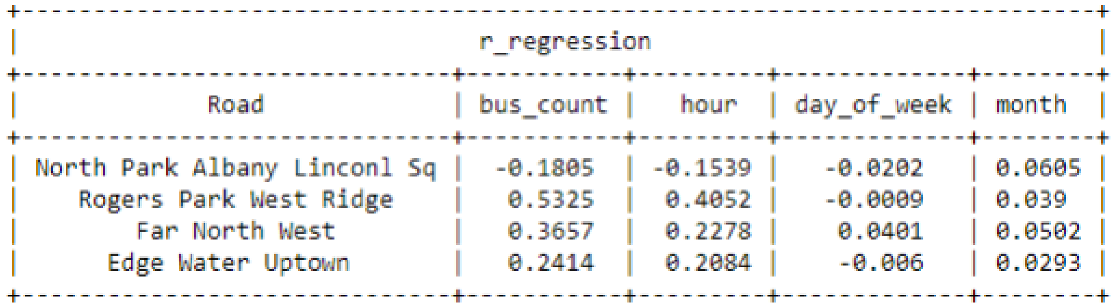
Mutual information [6] and r-regression [7] were used to select significant features and per table 1 and 2, only bus count and hour of the day were statistically significant on the speed level. Thus, the multivariate time series model was taken into consideration. Since the time series were found stationary using augmented Dickey Fuller test, VAR and ARIMA time series models were tried to predict speed in 10, 20, and 30 minutes ahead. VAR is a multivariate time series model which combines different time series to predict the future [8]. ARIMA on the other hand, is a univariate-based model that learns the variation of time series overtime and generalizes the future [9]. ARIMA is fitted using only speed time series while VAR uses speed, bus counts, and hour of the day to forecast speed and the both models were evaluated using Root Mean Squared Errors (RMSE).

Table 1: *Feature importance based mutual information approach*



```
+--------------------------------------------------------------------------+
|                           Mutual information                             |
+-----------------------------+----------+--------+-------------+--------+
|            Road             | bus_count |  hour  | day_of_week | month  |
+-----------------------------+----------+--------+-------------+--------+
| North Park Albany Linconl Sq |  0.7541  | 0.5779 |    0.0976   | 0.0183 |
|    Rogers Park West Ridge    |  0.4371  | 0.3963 |    0.0071   | 0.0126 |
|        Far North West        |  0.2531  | 0.1449 |    0.0145   | 0.0046 |
|      Edge Water Uptown       |  0.556   | 0.4878 |    0.0189   | 0.0041 |
+-----------------------------+----------+--------+-------------+--------+
```

Table 2: *Cross-correlation based on r-regression approach*

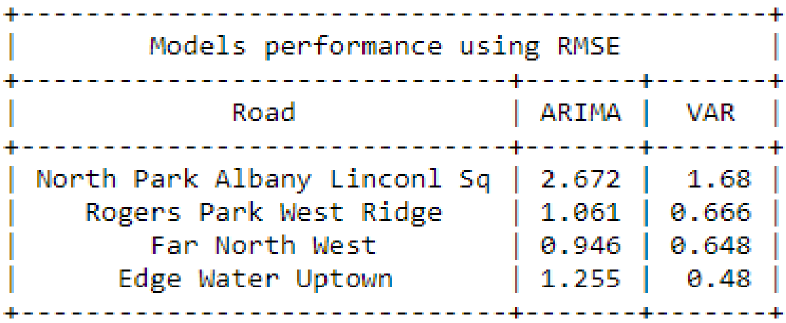


```
+--------------------------------------------------------------------------+
|                              r_regression                                |
+-----------------------------+-----------+---------+-------------+--------+
|            Road             | bus_count |  hour   | day_of_week | month  |
+-----------------------------+-----------+---------+-------------+--------+
| North Park Albany Linconl Sq |  -0.1805  | -0.1539 |   -0.0202   | 0.0605 |
|    Rogers Park West Ridge    |   0.5325  |  0.4052 |   -0.0009   | 0.039  |
|        Far North West        |   0.3657  |  0.2278 |    0.0401   | 0.0502 |
|      Edge Water Uptown       |   0.2414  |  0.2084 |    -0.006   | 0.0293 |
+-----------------------------+-----------+---------+-------------+--------+
```

III. Results and Discussion

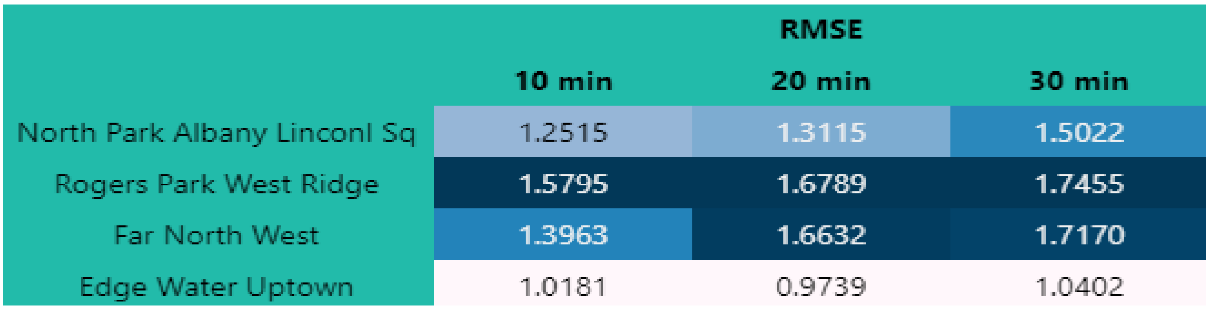
Two different time series forecasting models were examined for each of the selected regions. Auto-ARIMA and VAR models were compared to identify the optimal model for forecasting speed in each of four selected regional roads of Chicago. The results showed that VAR which uses bus count, hour and speed performed better than Auto-ARIMA as shown in Table 3. Therefore, VAR was considered for speed forecasting where it used a rolling technique where the model forecasts the speed in 10, 20, and 30 minutes ahead and then the model is updated as the new data becomes available. The resulting model performance using rolling method is presented in Table 4. These results show that the forecasting error tends to increase when the model tries to predict speed in the far future. Additionally, EdgeWater Uptown region has the smallest root mean squared error among all roads for all time frames whereas Rodgers Park West Ridge has the worst.

Table 3: *RMSE results for ARIMA and VAR models*



```
+---------------------------------------------+
|         Models performance using RMSE       |
+------------------------------+-------+-------+
|             Road             | ARIMA |  VAR  |
+------------------------------+-------+-------+
| North Park Albany Linconl Sq | 2.672 |  1.68 |
|    Rogers Park West Ridge    | 1.061 | 0.666 |
|        Far North West        | 0.946 | 0.648 |
|      Edge Water Uptown       | 1.255 |  0.48 |
+------------------------------+-------+-------+
```

Table 4: *RMSE results for vector auto regressive (VAR) model for speed forecasting.*



		RMSE	
	10 min	20 min	30 min
North Park Albany Linconl Sq	1.2515	1.3115	1.5022
Rogers Park West Ridge	1.5795	1.6789	1.7455
Far North West	1.3963	1.6632	1.7170
Edge Water Uptown	1.0181	0.9739	1.0402

IV. Reflection & Recommendations

Congestion forecasting is important to an efficient transportation system. Many governments, today, allocate a big amount of budget in expanding existing transportation infrastructure; but still car growth rate is high, especially in developing countries, thus, resulting in heavy road traffic. This depicts that road construction cannot solve traffic congestion issues alone because road users need to have knowledge of road status in the near future to optimize road usage as there might be severe traffic congestion in certain roads while their alternatives are less used. Therefore, implementation of the traffic congestion forecasting system, as developed in this project, could improve efficiency of using the existing transport infrastructure and thereby adding value on the efforts currently being put in roads upgrading. We would like to recommend this project to Rwanda transport development agency (RTDA) where it can be used as the starting point of building real time traffic congestion forecasting and road recommendation systems that can help optimize roads usage and support decision making.

Conclusion

This project utilized time series machine learning models (ARIMA and VAR) to forecast the speed on four regional roads in Chicago. Historical traffic data was utilized, and the results indicated that the VAR model provides superior results in comparison to the ARIMA model. Consequently, the VAR model was utilized for speed prediction, using a rolling method in which the model is updated when new data becomes available. As a result, this project can aid in the efficient utilization of existing transport infrastructure and facilitate future planning. The project was developed to explore the potential of collecting real-time traffic data and its implementation in Africa to address traffic congestion problems that are common in the cities of developing African countries, specifically Kigali, Rwanda.

References

[1] D. Kim, "Traffic Congestion," Story maps, 10 December 2019. [Online]. [Accessed 20 January 2023].

[2] J. Karuhanda, "KIGALI - Traffic Boss, Chief Supt Vincent Sano, has said that the increasing number of vehicles is likely to cause congestion, especially in urban areas like Kigali city.On average, 250 automobiles come into the country every month," The new times, March 2011. [Online]. Available: https://www.newtimes.co.rw/article/51912/National/officials-aconcerneda-over-surgingnumber-of-vehicles. [Accessed 20 September 2021].

[3] Mininfra, "STRATEGIC PAPER ON ELECTRIC MOBILITY ADAPTATION IN RWANDA," REPUBLIC OF RWANDA-MINISTRY OF INFRASTRUCTURE, Kigali, 2021.

[4] R. O. Oduku, "Rwanda Launches Africa’s First Public Bike-Share Transport System," SISi AFRIKA MAGAZINE, 20 September 2021. [Online]. Available: https://www.sisiafrika.com/rwandalaunches-africas-first-public-bike-share-transport-system/. [Accessed 20 February 2023].

[5] C. D. Portal, "Chicago Traffic Tracker - Congestion Estimates by Regions," Chicago Data Portal, [Online]. Available: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Re/t2qc-9pjd. [Accessed 01 March 2023].

[6] "sklearn.feature_selection.mutual_info_regression," Scikitlearn, [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html. [Accessed 21 April 2023].

[7] "sklearn.linear_model.LinearRegression," Scikitlearn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. [Accessed 21 April 2023].

[8] "Multivariate Time Series Analysis With Python for Forecasting and Modeling," Analytics Vidhya, 13 April 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/. [Accessed 03 April 2023].

[9] "Univariate Time Series Analysis and Forecasting with ARIMA/SARIMA," Section, 11 May 2022. [Online]. Available: https://www.section.io/engineering-education/univariate-time-series-analysis-with-arima-in-python/. [Accessed 01 April 2023].