ESTIMATING & COMPARING PUBLIC TRANSPORT EMISSION USING GTFS2EMIS 'R' PACKAGE

Prayas Baliyan

DEPARTMENT OF DATA ANALYTICS OF BUSINESS

ST. CLAIR COLLEGE, WINDSOR, ONTARIO, CANADA, N9A5K3

EMAIL: W0790447@MYSCC.CA

Hemant Chowdary Gorantla

DEPARTMENT OF DATA ANALYTICS OF BUSINESS

ST. CLAIR COLLEGE, WINDSOR, ONTARIO, CANADA, N9A5K3

EMAIL: W0788804@MYSCC.CA

Nishi Shrivastava

DEPARTMENT OF DATA ANALYTICS OF BUSINESS

ST. CLAIR COLLEGE, WINDSOR, ONTARIO, CANADA, N9A5K3

EMAIL: W0770047@MYSCC.CA

Sai Krishna Yarlagadda

DEPARTMENT OF DATA ANALYTICS OF BUSINESS

ST. CLAIR COLLEGE, WINDSOR, ONTARIO, CANADA, N9A5K3

EMAIL: w0789428@MYSCC.CA

Eswar Kiran Pathuri

DEPARTMENT OF DATA ANALYTICS OF BUSINESS

ST. CLAIR COLLEGE, WINDSOR, ONTARIO, CANADA, N9A5K3

EMAIL: W0788366@MYSCC.CA

Supervised By:

Prof. Umair Durrani

St. Clair College

GitHub Repository - https://rb.gy/g6tym

ABSTRACT

This study analyzes public transit bus emissions using General Transit Feed Specification (GTFS) (João Pedro Bazzo, Rafael H. M. Pereira,Pedro R. Andrade, 2022) data and three machine learning models: Random Forest Regressor, Facebook Prophet Model, and Time series using ARIMA. The study aims to show the main pollutants emitted by public transit buses, their highest and lowest emissions from 2010 to 2022, pollutants emissions at different speed intervals, and the fuel types that emit major pollutants. The study reveals that EC was the main pollutant emitted, followed by CO2. Emissions remained constant from 2010 to 2014, sharply increased in 2015, peaked in 2018, and gradually declined until 2020. Emissions were highest for buses built between 1980 and 1990 and decreased thereafter. Emissions were highest at lower speed intervals (0-8 miles/hour) and lowest at upper speed intervals (65-72 miles/hour). EC and CO2 had the highest emission levels for all fuel types of transit buses. The Random Forest model had high accuracy of 0.96 and 0.99 for test and train data, respectively. The Facebook Prophet model predicts emissions ranging from 0 to 3 gm/mile for the years 2023 to 2027, while the Time Series/Forecasting using ARIMA model predicts emissions ranging from 1.0 to 1.15 grams/mile for the same years. These results can help policymakers and public transit operators make informed decisions to reduce emissions and improve air quality. Overall, this study supplies a comprehensive analysis of public transit bus emissions and their trends using machine learning models.

TABLE OF CONTENTS

List of Figures

1. INTRODUCTION

The excessive reliance on buses for transportation has led to lot of adverse effects, such as traffic congestion, unfavorable health outcomes, increased frequency of accidents, air pollution, smog, and the emission of greenhouse gases. (Santos, G., Behrendt, H., Maconi, L., Shirvani, T., & Teytelboym, A. (2010a). Part I: Externalities). This study adds to the existing research on emissions associated with public transit by developing a time series forecasting model for transit bus emissions at the urban regional level. Although public transit is a sustainable transportation mode, emissions can vary significantly due to a range of factors, such as passenger load, fuel type, service type, vehicle configuration, type, and age. (Alam, A., Diab, E., El-Geneidy, A. M., & Hatzopoulou, M. (2014). A simulation of transit bus). The transit bus emissions are primarily a result of fuel combustion during vehicle operation, including fuel combustion during passenger transport between destinations (via bus stops), fuel combustion while idling (for passenger boarding and alighting), and fuel combustion during empty vehicle transport (such as driving from the bus depot to the bus route starting point). However, there is limited research that tries to quantify and understand the emissions produced by the entire bus transit system. To address this gap, this study aims to enhance the method for evaluating bus transit emissions by conducting a disaggregated level analysis of bus transit emissions.

The combustion process in automobiles, buses, and trucks generates traffic-related air pollution, which includes particulate matter, nitrogen oxides (NOx), volatile organic compounds (VOC), and other pollutants. Infinite studies conducted in recent years have proven a connection between exposure to these pollutants and various chronic and acute health effects. (Brauer); (Gan); (Selander, J., Nilsson, M. E., Bluhm, G., Rosenlund, M., Lindqvist, M., Nise, G., & Pershagen, G.). Transportation and health research communities have found how to enhance urban air quality. To achieve this goal, it is crucial to accurately quantify emissions. In recent years, the transportation sector has made considerable progress in developing quantitative frameworks that enable the estimation of disaggregated automobile emissions, while considering travel patterns, vehicle characteristics (such as age, type, and fuel type), and land use patterns that change automobile emissions. (Beckx, C., Panis, L. I., Janssens, D., & Wets, G. (2010). Applying activity-travel data for the); (Sider); (Sider, T. M., Alam, A., Farrell, W., Hatzopoulou, M., & Eluru, N. (2014). Evaluating vehicular); (Dons).

Typically, studies that examine emissions for large transportation networks rely on large-scale inventories that are based on traffic simulations or GPS data of vehicles. While it can generate traffic emission inventories for large transportation systems, they concentrate solely on private vehicles and overlook public transportation networks. Studies have investigated bus emissions in urban areas, focusing on bus fleet and fuel choices for cities like London, Hong Kong, and Macau (Liu, n.d.); (Shan, n.d.); (Chan, n.d.). However, these studies estimate bus emissions at a general level and are unlikely to supply individual-level emissions data. Therefore, the research on bus emission estimation lags the methods used to estimate emissions from private vehicles.

In this project, we used Machine learning Models. We have used latest development in building models including Random Forest Regressor, Facebook Prophet Model, and Time series using ARIMA The project is hoping to answer the following questions:

- Main Pollutants.
- Highest & Lowest Emissions from transit buses from 2010 to 2022

- Pollutants Emission at Upper and Lower (0-8) speed interval.
- Fuel (CNG, Gas, Diesel) Emitting major pollutants.
- Predicting future emissions.

2. DATA DESCRIPTION & PRE-PROCESSING

In this study we use General Transit Feed Specification (GTFS) data from a R package called gtfs2emis, to estimate public transport emissions. The gtfs2emis package uses emission factor models for Europe (EMEP/EEA), the United States (EPA/CARB and MOVES/EPA), and Brazil (CETESB) to calculate more than sixteen types of pollutants released by each vehicle. Fleet characteristics required by each emission factor models includes information for buses, including age, fuel, EURO standard, technology, load, slope, and whether they are Micro, Standard, or Articulated. However, the required characteristics may vary depending on the emission factor model. (João Pedro Bazzo, Rafael H. M. Pereira,Pedro R. Andrade, 2022)

The package requires two main inputs: First is the **transport model**, which transform a GTFS data input into a table of trajectory data comparable to GPS recordings that track the spatial and temporal location. Input data includes information on routes, stops, trips, time periods and schedules. In addition, provides estimates of multiple pollutants with high precision in both spatial and temporal dimensions. Second is the **emission model**, which involves approximating the emissions of pollutants by vehicles at every time and road segment by using information from the transportation model, the user's fleet characteristics, and the emission factors contained within the gtfs2emis package. The emission data output allows for detailed analysis of how emission levels differ based on fleet characteristics, as well as variations across time and space. (Rafael H.M. Pereira[1], João Pedro Bazzo Vieira[1],Pedro R. Andrade[2], 2023)

This study uses Detroit & California data which follows MOVES/EPA ((Technical, n.d.) and EMFAC emission factor model. The data has undergone changes, including the removal of the "source_type" and "id_speed" columns. We added a new column called "average_speed_interval" to the data. We filtered records where the speed interval is less than or equal to 68.4 miles per hour. In addition, we applied a specific transformation technique to the emission factor columns of the data, resulting in the creation of a new column named "log_ef" to store the transformed values.

3. METHODS

In this study, we examine the environmental impact of public transportation and explore ways to reduce emissions and promote sustainable transportation. We find air pollutants by analyzing the data. The focus of the analysis is to find which vehicles emits pollutants at lower and upper-speed intervals. By doing so, we gained a more comprehensive understanding of the pollutants in the city and generate insights that can inform policy decisions aimed at reducing emissions and improving air quality. This approach allows us to concentrate on specific vehicle types and speed intervals that are significant contributors to air pollution, which can help prioritize efforts to reduce emissions in these areas.

In this study, we used data from Toronto to conduct our analysis, and we gained insights into the emission patterns and trends of different pollutants in the city. Our findings showed that the average emission factor by pollutant shows that carbon dioxide ($CO2$) has the highest emissions, followed by carbon monoxide (CO) and nitrogen oxides (NOx). Furthermore, we noticed a significant decline in emissions in

2016, followed by a surge in 2017, and then another decline in 2018. These observations suggest that the city of Toronto may need to implement more robust emissions reduction strategies to address the fluctuation in emission levels over time. To mitigate the impact of fluctuating emission levels over time, we implemented a log transformation of the emission factor data. By taking the logarithm of the emission factor values, we were able to normalize the data and reduce the influence of outliers or extreme values, resulting in a more stable representation of emission trends over time. The log-transformed emission factor data can supply more reliable insights into the underlying patterns and drivers of emissions, which can inform the development of targeted and effective emissions reduction strategies for the city of Toronto.

To help us with evaluating the average emissions based on reference year, emission at upper and lower speed interval, major pollutants, and fuel types, we have created explanatory visualization for each. In addition, we aim to predict future emissions/forecast using three different models: Random Forest Regressor, Facebook Prophet, and Time Series ARIMA.

4. EXPLANATORY DATA VISUALIZATION

4.1 POLLUTANTS EMITTED AT HIGHEST LEVEL

Bar plot showing the average emission level (in grams per miles) of various pollutants emitted by transit buses. The x-axis shows the different pollutants, and the y-axis is the average emission level for each pollutant. We sorted bar plot in descending order, pollutant with the highest average emission level at the top. With 16.52 g/miles of emission, pollutant EC (Energy Consumption) accounted for the majority with pollutant CO2 being the second with 6.90 g/miles and other pollutants were under 2 g/miles. (See Figure 1)

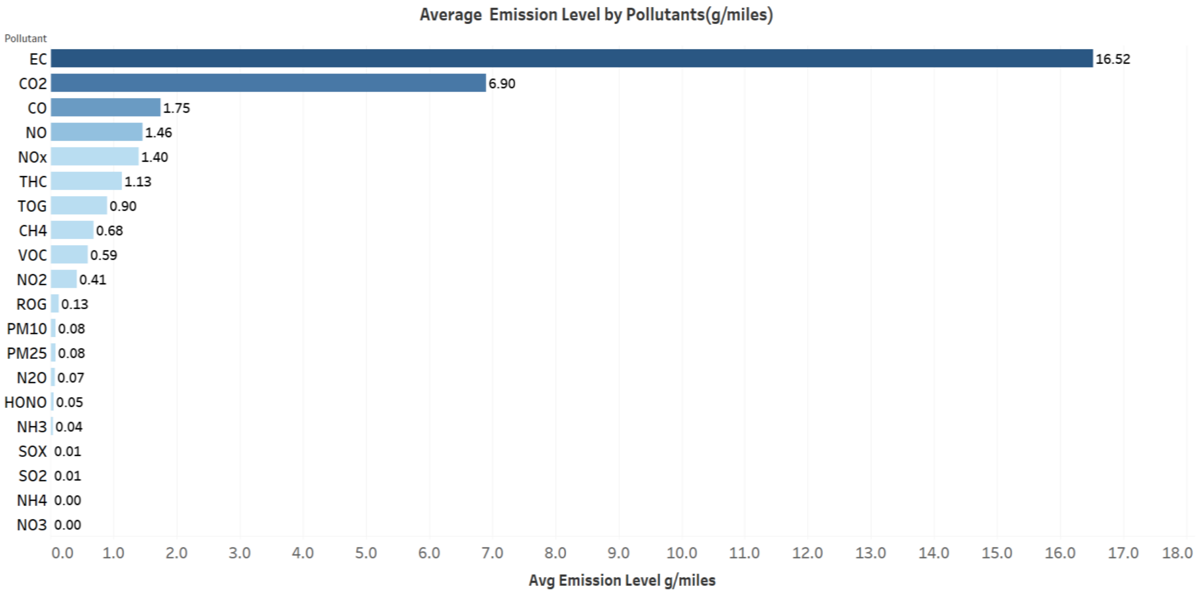

FIGURE 1

4.2 AVERAGE EMISSIONS FROM 2010 TO 2022

The plot shows the trend of average emissions for different reference years from 2010 to 2022. The y-axis is the average emission levels, while the x-axis stands for the reference year. The plot is a line chart with markers standing for the average emission values for each reference year. From 2010 to 2014, emission levels stayed constant, and then increased suddenly in 2015, reaching its highest level in the reference year 2018, 1.92 g/mile. The emission level then gradually declined until 2020, then remained constant with 1.88 g/miles in 2023. (See Figure 2)

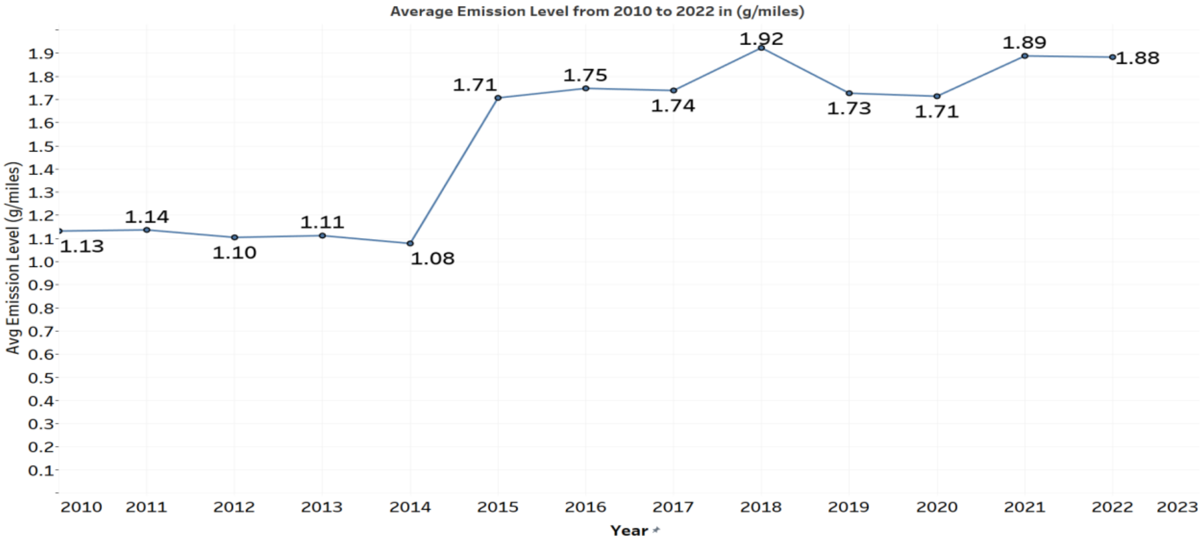

FIGURE 2

4.3 AVERAGE EMISSIONS FOR DIFFERENT MODEL YEAR

The plot shows the trend in average emissions by model year for different fuel types. The y-axis stands for the average emission levels, while the x-axis is the model year. The plot is a line chart with markers standing for the average emission values for each model year. With 0.96 g/miles for the 1980 model transit bus, emissions increased steadily until 1990, when they reached their peak of 2.21 g/miles. From then on, the

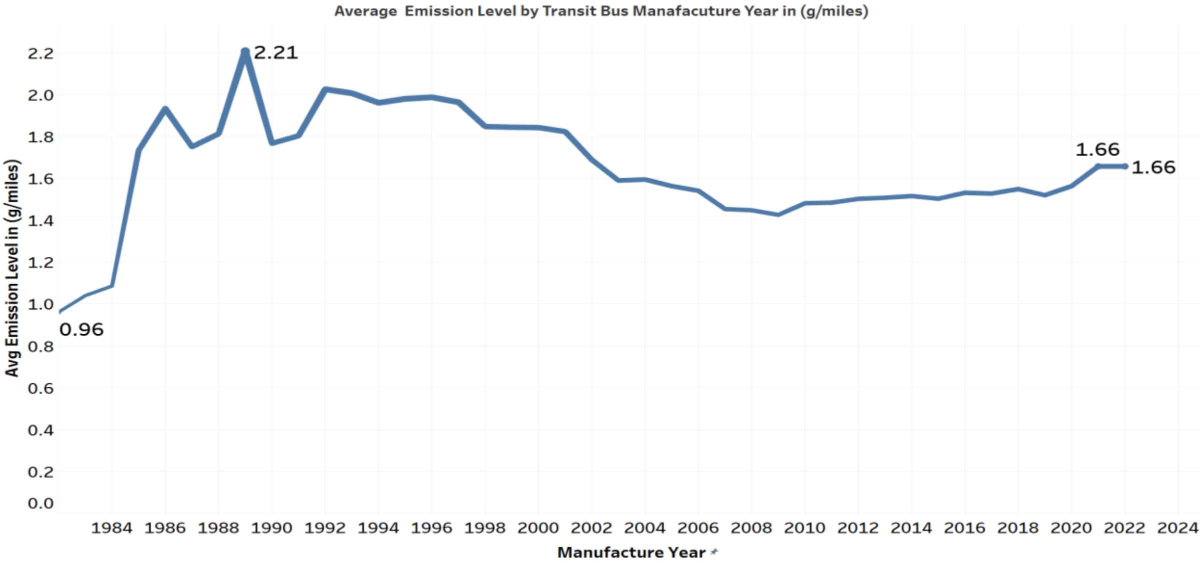

FIGURE 3

emission levels declined and remained constant for transit buses. (See Figure 3)

4.4 EMISSIONS LEVEL AT UPPER & LOWER SPEED INTERVAL

With the X axis showing the upper speed interval, the Y axis showing the lower speed interval, and the middle part showing the emission levels with colors, the heat map depicts the average emission levels at different speed intervals. In transit buses, emissions were at their highest between 0 and 8 miles with 1.35 g/miles, while they were at their lowest between 65 and 72 miles with 1,0 g/miles and emission levels range from 1.05 to 1.30 g/miles at other speed intervals. (See Figure 4)

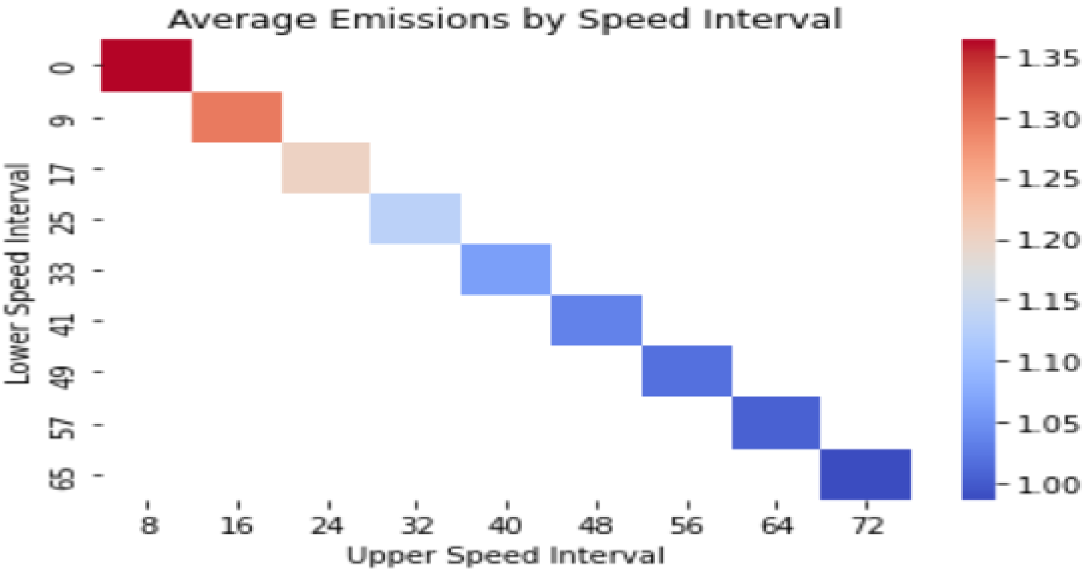

FIGURE 4

4.5 EMISSION FROM DIFFERENT FUEL TYPES

The plot shows the top five pollutants by emission factor for each fuel type. The x-axis is the different fuel types, and the y-axis is the emission factor measured in g/miles. Pollutant EC has the highest emission level of more than 16 g/miles for all fuel types of transit buses, followed by CO2, with more than 6.80 g/miles. For all fuel types of transit buses, the emission levels are the same. (See Figure 5)

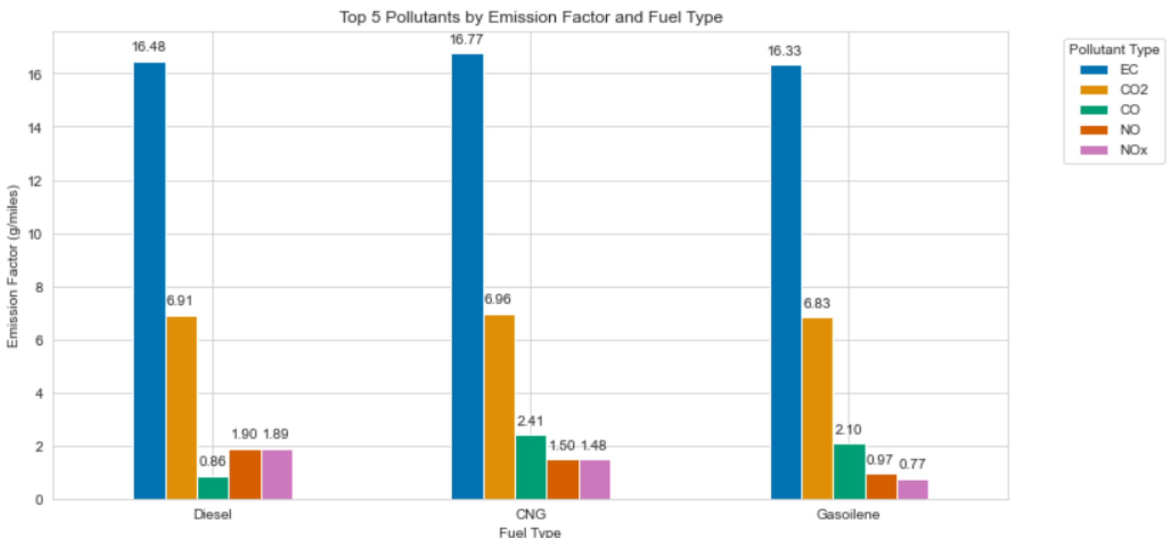

FIGURE 5

5. MACHINE LEARNING MODELS

5.1 RANDOM FOREST MODEL

This Study uses Random Forest Regressor machine learning model for the analysis to predict future emissions. We combined the USA EMFAC (EMFAC, 2014) and USA MOVES data into a single data. We split the combined dataset into training and test sets based on the reference year, where the training set includes all the data before 2020 and the test set includes all the data from 2020 onwards. We use features variables, i.e., reference year, pollutants, fuel types i.e., Gasoline, CNG, Diesel lag, and emission factor as target variable, which is crucial for capturing the complexity of emission patterns.

We created a lag column for the emission factor in train and test data to incorporate the effect of past values of the variable on its current value. Creating a lag column for a variable means shifting its values by a certain number of time periods. By introducing a lag column for the emission factor, we investigated whether they relate the emission factor at a time to its values in earlier time periods. It is useful for modeling time series data, as it helps to find patterns and seasonality in the data. By including lagged variables in the model, we reduced autocorrelation and enhance the model's accuracy.

In addition to Random Forest Regressor model, this study also incorporated Lasso regularization technique. Lasso is a regularization method that can help prevent overfitting in models by adding a penalty term to the objective function. After combining Lasso regularization with the Random Forest model, we improved the accuracy and generalization ability of the model. This approach allowed us to strengthen prediction based on test and train data. It created a more comprehensive model that can supply more correct predictions about future emissions.

The scatterplot illustrates for test (See Figure 6) and train (See Figure 7) data the actual and predicted values of emissions from a random forest model. The X-axis displays the actual values of emissions, while the Y-axis stands for the predicted values of emissions. The orange points show the actual values, while the blue points are the predicted values. The proximity of the points to the diagonal line shows the accuracy of the model. Points closer to the line implies that the model is more precise in predicting the emissions. Blue points consistently fall below the orange points shows that the model under-predicts the emissions. Conversely, blue points are consistently higher, it suggests that the model over-predicts the emissions.

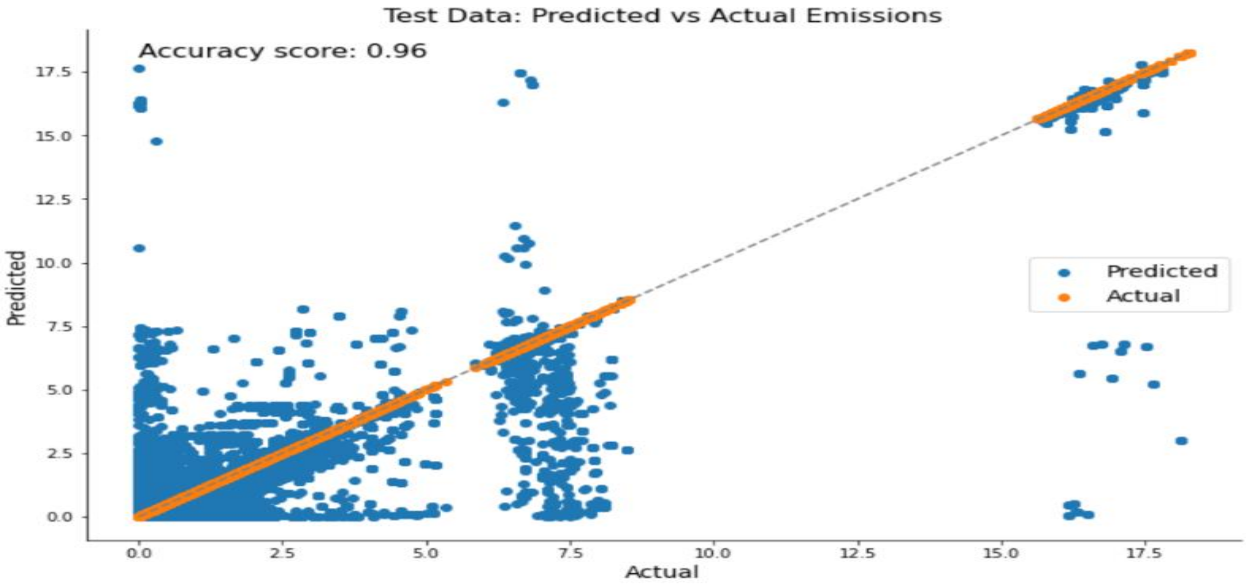

FIGURE 6

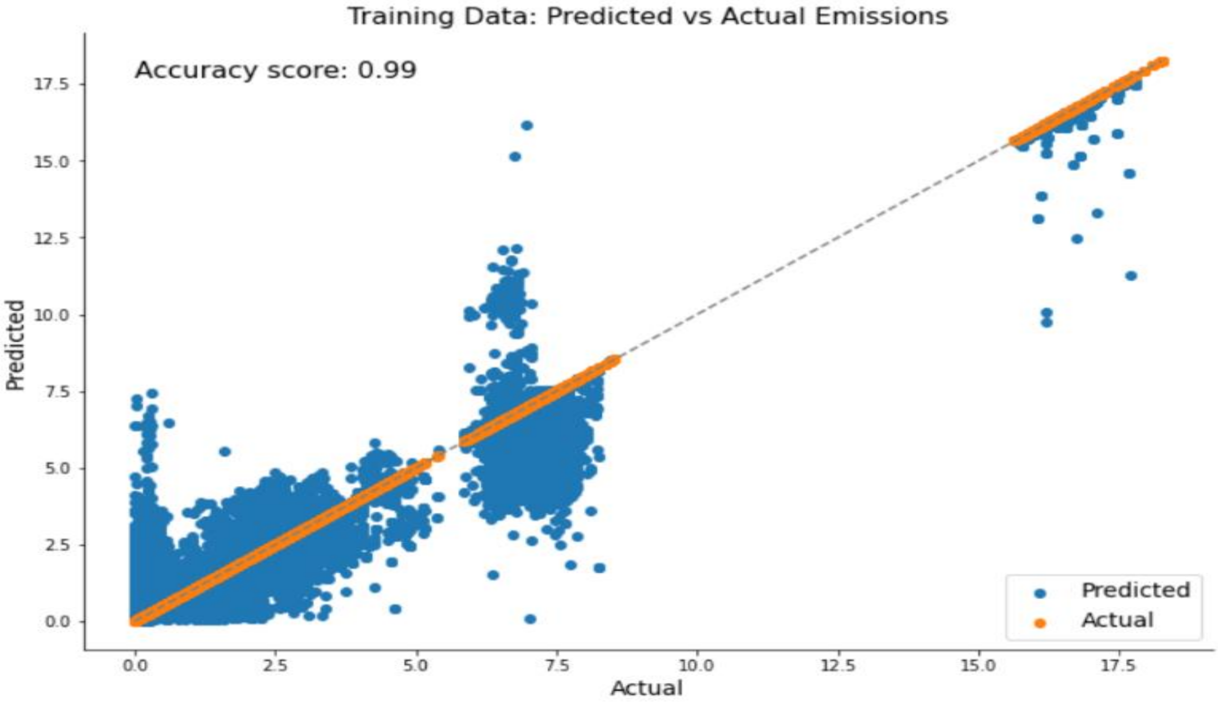

FIGURE 7

5.2 FACEBOOK PROPHET

This research employs Facebook's Prophet library to perform time series forecasting. The dataset used in this study is the combined data from USA, from which we selected the columns "reference_year" and "ef" and dropped the remaining columns. Then, we used a Prophet model to forecast future values with a yearly frequency.

We used FB Prophet model to create a plot (See Figure 8) that shows the yearly trend of emission values. The x-axis of the plot is years, while the y-axis shows the emission factor values. The plot shows that the emission values will range between 0 to 3 grams/mile for the years 2023 to 2027.

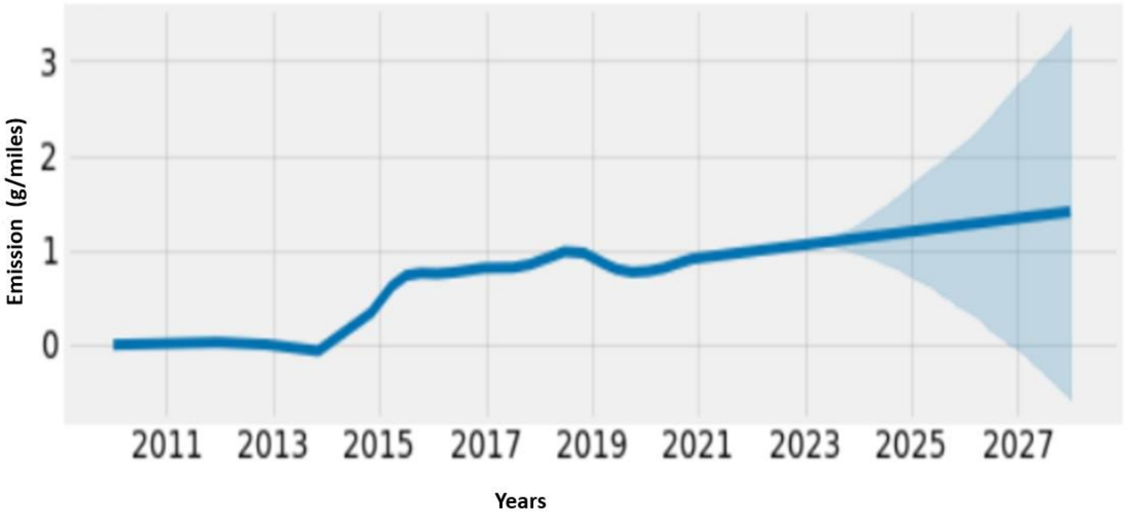

FIGURE 6

5.3 TIME SERIES/FORECASTING USING ARIMA

In this research, we have used the ARIMA model for time series forecasting. We have combined USA MOVES & USA EMFAC data into a one csv file. Initially, we have selected numerical columns from the dataset, and applied one-hot encoding to them. Developed a feature importance model to figure out the top ten essential features. To see trends, applied Simple Moving Average for 10 and 20 years. We use ARIMA model by considering the reference year and the mean of the emission factor values to predict future 5-year emission values.

The trend of actual and forecasted values for the next 5 years. The x-axis of the plot stands for the year, while the y-axis shows the average emissions in grams/miles. The emission factor values for the years between 2023 and 2027 forecast which ranges from 1.0 to 1.15 grams/miles. (See Figure 9)

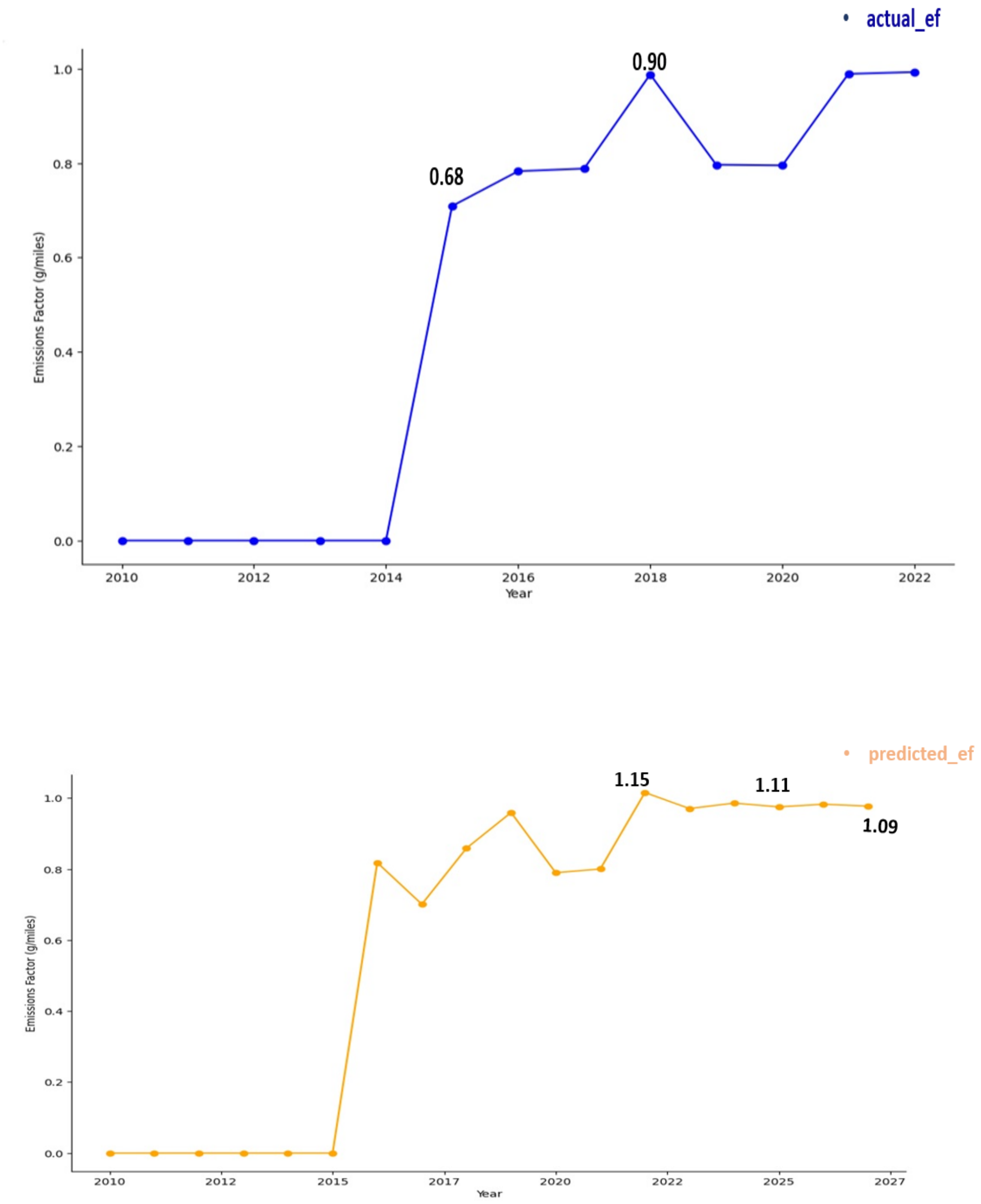

FIGURE 7

6. RESULTS

1. The Energy Consumption (EC) pollutant accounted for most emissions, with CO2 being the second highest.
2. Emissions remained constant between 2010 and 2014 before sharply increasing in 2015, reaching a peak in 2018, and gradually declining until 2020.
3. Emissions were highest for buses built between 1980 and 1990 and gradually decreased thereafter.
4. Emissions were highest for lower speed interval (0-8) miles per hour and lowest for upper speed interval (65-72) miles per hour.
5. For all fuel types of transit buses, EC and CO2 had the highest emission levels of more than 16 and 6.80 grams/mile, respectively.
6. The Random Forest model had an accuracy of 0.96 and 0.99 for test and train data, respectively.
7. The Facebook Prophet model predicts emissions between 0 to 3 gm/mile for the years 2023 to 2027.
8. The Time Series/Forecasting using ARIMA model predicts emission values ranging from 1.0 to 1.15 grams/mile for the years between 2023 and 2027.

7. CONCLUSION

In conclusion, this study sheds light on the significant impact of public transit emissions on air quality and public health. The results of the analysis suggest that transit bus emissions, particularly those of Energy Consumption (EC) and CO2, have remained high over the past decade, with a sharp increase in 2015 and a gradual decline since 2020. The study highlights the importance of reducing transit bus emissions to improve air quality and public health and offers insights that can guide policymakers and transit operators in developing effective strategies and policies. Machine learning models used in the study, such as the Random Forest Regressor, Facebook Prophet Model, and Time series using ARIMA, supply detailed and correct insights into the emission levels of transit buses, fuel types, and speed intervals. These models offer a powerful tool for transportation data analysis and decision-making. The study recommends strategies for reducing transit bus emissions, such as transitioning to more sustainable fuel types and technologies, improving maintenance practices, and enhancing transit efficiency. By implementing these measures, policymakers and transit operators can significantly reduce transit bus emissions and improve air quality. Overall, this study shows the value of using machine learning models to analyze transportation data and supplies a valuable resource for policymakers and transit operators looking to reduce transit bus emissions and improve air quality.

ACKNOWLEDGEMENT

We would like to take this opportunity to express our heartfelt gratitude to Professor Umair Durrani, for his invaluable guidance and support throughout our work on emissions predictions. His extensive knowledge and ability in the field of environmental modeling helped us to understand the complexities of emissions forecasting and develop a deeper appreciation for the importance of sustainability in today's world.

Professor unwavering commitment to teaching and mentorship was clear in the time and effort they dedicated to reviewing our work, supplying constructive feedback, and encouraging us to think critically about our assumptions and methods. His guidance and support were instrumental in helping us to develop skills in statistical modeling and interpreting complex data sets.

His contributions to our academic and personal growth have been immeasurable, and we are grateful for the opportunity to gain experience from such an outstanding professor and mentor. We will always treasure the knowledge and insights you shared with us and will carry them with us throughout our academic and professional journey.

Thank you, Professor, for all you have done for us and for your unwavering dedication to teaching and mentorship.

BIBLIOGRAPHY

(Technical, E. (.-D. (n.d.). Retrieved from https://www.epa.gov/moves/moves-onroad-technical-reports#moves3

Alam, A., Diab, E., El-Geneidy, A. M., & Hatzopoulou, M. (2014). A simulation of transit bus. (n.d.). *Alam et al., 2014.*

Beckx, C., Lefebvre, W., Degraeuwe, B., Vanhulsel, M., Kochan, B., Bellemans, T., . . . Int Panis,. (n.d.).

Beckx, C., Panis, L. I., Janssens, D., & Wets, G. (2010). Applying activity-travel data for the. (n.d.).

Brauer, M. L. (n.d.). *Brauer et al., 2008.*

Chan, E. C. (n.d.). Retrieved from https://gmd.copernicus.org/preprints/gmd-2022-147/gmd-2022-147.pdf

Dons, E. B.-b. (n.d.).

EMFAC. (2014). Retrieved from https://arb.ca.gov/emfac/2014/

Gan, W. Q.-t. (n.d.). *Gan et al., 2012.*

João Pedro Bazzo, Rafael H. M. Pereira,Pedro R. Andrade. (2022, May). *gtfs2emis: Estimating Public Transport Emissions from GTFS Data*. Retrieved from https://github.com/ipeaGIT/gtfs2emis

Liu, H. C.-R. (n.d.). Retrieved from https://journals.sagepub.com/doi/10.3141/2340-05

Rafael H.M. Pereira¹, João Pedro Bazzo Vieira¹,Pedro R. Andrade². (2023, March 08). Retrieved from https://osf.io/8m2cy/

Santos, G., Behrendt, H., Maconi, L., Shirvani, T., & Teytelboym, A. (2010a). Part I: Externalities. (n.d.). *(Santos et al., 2010a).* (Santos et al., 2010a).

Selander, J. N. (n.d.).

Selander, J., Nilsson, M. E., Bluhm, G., Rosenlund, M., Lindqvist, M., Nise, G., & Pershagen, G. (n.d.).

Shan, X. C. (n.d.). Retrieved from https://www.mdpi.com/2071-1050/11/10/2936

Sider, T. A. (n.d.).

Sider, T. M., Alam, A., Farrell, W., Hatzopoulou, M., & Eluru, N. (2014). Evaluating vehicular . (n.d.).