Hybrid Modelling for Public Transportation Wait Assessment Forecasting: A Comparison with Linear and Nonlinear Methods

Paul Ryan

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

February 2024

Supervisor: Marina Iantorno

**Abstract**

*This study looks at public bus wait assessment data as provided by the Metropolitan Transportation Authority who run the buses for New York City. Public transit delays can cause a range of problems for the users and operators as well as leading to an erosion of trust and falling ridership levels. With the knowledge that increasing public transit usage is beneficial in many areas it is proposed that a number of forecasting models are used to attempt to predict the wait assessment values and use this information to help alleviate future delays. A literature review showed that a Seasonal Autoregressive Integrated Moving Average Exogenous (SARIMAX) model was suitable to be used with the exogenous variables, which in this case were traffic and weather features obtained and transformed to align with the wait assessment values. The models examined were based on either SARIMAX or Random Forest (RF) forecasters, with a novel approach to creating a SARIMAX-RF hybrid using the time series residuals as inputs also employed. All models also have their hyperparameters tuned to optimise performance. A sliding window method was tested for its performance in conjunction with a Random Forest model. The tuned SARIMAX model using the weather and traffic exogenous variables was the best performing with a Mean Absolute Error of 0.34 and a Root Mean Square Error of 0.41. The study was a thorough examination of the proposed methods and models and provides a strong foundation for future research.*

**Keywords:** Wait assessment, Seasonal Autoregressive Integrated Moving Average Exogenous (SARIMAX), Random Forest, residuals, exogenous variables, sliding window method

[**1 Introduction 4**](#_5jtyaqyoccq9)

[**2 Research Design 5**](#_gu569kve524e)

[2.1 Primary Data 5](#_lkwxmkt38x5p)

[2.2 Secondary Data 5](#_yt2l0g9u06f2)

[2.3 Problem Identification and Clarification 6](#_vbz0nh5dthua)

[2.4 Research Objectives 7](#_svumh3ot8y8l)

[2.6 Ethics 8](#_9tm8qa7oe1lc)

[**3 Literature Review 9**](#_e0j6tgww12fr)

[3.1 Introduction 9](#_jze3jbrkrhdc)

[3.2 Bus Delays Causes and Impacts 10](#_bosshg5rmbhe)

[3.3 Exogenous Time Series Variables 11](#_rw1f6ygkopnr)

[3.4 Sliding Window Forecasting 14](#_jvpmxtzi3bvk)

[3.5 Hybrid Residual Models 16](#_qxfnqa1cd6n8)

[3.6 Conclusion 18](#_ueimgwahea8f)

[**4 Methodology 20**](#_b6rnv4vryztn)

[4.1 Random Forest 20](#_b5hh2thgwjzn)

[4.2 Time Series Analysis 22](#_m5tejmowhb55)

[4.3 Hybrid Model using Residuals 24](#_39bwq3mvgzcq)

[**5 Implementation and Results 25**](#_vsp4hyppv7kz)

[5.1 Data 25](#_z0y7ojgxc8vi)

[5.2 Exploratory Data Analysis (EDA) 25](#_lidgelmejzqd)

[5.3 Statistics 32](#_gr1ogiinbq2l)

[5.4 Time Series Analysis 40](#_cnzsolcuk0o3)

[5.5 Random Forest 43](#_m7diwni8bwrh)

[5.6 Hybrid Model 49](#_b3qk9vx07nq6)

[5.7 Interactive Dashboard 52](#_iaesmmtvz32z)

[5.8 Table of Results 54](#_4zqwbrvbm65r)

[**Discussion 57**](#_aciw2gxgqs2z)

[**Conclusion 59**](#_fd9ifv4tzehh)

[**A Model Plots 61**](#_kqbjw0ung8zb)

[**B Transcripts 75**](#_n52o14sl16u9)

Table 1: List of Acronyms

| **Acronym** | **Full Form** |
| --- | --- |
| ANN | Artificial Neural Network |
| ADF | Augmented Dickey-Fuller |
| ATR | Automated Traffic Recorders |
| ARMA | Autoregressive and Moving Average |
| ARIMA | Autoregressive Integrated Moving Average |
| BP | Back-Propagation |
| CELA | Cluster Based Ensemble Learning Approach |
| CNN | Convolutional Neural Network |
| c.d.f | Cumulative Probability Distribution Function |
| DELA | Decomposition Based Ensemble Learning Approaches |
| DBN-DNN | Deep Belief Network - Deep Neural Network |
| EDA | Exploratory Data Analysis |
| GELA | General Ensemble Learning Approaches |
| HONN | Higher Order Neural Networks |
| HST | Hydrostatic-Seasonal-Time |
| IQR | Interquartile Range |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MAE | Mean Absolute Error |
| MTA | Metropolitan Transportation Authority |
| MLP | Multilayer Perceptron |
| NOAA | National Oceanic and Atmospheric Administration |
| NYC | New York City |
| NYC DOT | New York City Department of Transportation |
| NRMSE | Normalised Root Mean Square Error |
| PCA | Principal Component Analysis |
| RBF | Radial Base Function |
| RF | Random Forest |
| RELA | Residual Ensemble Learning Approaches |
| RMSE | Root Mean Squared Error |
| SARIMAX | Seasonal Autoregressive Integrated Moving Average Exogenous |
| SVM | Support Vector Machines |
| TDNN | Time Delay Neural Network |

# **1 Introduction**

Wait Assessment (WA) is a tool used to understand the percentage of bus arrivals that occur within three minutes of the scheduled interval period for peak times or within five minutes of the scheduled interval time during non-peak hours. It is currently used by the Metropolitan Transportation Authority (MTA) to report on the performance of its buses, which run throughout the state of New York, with service within New York City (NYC) being handled by two MTA agencies, NYC Transit and MTA Bus Company. In 2022 there were over 425 million rides taken using these two agencies (Metropolitan Transportation Authority, 2022). When a service impacts such a large number of people, and in a way which is meaningful to how many of them live their lives, any disruption can have large ramifications for the individuals affected, as well as secondary negative effects for society in general. Delays and unreliable services are one such disruption that all public transportation systems have to tackle. Encouraging widespread public transportation usage can be beneficial on many different levels, from giving an individual the freedom of mobility, to take job or education opportunities, or reducing the number of private vehicles on the road, decreasing carbon emissions and benefiting the environment and society as a whole. Rissel et al. (2012) even found that using public transport gave a substantial boost to the likelihood that an individual is “sufficiently active”, which confers a number of health benefits. By accurately predicting wait assessment, or another delay measure, advanced warning is given to the transportation operators, allowing them time to respond and potentially alleviate the delay. The information can also be passed to the rider, who with more advanced notice can modify their plans, or notify others of their potential delay. Forecasting the delays is one step, it does not address the root cause, but it can give stakeholders time, and be used as a tool to help manage the issue.

This study aims to tackle the issue of delay prediction using a range of techniques and models, with a view to understand the best combination for returning a solid prediction. Included in the study will be factors which would influence the delays, in this case traffic and weather related data. These were chosen based on a number of factors including the primary research results, the literature review and data availability. There will be three separate model types tested, a SARIMAX, which is a linear model, a Random Forest Classifier (RF) which is nonlinear, and a hybrid mode which takes SARIMAX residuals as additional inputs for the RF. In addition a technique known as a sliding window will be tested. This has shown in studies like Davtyan et al. (2020) to improve the performance of forecasting models.

Section 2 will discuss the research design, including a deeper explanation of the problem being solved for and a set of research objectives to guide the study. Section 3 is a literature review which looks at the current state of the art in regards to a number of facets of the study. The sliding window technique will be investigated here as well as the choice to include exogenous variables into a time series analysis. Section 4 goes through the methodology of the models being used in the study, touching on the high level concepts involved in making them effective. Section 5 then goes through the implementation of the work and the results obtained. Section 6 will take the results in the context of all the previous work done in the study and assess them giving possible implications and recommendations. Section 7 is the conclusion to the study.

# **2 Research Design**

## 2.1 Primary Data

The primary data for this research was in depth interviews with public bus users who deal with accessing services on a regular basis. The interviews were used to support the insights gained from the secondary data as well as understand and clarify the problem statement and research objectives. The individuals were selected using a non-probability sampling method known as self-selection, or judgement sampling. This allows for the use of bias to control the sample and select cases by a certain criteria, in this case the impact that the problem statement has upon them. The information gathered is used throughout the sections but it was of note that all participants mentioned delays, and weather was a prominent factor in why they felt the issues were occurring.

## 2.2 Secondary Data

The secondary data used in the study was collected from a number of different publicly available datasets. These covered areas of bus delays, provided by the Metropolitan Transportation Authority (MTA), weather, provided by the National Oceanic and Atmospheric Administration (NOAA) and traffic, which was supplied by the New York City Department of Transportation (NYC DOT). The bus delay data was on a monthly frequency, as was the weather data. The traffic data was more granular but any periods of less than a month were aggregated so as to fit with the target variable frequency. Variables included for exploration in the study were:

**Wait Assessment** **:** The wait assessment measures how evenly buses are spaced. It is defined as the percentage of actual intervals between buses that are no more than three minutes over the scheduled interval for peak periods, and no more than five minutes over the scheduled interval for off-peak periods.

**Vehicle Volume Counts :** NYC DOT uses Automated Traffic Recorders (ATR) to collect traffic sample volume counts at bridge crossings and roadways.

**Average Precipitation :** The average rainfall for that month

**Average Temperature :** The average temperature for that month, measured in celsius.

## 2.3 Problem Identification and Clarification

One way of measuring the performance of public buses, as it relates to delays in scheduling experienced by the users, is the wait assessment calculation. It is a measure of how evenly buses are spaced and can be taken as a performance indicator. Delays to scheduled times can cause serious disruption for users who depend on the services for transport to work, educaction, medical services, and other basic amenities. Bus delays and unreliability are one of the main factors discouraging use, which, were it to be improved, could have a significant positive impact on the individuals and society at many different levels. Being able to predict disruption would allow for more flexibility in the response of both the transport supplier and user, improving the customer experience and increasing the likelihood of future usage.

Past performance alone was not deemed sufficient for the highest level of prediction and so factors which are likely to impact the level of delays, specifically traffic and weather measurements are included in the study. Bus users identified these in the in-depth interviews as the two factors which they felt impacted the likelihood of a delay.

This study, therefore, will attempt to use past performance data, as well as traffic and weather factors, to forecast future wait assessment values. It will look at using liner, nonlinear and hybrid models and assessing which is the most valuable in terms of accuracy as compared to the actual values.

## 2.4 Research Objectives

The objectives of this study are to:

* Investigate the level of delays, as represented by a wait assessment, for a public bus provider, as well as corresponding traffic and weather conditions for the area, focusing on the impact they have on the dependent delay data.
* Develop and train multiple forecasting models in order to predict the wait assessment using the variables identified in the first step. Models should include a linear model such as an ARIMA, nonlinear models such as a RF, and a hybrid model to capture linear and nonlinear patterns.
* Compare the results of the various models developed in terms of their accuracy as shown by standard metrics such as MAE and RMSE when predicting the wait assessment values.
* Including the results obtained from the models, evaluate the pros and cons of the various models and create a recommendation for implementing the research in a beneficial manner.

2.5 Validity Type

Different validity types need to be considered for the study. Construct validity refers to the extent a particular construct captures the thing being studied, which in this case is the forecasting accuracy of bus wait assessment time. As these are directly observable there is no construct per se, and the forecast values can be compared with the actual values using an appropriate metric, which in the case of this study was the MAE and the RSME. The MAE is scale invariant, meaning that it does not depend on the scale or units of data.

Content validity assesses the extent to which a measure is representative of all relevant aspects. For this study, the features which may affect bus delays were taken from a number of traffic and weather datasets, with exploratory data analysis and principal component analysis being performed to verify content validity.

Face validity is a subjective judgement and does not necessarily rely on statistical analysis but rather intuitive impression, that the measure tool appears to be relevant and appropriate for the intended purpose. In this study, all of the implementations have been researched in the literature review and the methodology broadly follows the best practices established therein

Finally Criterion validity is an evaluation of how much the scores or values obtained correlate with the expected outcome, in this case, are we making a strong prediction of the bus wait assessment values compared to the actual values.

## 2.6 Ethics

The ethical concerns of the paper centre around the primary research section, where in-depth interviews were performed with individuals to gain knowledge and perspective around the issue being studied. To ensure that no ethical issues would arise while carrying out this research, a strict set of procedures were followed. All interviews were required to give informed consent. This was done after carefully explaining the purpose of the interview and how it related to the study, as well as how the interview and the interviewees information would be used. All participants would remain anonymous, and of the interviews which were recorded, the recordings were destroyed after the transcript was created.

While the participants were not asked to reveal personal information or anything that could be deemed to be of a personal nature, to prevent any issue with revealing such information, participants were given a copy of the transcripts to review.

This also addressed another ethical issue, which is that of research integrity. Despite our best efforts to remain impartial, cognitive bias affects everyone and so it is important to verify that our recollection and interpretation of events is the same as how the event actually occurred. In addition to ensuring that no unwanted information was included, the participants also had an opportunity to ensure that the information contained in the transcript gave an accurate account of how the interview took place and the participants responses to the questions posed.

As the secondary research for this study consists solely of publicly available datasets there were no ethical concerns around this area.

# **3 Literature Review**

## 3.1 Introduction

Public transport delays affect large portions of the population and impact wider concerns around environmental concern and urban planning. Understanding the causes and impacts of the delays helps to inform the necessity of the study and which areas have so far been deemed relevant to the prediction of future delays. Information around the performance of bus transit has been provided and due to the time series nature of the data is suitable for autoregressive models.

In the case of our data we also want to look at the impact of traffic and weather variables on the bus performance, and so the best method of incorporating the extra variables was examined. Time series models which take exogenous variables such as SARIMAX would be suitable and a range of studies comparing different implementations were examined.

For the random forest machine learning model, the consideration was around the fact that for time series problems, if only using the series order and the bus performance variable, new models may need to be constantly created. The sliding window method was examined to understand the benefits and impacts of its use on the random forest classifier. Papers examining it use it in a variety of different domains, as well as with different selections around parameters such as window size. The current state of the art is examined for best practices.

Finally, as the residuals of the time series are to be used to create a hybrid model with the random forest classifier, hybrid models are investigated. The different types are looked at, with a focus on those using residuals from a linear model as inputs for a nonlinear model. Hybrid models combining autoregressive models and random forest models are also examined for the methodology used and lessons learned.

## 3.2 Bus Delays Causes and Impacts

In the project being proposed, factors which can influence the likelihood and severity of delays are being considered, specifically measurable weather and traffic data. Both of these can contribute to congestion and impact the reliability of a scheduled bus service.

Taking the weather as the first factor, “adverse weather conditions have an impact on the level of service an operator provides. They also result in higher levels of congestion due to an increase of personal car usage.” (Hofmann and O’Mahony, 2005). This paper took the approach of looking at the effect of just weather conditions on the reliability of bus operators and showed that travel time was longer on rainy days than non-rainy days, leading to delays.

Common sense would say that weather will have an impact on the number of cars on the road, leading to an increase in traffic. It is also conceivable that in extreme weather conditions, people without access to cars are more likely to avail themselves of public transport, as opposed to walking or cycling to their destination. Studies looking at the effect of weather on traffic confirm this assumption, clearly stating “Weather intrinsically influences the transport sector all the time” (Bi, Ye and Zhu, 2022). There are of course several factors which can influence the severity of traffic delays and one study by Bi, Ye and Zhu (2022) found that the day of the week also played a factor, with the impact of meteorological variables being more severe on non-weekdays. Likewise, the correlation between the variables and traffic conditions is relatively weak on weekdays, especially during commute hours, with the authors suggesting the flexibility for travel mode decision at these times as a possible reason for the weaker correlation.

The type of weather conditions will also factor into the level of delays; rain, wind, snow, ice, fog, extreme heat, all contribute, with some posing particular issues, such as reduced visibility or road adhesion. An article by Romanowska and Budzyński (2022) describes a study performed on Spanish motorways, analysing the effect of temperature, rain and snow, wind speed and thickness of snow on free-flow speeds, it was found that depending on the intensity of rainfall, speeds could be reduced by 5.5 - 7 km h-1, and snow causing a larger drop of 9 – 13.7 km h-1. Wind speed had a minimal effect (>8 m s-1) on vehicle speed, while visibility caused a drop in speed only when restricted to several hundred metres.

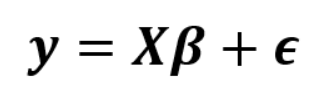
Traffic is the other factor which will be included in its impact on bus delays. The number of vehicles on the road will have a significant impact on the ability to achieve free flow speeds. Research by Yaiza Montero-Lamas et al. (2023) states that traffic delay is a critical factor affecting bus travel time performance. When the traffic in a city grows, more transit vehicles are required to make headway through the congestion, leading to an increase in operation costs. This also results in increased travel time for transit users, which results in ridership loss.

One cause of traffic can be road accidents and construction. These are very different in that the first is a random occurrence which we assume there is no forewarning of, whereas construction can be planned for along a specific timeline. The effect of construction on traffic flow has been looked at with a “moderate effect” being found in one specific study on the impact of an office building in Shanghai (Wang et al., 2019). Information of the impact of traffic accidents on traffic flow is harder to find, with a lot of the current information focused on the cause and outcomes of the accidents themselves. Technology is currently used to attempt to regulate and manage the flow of public transport. Transportation agencies rely on software for traffic planning, “it’s essential to plan services efficiently and be able to estimate expected demand” (Baeli, 2021).

## 3.3 Exogenous Time Series Variables

Exogenous and endogenous variables can be found in the context of regressive models, where an “exogenous variable is an explanatory variable that is not correlated with the error term” (Date, 2022).

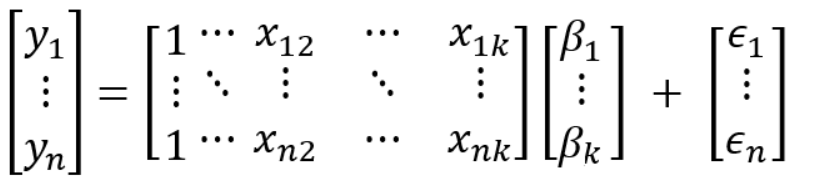
To explain this, if we take the equation of a linear regression model

(1)

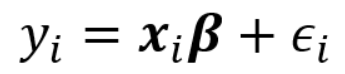
where

* *y* is the dependant variable
* X is the matrix of explanatory variables, including the placeholder for the intercept
* β is the vector of regression coefficients
* Є is the vector of error terms

In the previous equation, some of the variance in *y* is explained by X, but the remaining amount which is not explained needs to be accounted for. This is where the error Є is used. It represents the effect of all factors on *y* that X have not been able to account for. Now if there are *n* data points with *k* regression variables in the data sample, X is a matrix of size [*n* x *k*], *y*  and Є are column vectors of size [*n* x1] and β is a column vector of size [*k* x 1] as follows

(2)

For the *ith* row in the sample (where *x\_i\_k* is the value of the *kth* regression variable ***x****\_k*) the models equation can be expressed as follows

(3)

For the above model, the regression variable ***x****\_k* can be considered exogenous if it is not correlated with Є. It follows that this applies for any given row in the dataset. As the regression variable is not correlated it means that the mean value of the error term is not influenced by it. Therefore an exogenous explanatory variable has no information regarding the model errors, and so cannot be used to predict them.

There are a large number of time series models which will ignore exogenous variables, even if they are present, given the strong focus on the time series (Brownlee, 2018). However, including exogenous variables can help obtain better results in the representation of the transfer function. Athanasopoulos et al. (2011) take tourism data with monthly, quarterly and annual series’ and implement univariate and multivariate time series approaches. They used three fully automated time series algorithms (Forecast Pro, ARIMA and exponential smoothing based algorithms), two method specific approaches (the Theta method and the damped trend), as well as five general frameworks which incorporate exogenous variables. The results of their study showed that methods using exogenous variables were less accurate than those which took only the endogenous variable. This result was contradicted by Allen and Fildes (2001) as well as (Song and Witt, 2003). Some of the explanation for the result was possible model misspecifications, but they still conclude that models with exogenous variables should be evaluated against pure time series alternatives before they are put to use.

If the decision is made to include the exogenous variables as part of the time series analysis a study by Reddy Beeram and Kuchibhotla (2021) looked at the current state of the art when it came to some of the possible prediction and forecasting machine learning models. They looked at what the dominant models have traditionally been, specifically the Box-Jenkins method for time series forecasting using an autoregressive integrated moving average (ARIMA) model, using this as a benchmark. The Box-Jenkins method typically involves three stages. First a model selection stage where the time series data is examined. It is looked at for stationary variables, any seasonality and evaluated to determine which components should be selected. Then the parameters are tuned to find the best fitting ARIMA model. Diagnostic checks are then performed, which can involve checking residuals for auto or partial autocorrelation.

Some studies looked at neural networks, with a stock forecasting paper by Zhang, Aggarwal and Qi (2017) using a novel State Frequency Memory (SFM) recurrent neural network, tested on real-time trading data, and giving an accuracy of predicted trading patterns that outperformed autoregressive models. Another study, also looking at stock data, was performed by Nayak (2017), where four Higher Order Neural Networks (HONN) were combined in hybrid models to perform a one-step-ahead prediction of exchange rates. In the study when compared to a single neural network and a regressive model the hybrid models performed favourably. The Functional Link artificial neural network (ANN) was used as the base model in this study, with Pi-Sigma, Sigma-Pi and Jordan Pi-Sigma as the other three. Rout and Dash (2016) also used the Functional Link method, but this time integrated with the radial base function (RBF) neural network, where it outperformed the individual models, as well as an ARIMA.

There were other studies focused on the best time series model, one looking to forecast the long term performance of the electricity sector (Alharbi and Csala, 2022). The SARIMAX model was chosen, with the ability to include exogenous variables and account for seasonal influence designed to reduce the error values. The results found the model both capable in performance and flexible in its implementation. It’s performance was significantly above other autoregressive models, with the authors praising the ability to deal with different sized sequential datasets. Vagropoulos et al. (2016) compared the SARIMAX to both other autoregressive models as well as ANN-based modelling. The ANN models compared favourably to the more basic ARIMA models but the SARIMAX was chosen as the best performing with a Normalised Root Mean Square Error (NRMSE) of 10.93, compared to the best performing ANN of 11.26. In both studies the SARIMAX models are trained and then optimised using hyperparameter tuning. Sah et al. (2022) look at tuning SARIMAX hyperparameters, and use a grid search cross validation. A grid of possible combinations of a selection of hyperparameters were chosen, and models were fitted to each one. Cross validation was then performed to assess how the predictive model performed. Once the optimal hyperparameters were obtained, significant improvement in the performance of the model was observed.

## 3.4 Sliding Window Forecasting

Time series forecasting is concerned with predicting a future step based on the available past data. This data can contain value in different areas, for instance the most recent values can hold a large amount of importance when it comes to predicting the next step. At the same time, by taking in longer periods of data it is possible to capture more trend and seasonality from the data. A sliding winding approach involves taking a window of fixed size and moving it one step at a time across a range of data, performing calculations on that window, and then creating an output. As the window size is fixed, it can be specified to encapsulate the seasonality of the data, while still keeping the most recent observations as the main basis for future predictions.

This is a method that has become useful in forecasting, Kapoor and Bedi (2013) look at implementing it in a weather forecasting research article. They applied the sliding window method to predict weather variables using a week-long window size based on the likelihood of a correct prediction. It found the approach to be highly accurate, except in months where there were high seasonal changes due to unpredictable weather conditions. It was found that adjusting the window size to a month increased the accuracy of that volatile monthly period.

Selecting the correct window size can be another factor to consider when choosing how to implement the sliding window. A certain level of domain knowledge should inform the choice, what is the date related to, and in what current interval format is it? Alberg and Last (2018) looked at short-term load forecasting in smart metres using sliding window based ARIMA models. They used both seasonal and non seasonal models in the study and looked at different window sizes for producing different results. It found that on a 24-, 48-, and 72-hour window scale, for seasonal models the 24- and 48-hour window sizes were optimal, with the non-seasonal models performing better with the 48- and 72-hour windows. Again the size of the window seems to be very dependent on the data being forecasted. Dong et al. (2020) looked at forecasting equity returns using the sliding window method and ARIMA models, and emphasised the importance of ensuring “the data windows are carefully selected”. They faced the dilemma that shorter windows containing the most recent transaction data would likely fit the future price trend the most, but any sudden change which is not consistent with the recent price developments could generate a large deviation in the forecast outcome. A long window however, would include previous trends that might explain a sudden change, but may be less relevant with predicting the future price. After creating a number of ARIMA models with different window sizes they concluded that the better forecast precision was generated by a longer sliding window size, saying that it “avoids the accumulation of stale information but encourages the inclusion of complete patterns of asset return data”.

The sliding window method can be used with different models, in the previous examples it was implemented with both ANN and ARIMA type models. Other studies have looked at its implementation with a random forest (RF) classifier. When forecasting rainfall distribution Chen et al. (2022) chose a random forest model as it had a number of benefits, namely the potential for quick training, with high flexibility which can work well for all types of data, both balanced and unbalanced. The study was performed on two climate conditions and the results of the RF were found to be satisfactory in rainfall forecasting for both climates. The advantages mentioned were handling many inputs without a variable selection as well as acting as a feature selection technique, while they found some disadvantages to be inaccurate predictions for data outside of the training data range.

Li, Shao and Zhao (2018) use a random forest model to forecast air pollution concentration. For the sliding window method they take the approach of setting a window size *T*, containing the samples, step size is set to 1, and the horizon, which is the number of time steps predicted in advance, is *P.* For every window of *T* they used a RF model to predict the dependent variable of the *P-th* sample in front of the sliding window. The window will then move one step forward and the process repeats. Results using the Random Forest were compared with other models including a Deep Belief Network - Deep Neural Network (DBN-DNN) model, and the RF model with a sliding window achieved the best results. There were three different air pollutants predicted for, and the RF model had RSME’s of 60.0, 24.88 and 17.70 compared to the DBM-DNN’s 71.59, 22.08, and 19.36. The accuracy and MAE of the RF also outperformed the DBN-DNN’s.

Another study looked at predicting dam displacement with Su et al. (2021) noting the specific issue of nonlinear characteristics and how they can reduce accuracy of model predictions. Given RF’s robustness to nonlinear data it was chosen with a sliding time window introduced to alleviate the time-lag effect of impact factor phenomenon and improve the time sensitivity. As the prediction accuracy of RF also depends on the reasonable setting of hyperparameters, they used a grid search optimisation to fine tune the selection. In this study the sliding window strategy is used to directly enter new data into the window, removing the need to delete expired data, as well as improving the model learning accuracy through the change trajectory in the captured time length, which can then further improve the data processing efficiency. The size of the sliding window is fixed at seven, with a step size of one. The combination of the sliding window method and the RF model yielded better results than the comparable Hydrostatic-Seasonal-Time (HST) model and an optimised and regular Multilayer Perceptron (MLP) model. It was also comparable or better than the convolutional neural network (CNN) and the Long short-term Memory (LSTM) model, while requiring less time for model training.

## 3.5 Hybrid Residual Models

When dealing with time series data forecasting, one of the issues faced is how to extract all of the patterns present in the data. Xiao, Xiao and Wang (2012) describe how a single linear or nonlinear model is inadequate as most contain both linear and nonlinear patterns. A solution to this problem is to use a hybrid model which combines more than one forecasting model in order to enhance the predictive capabilities. This can be done in a number of ways, with ensemble models taking an aggregate of the individual to predict a final value. Alternatively models can take outputs from one model stage as inputs for another.

When it comes to forecasting hybrid models have been implemented in different ways across a number of domains. Guermoui et al. (2020) produced a review of papers where hybrid models were used and categorised the models into different classes.

For example one category was the General Ensemble Learning Approaches (GELA), based around the idea that each model contributes differently to the forecasting process, and so several models are fused using different methods to boost the performance of the final forecast. In one such instance Mellit et al. (2005) used a hybrid (MLP-MTM) model, combining a MLP with a library of Markov Transition Matrices (MTM). It was trained with minimum input data and returned a RMSE of 8%.

Cluster Based Ensemble Learning Approach (CELA) is where datasets are divided into clusters and then a linear or nonlinear model is assigned to each cluster. By aggregating the estimated signal from each cluster the final forecasting signal is found. These are featured in studies by Boata and Gravila (2012) who using a C means clustering and a RMSE(kWh/m2) of [0.1374-0.2], with Jiménez-Pérez and Llanos Mora-López (2014) using a K-means clustering, fed with the cumulative probability distribution function (c.d.f.) vectors.

Decomposition Based Ensemble Learning Approaches (DELA) try to decompose the non-stationary signal into a set of meaningful signals in order to render the time series data stationary. The components are all predicted separately with high frequency component signals estimated through nonlinear models and low frequency with linear. Aggregation of the individual components results in a single forecast. Cao and Cao (2006) decomposed solar irradiance into three time-frequency domains using wavelet transformation, and applied a recurrent back-propagation (BP) network. The results showed a more accurate forecast than those obtained from using the ANN’s alone, with a MAE(MJ/M2) of 0.72. Deo, Wen and Qi (2016) also performed a wavelet transformation, but applied support vector machines (SVM). Results for daily forecasts showed that the hybrid model outperformed the classical SVM model for optimum input combinations. It returned a MAE(MJ/M2) of [1.81-2.08].

There are also other hybrid model classifications, but the one most applicable to the current research is the Residual Ensemble Learning Approaches (RELA). Like other hybrid categories it is based on the idea that the time series is composed of linear and nonlinear components. For the linear components, a linear model can be applied, with the residual component modelled using non-linear means. The forecast can be obtained from the combination of the two models. Most of the studies here use some sort of ARIMA linear model, with a variety of nonlinear models having been tested. Ji and Chee (2011) applied an Autoregressive and Moving Average (ARMA) model, and used the resulting residuals with a Time Delay Neural Network (TDNN), with a resulting model which performed well, although performance suffered when predicting negative results, with a RMSE(Wh/m2) of [20- -300]. The linear ARMA model was also used with an ANN by Voyant et al. (2013) where they found that in its best configuration the hybrid model allows an improvement of more than 1% than the ANN alone with a nRMSE(%) of 36.59.

Kumar and Thenmozhi (2014) forecasted stock index returns using an ARIMA and RF model. Their study validated the relevance of hybrid models by comparing them with independent models and they were found to outperform all other models. The ARIMA-RF model returned a RMSE of 0.0173 which was better than the SVM (0.0174), NN (0.0178), RF (0.0188) and ARIMA (0.0186). Ashoke Kumar Biswas et al. (2021) looked at short and mid-term wind power prediction, again using an ARIMA-RF hybrid model to attempt to improve the predictive power. The predictor variables are used to train the ARIMA; if the relationship between wind power generation and the atmospheric variables is nonlinear, the model's residual will contain non-linear information, with that information being used in the RF model. The hybrid ARIMA-RF significantly improved the accuracy with five weather variables with a nRMSE of 34.48% in one case, while also decreasing the error rate.

## 3.6 Conclusion

In reviewing the studies around how traffic and weather data is currently collected as well as the impact it can have upon travel time, the choice of variables to include as factors impacting public transportation delays was justified, with the importance of reliable public transportation also playing a large role in people's lives giving rise to the necessity of the study.

Using exogenous variables as part of a time series analysis is not novel but a large number of studies have been focused on endogenous only, univariate data. Studies such as Athanasopoulos et al. (2011) helped to inform whether including exogenous variables was helpful in its contribution to the model, with Vagropoulos et al. (2016) being one of the papers that would support the inclusion, showing how it could add additional layers of complexity and help the linear model outperform a comparison ANN, while also being flexible to handling datasets of differing sizes. Sah et al. (2022) help show how to best perform a grid search cross validation to tune the hyperparameters of the SARIMAX model and further improve its performance. Limitations around this research are conflicting reports on the effectiveness of including the exogenous variables as opposed to using a simple time series.

The sliding window method when implemented for forecasting has a number of different variables to be accounted for, namely the window size, step size, and horizon. The method was found to improve accuracy in studies by Chen et al. (2022) and Kapoor and Bedi (2013), with the latter encountering the issue of window size. In their study, increasing the window size to one month, increased the accuracy of weather predictions particularly around irregular periods. Dong et al. (2020) investigated this and found that longer window sizes did tend to return more accurate results but that it did depend on the data set, which follows given the trend capturing nature of the sliding window approach. Other studies looked at its implementation in combination with a random forest classifier and showed promising results with Su et al. (2021) finding it outperforming the deep learning models it was compared to. The window size selected for that study was seven, which again underlines the importance of understanding the data and trends within it when choosing the sliding window parameters.

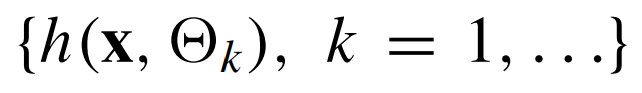
The paper by Guermoui et al. (2020) reviewing hybrid modelling was helpful as a high level view of the current different types of hybrid predictive forecasting models. In the review of the model categories as well as the papers studying individual models, it can be seen that a large number of models attempt to utilise both the linear and nonlinear nature of the time series data, and do so with multiple models in a variety of ways, including clustering such as with Boata and Gravila (2012), or decomposition such as with Cao and Cao (2006). The RELA models showed that residual based hybrids could offer high performance and studies like Kumar and Thenmozhi (2014) acted as a proof of concept that RF models could be used as a nonlinear predictor. Drawbacks from the research in this area is a lack of studies performed using a SARIMAX model as part of the hybrid model, all linear models seen were ARIMA or ARMA class.

# **4 Methodology**

Different models were trained on bus delay data, with date, traffic and weather variables, in an effort to predict future delay times. An overview of how these models work is presented here in an effort to aid the understanding of how the overall analysis was implemented.

## 4.1 Random Forest

According to Breiman (2001) a random forest is defined as a classifier, which itself consists of a collection of tree-structured classifiers.

 (4)

Where:

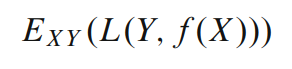
* 𝛉*k* are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**

Classification accuracy can be improved by growing ensembles of trees, and letting them vote for the most popular class. To grow the ensembles, random vectors can be generated which determine the growth of each tree in the ensemble.

Bagging is one example of this, where a random selection, without replacement, was taken from the examples in the training set to grow each tree. The random split selection is another method, which at each node selects the split at random from among the K best splits.

For all of these methods, the same underlying process is followed whereby for each tree, a random vector is generated, which is independent of the past random vectors but with the same distribution. A tree is then grown using both the training set and the random vector, which results in a classifier containing an input vector. Once a large number of trees are generated, the results are aggregated in a ‘vote’ for the most popular class.

To process it from a mathematical standpoint, the goal is that given predictor variables represented by a p-dimensional random vector X = (X1,...,Xp)T, a prediction function f(X) will be found to predict Y, representing the response values. Cha Zhang and Ma (2012) explain that the prediction function is determined by a loss function, L(Y,𝑓(X)) and is defined so as to minimise the expected value of the loss.

 (5)

Where

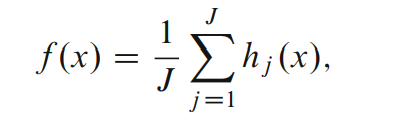
* EXY denotes expectation with respect to the joint distribution of X and Y.

L(Y, 𝑓(X)) will penalise values of 𝑓(X) that are a long way from Y, with the choice of L depending on whether it is a regression or classification problem. The regression function becomes

(6)

Which minimises EXY(L(Y, 𝑓(X))) for squared error loss.

Ensembles collect “base learners”, *h*1(*x*),...,*hJ*(*x*), which are then combined to give the “ensemble predictor” 𝑓(*x*). For a regression problem these base learners are then averaged.

(7)

For a Random Forest the *j*th base learner is *hj(X,* 𝛉j*)*, where 𝛉j is a collection of random variables and the 𝛉j’s are independent for j = 1,..., *J*.

The hyperparameters of the random forest control the structure of each individual tree, the structure and size of the forest and the level of randomness (Probst, Wright and Boulesteix, 2019). The hyperparameters used in the models that are part of the analysis are as follows:

* n\_estimators : The number of trees in the random forest ensemble.
* max\_depth : The maximum depth of the individual trees. A higher depth captures more variance but can cause overfitting.
* min\_samples\_split : The minimum number of samples needed to split an internal node when a tree is being constructed. It prevents the tree from growing if the number of samples falls below the threshold specified.
* min\_samples\_leaf: The minimum number of samples required to be a leaf node.Requires a certain number of samples in each leaf to prevent overfitting.
* max\_features : This helps to control the balance between bias and variance within the model. Like the max\_depth a higher value can capture more information but risks overfitting.

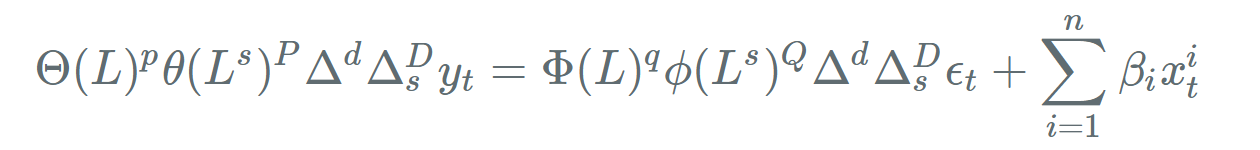
While individual decision trees can be easily interpretable, Schonlau and Zou (2020) state that in return for increased performance, random forests lose the interpretability as many decision trees are aggregated.

## 4.2 Time Series Analysis

Time series data can be considered as data which is indexed in a form where the data points are the “magnitude of changes occurring with time” (Verma, 2021). Because in time series, data points are gathered at adjacent time-spaces, there is a relationship between observations, whether they can be proportional or unproportional, this is the feature that differentiates time series data from other data. Time can be considered the independent variable with the goal being to model changes in a data series, the dependent variable. The patterns in the changes over time can be referred to as components. Time series data is also unique in that time only moves in one direction, and so future observations cannot affect past ones, while events closer together will have a stronger association than more distant observations (Frost, 2020).

When examining the time series data several things should be considered. First the assumption that the summary statistics of observations are consistent. A time series which meets this assumption can be said to be stationary. The “first step in an analysis should be to check whether there is any evidence of a trend or seasons; effects and, if there is, remove them” (Cowpertwait and Metcalfe, 2009). Where a time series is stationary can be determined using a number of techniques, in this analysis statistical testing was used to determine this. Moving averages, or rolling averages, can be used to smooth time series data, and can turn non-stationary time series into stationary ones. According to Frost (2020b), the method relies on the idea that observations close in time are likely to have similar values. Therefore, by averaging you remove random variation, or noise, from the data. As you set the length of the rolling average to the season length that appears in the data, it can remove seasonal patterns to reveal underlying trends.

The time series performed in the analysis is a Seasonal Autoregressive Integrated Moving Average Exogenous model (SARIMAX). The SARIMAX model is defined by

(8)

Where

* *p* is the number of time lags to regress on
* 𝛉(*L*)*p* is an order *p* polynomial function of the lag operator, *L*
* 𝛜t is the noise at time *t* and 𝛃 is a constant.
* *q* is the number of time lags of the error term to regress on.
* Φ is defined analogously to 𝛉
* 𝝙 is an integration operator where *d* is the order of differencing used.
* *P, D,* and *Q* have a similar meaning to *p, d* and *q* but apply to seasonal lags
* 𝑛 are the exogenous variables not auto regressed on

As can be seen, the parameters of a SARIMAX time series which can influence the results of the model are p, d, q, P ,D, Q and s. The p, d, q values are the autoregressive order, the differencing order and the moving average order respectively. These are the same but for seasonal lags for the P, D and Q values. The s order is the length of the seasonal cycle, as the data being used in the analysis is monthly, this will be set to 12 to cover a yearly seasonal cycle.

## 

## 4.3 Hybrid Model using Residuals

The proposed hybrid model is a random forest classifier, incorporating the residuals which were obtained from a SARIMAX time series model. Residuals are what remain after fitting a model. Hyndman and Athanasopoulos (2018) explain that each observation in a time series can be forecast using all previous observations, and these are known as fitted values, represented by *ŷt|t-*1which is the forecast of *yt* based on observations *y1*,...,*yt*-1.

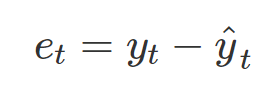
Fitted values always involve one-step forecasts and *ŷt|t-*1 is so regularly used it is written as *ŷt*. They are not forecasts, because if a parameter is involved in the forecasting method, it is estimated using all available observations in the time series, which includes future observations as well. Using the average method the fitted values can be given by

(9)

where

* ĉ is the average calculated over all available observations, which includes those after time *t*.

In this case, the difference between the observations and the corresponding fitted values is equal to the residuals, which is represented as

 (10)

where

* е*t* are the residuals for each observation by timestep

Time series data on its own can show nonlinear behaviour which may not be fully captured. The residuals contain information about the deviation from expected behaviour. As random forests are robust to nonlinear relationships, adding the residuals as features can capture additional complexity in the data. It also provides the model with an indication of where the original model is under or over-fitted, allowing for the opportunity to correct for these errors.

# **5 Implementation and Results**

## 5.1 Data

All of the data sources used have been outlined in the Secondary Data section with links to how to obtain copies. In this case the data on MTA NYC bus delays was accessed via an API call to the New York State Open Data Program website where the data is hosted. This same process was followed by the traffic information datasets. The delay data, which was the target variable of the study, was only available in a monthly format. It was also split into two different time period datasets, and it was necessary to merge these after downloading.Due to the format of the delay data, the traffic data had to be transformed, as it was in various interval formats, to align with a monthly time series. Some of this transformation could be done when calling the api, grouping by certain columns of the csv file being provided, but in most cases, post processing using pandas was performed on the downloaded data.

The weather data was not available via an api and so was loaded in on csv files taken from the NOAA source website. This was already in monthly format but each weather condition was contained in a separate table, so these were all combined with a weather type indicator column being added.

## 5.2 Exploratory Data Analysis (EDA)

*Delay Data*

The delay data contain monthly information denoted by a single ‘date’ column with ‘wait assessment’ being the measurement. These were further grouped by categories such as ‘borough’, ‘route id’ and ‘period’, which denoted if it was for peak or off peak periods. Taking the wait assessment values from the full data set, the mean was 0.79, and it was found to be normally distributed using the Shapiro-Wilk test of normality. This calculates a p value which is the “probability of finding the observed - or a larger - deviation from normality in our *sample* if the distribution is exactly normal in our population” (Geert van den Berg, 2023). As the sample size was larger than 5000 there is a warning that the calculated p-value may not be accurate. At this size due to the central limit theorem, Frost (2018) explains how normality can be assumed as given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution. This can also be informed by a histogram with kernel density overlaid, which in this case does show a relatively bell shaped distribution associated with normality.

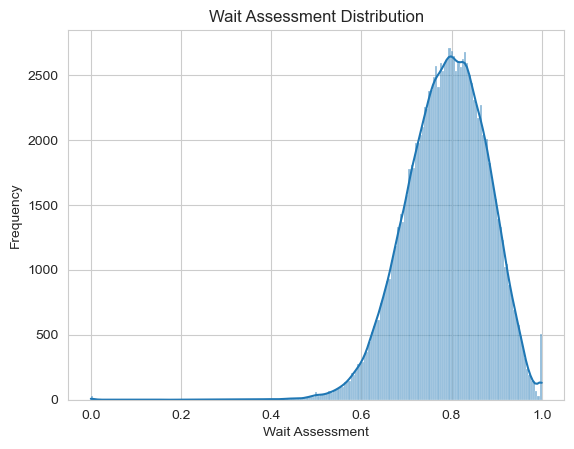


Figure 1: Wait Assessment kdf distribution

Both mean and median were 0.79, with 0.73 and 0.85 being the 25th and 75th percentile respectively.

The mean wait assessment across categories was plotted as a time series. The data covers a time period from January 2015 to December 2023.

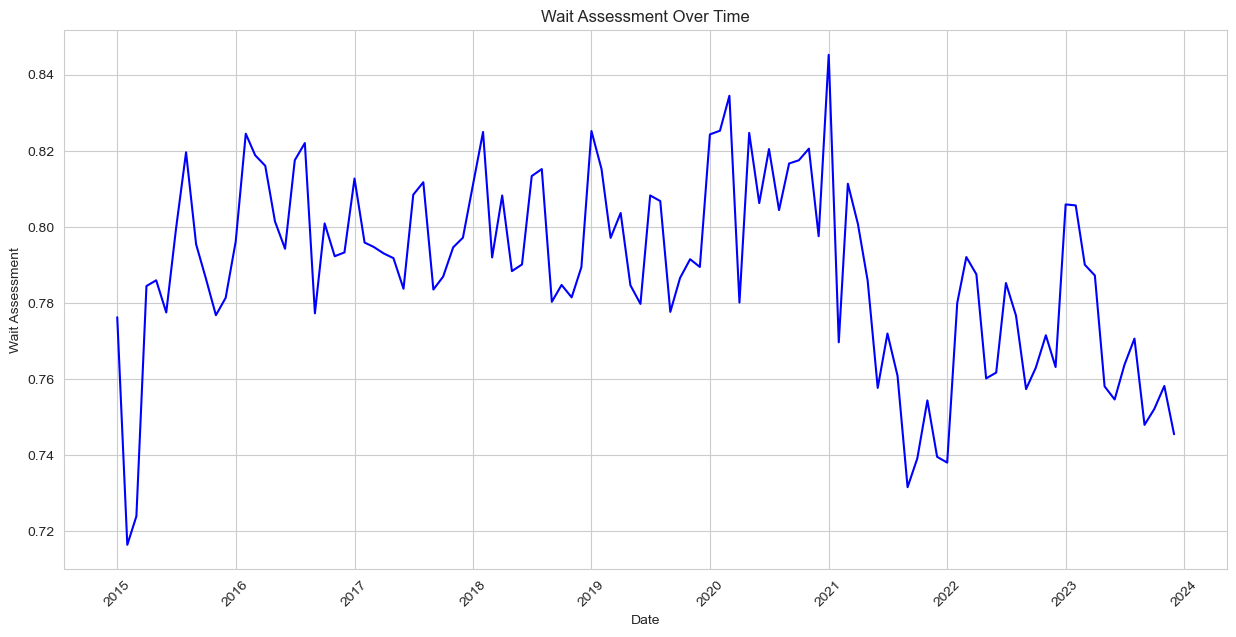


Figure 2: Wait Assessment over time

Overall there was a drop in the wait assessment starting in 2021 and while there have been spikes since the overall level has stayed below this previous period. There also looks to be a degree of seasonality in the data, this will be examined further when the time series model will be created.

*Traffic Data*

The traffic data has different metrics which include a vehicle count, speed, and travel time. These are demarcated under the ‘value\_type’ column with the corresponding value being held in the ‘value’ column. Both the speed and travel time values had date values ranging from 2017 to 2020, while the vehicle count had values ranging from 2000 to 2024. As the idea is to use traffic information as variables to inform the models predicting the delay values, it is important to cover the full range of dates in the delay dataset. For this reason the vehicle count values were selected as the ones to be considered for inclusion in the model. There were two separate groups of vehicle count data remaining however, and so principal component analysis (PCA) was performed to understand which of these should be included in the final models. PCA is a statistical technique that transforms the original dataset into a new coordinate system that is structured by the principal components. These principal components are linear combinations of the original variables that have the maximum variance compared to other linear combinations. These components capture as much information from the original dataset as possible (IBM, 2023). The results showed that the values from the ‘vehicle\_count\_automated\_traffic\_volume\_counts’ accounted for 0.99 of the total explained variance and so it was chosen as the single traffic data variable to be included in the delay prediction models.

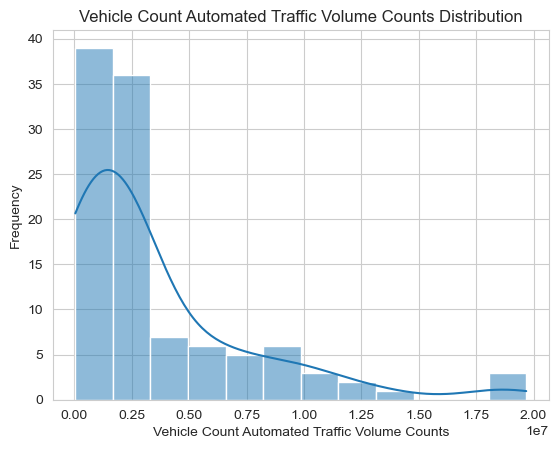


Figure 3: Vehicle Count distribution

When plotting the histogram of the values it can be seen that the distribution of values is right skewed, indicating that the mean will be greater than the median and the data may not be normally distributed. Printing the values shows a mean of 3479085 and a median of 2337810, and a Shapiro-Wilks test advised that the p-value was smaller than the 0.05 alpha and the null hypothesis was rejected.

*Weather Data*

The weather date included a number of different weather types, ten in total, with different measurements around temperature, precipitation and snowfall. The date range on the weather data selected went from January 1st 2000 to January 1st 2024, in monthly intervals. As there were multiple weather events to choose from, PCA was again performed to select those with the most variance. In this case the ‘average precipitation’ and ‘average temperature’ features had the highest explained variance ratio at 0.72 and 0.26 respectively, containing 0.98 combined. As a result these two were selected as the weather variables to be included in the predicted model.

The average precipitation data was right skewed with a mean of 0.14 and a median of 0.13

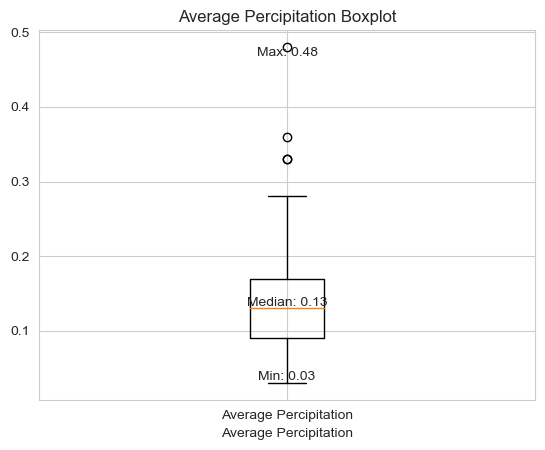


Figure 4: Average precipitation boxplot

With a minimum value of 0.03 and a maximum of 0.48, it can be seen that the maximum value is one of a handful of outliers, of which there are none for the minimum values. It can be concluded that there have been very few occurrences of precipitation this high through the time period, but that due to their extremity, they may be influential on the delay data.

The average temperature distribution was bimodal, with the peak at 80 and a slightly lower peak at 40. As we are measuring temperature over time, having two means and medians could be explained by the difference in temperature over seasons with one being the summer peak and the other being the winter peak.

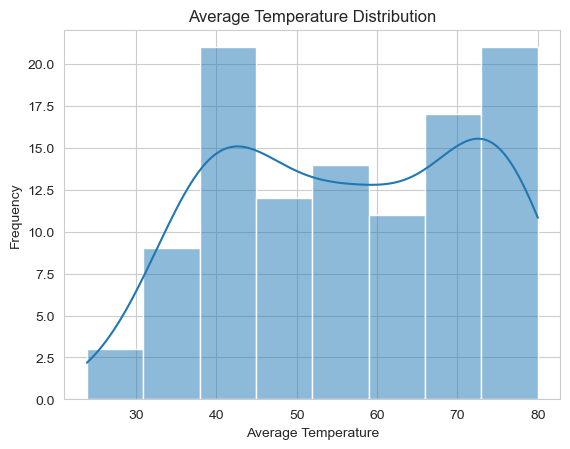


Figure 5: Bimodal Average temperature distribution

When the data is plotted as a time series, the seasonality can be seen with large differences in the values repeated in the same pattern throughout the years.

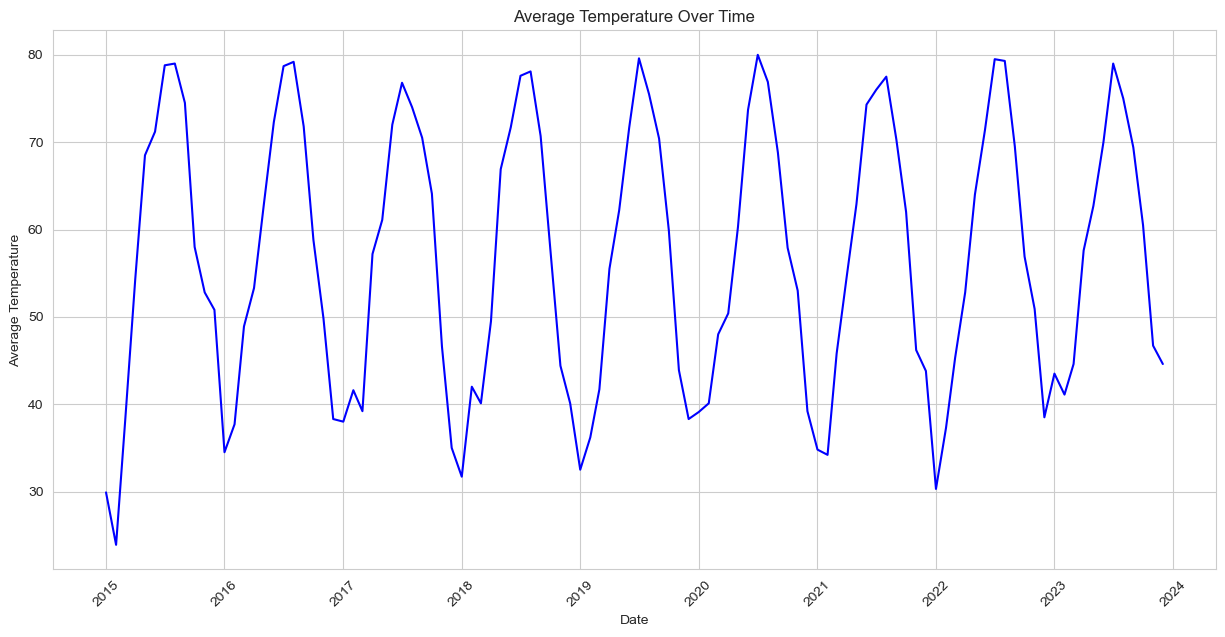


Figure 6: Average temperature over time

## 5.3 Statistics

*Assumptions*

Before performing statistical analysis it is important to test the values to see if they follow certain required assumptions. For the two-way anova planned, “the error model is the usual one of Normal distribution with equal variance for all subjects that share levels of both (all) of the explanatory variables.” (Seltman, 2018). This requires the values for each combination of delay categories to be normally distributed, have the same number of samples, and have homogeneity of variance. Value counts were performed to ensure that the sample sizes of all groups were the same, which they were. The Shapiro-Wilks test was performed on each grouping's values with the results showing a mix of normally and non normally distributed values.

Table 2: Group distributions

| **Group** | **p\_value** | **Distribution** |
| --- | --- | --- |
| ('Bronx', 'Off-Peak') | 1.00 | Normally Distributed |
| ('Bronx', 'Peak') | 0.00 | Not Normally Distributed |
| ('Brooklyn', 'Off-Peak') | 0.00 | Not Normally Distributed |
| ('Brooklyn', 'Peak') | 0.00 | Not Normally Distributed |
| ('Manhattan', 'Off-Peak') | 1.00 | Normally Distributed |
| ('Manhattan', 'Peak') | 0.00 | Not Normally Distributed |
| ('Queens', 'Off-Peak') | 1.00 | Normally Distributed |
| ('Queens', 'Peak') | 1.00 | Normally Distributed |
| ('Staten Island', 'Off-Peak') | 1.00 | Normally Distributed |
| ('Staten Island', 'Peak') | 1.00 | Normally Distributed |

To test if there was homogeneity of variance, Levene's test for equality of variances was used. Levines test is an alternative to the Barlett test and is less sensitive to departures from normality (National Institute of Standards and Technology, 2003). In the case of the delay data where some groups are normally distributed and some are not, this seems the most appropriate. In this case the homogeneity of variances was not satisfied, with a p-value of 0.00, lower than the 0.05 alpha value. With the failure of the homogeneity of variance and some groups having non normal distribution, a two way anova may not be the best fit, however in the case of the delay data there is a large enough sample size where the assumptions may be robust to violation. This could be true if the sample sizes are large enough and not too dissimilar in size (Feldman, 2022). The decision is made to proceed with the two way ANOVA, and then perform the non parametric Kruskall Wallis test on the borough and period groups separately.

*Two - Way Anova*

The groupings were visualised, in all cases the off-peak period values had a higher median than the peak period values. There were a number of minimum outliers across all boroughs and periods.

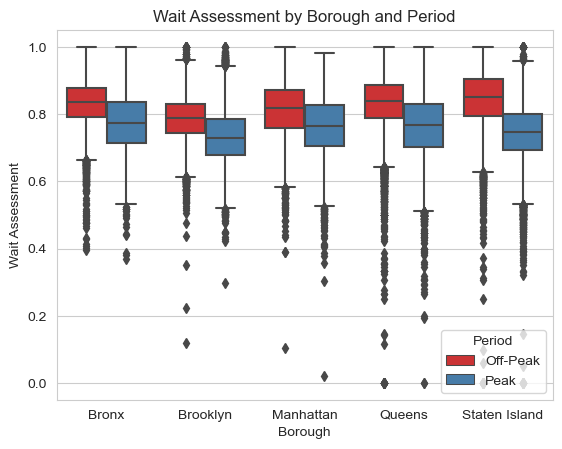


Figure 7: Wait assessment boxplot for borough and period

Off peak means were also higher and standard deviation values lower, than their peak counterparts.

Table 3: Group statistics

| **Borough** | **Period** | **Mean** | **Standard Deviation** |
| --- | --- | --- | --- |
| Bronx | Off-Peak | 0.8319 | 0.0663 |
| Bronx | Peak | 0.7751 | 0.0850 |
| Brooklyn | Off-Peak | 0.7879 | 0.0681 |
| Brooklyn | Peak | 0.7331 | 0.0782 |
| Manhattan | Off-Peak | 0.8113 | 0.0777 |
| Manhattan | Peak | 0.7655 | 0.0879 |
| Queens | Off-Peak | 0.8335 | 0.0803 |
| Queens | Peak | 0.7641 | 0.0919 |
| Staten Island | Off-Peak | 0.8428 | 0.0908 |
| Staten Island | Peak | 0.7411 | 0.0908 |

The Hypothesis Test was as follows:

Null Hypothesis (H₀): The mean of the wait assessment in one borough is equal to the mean of the wait assessment in other boroughs, and there is no difference between peak and off peak values.

Alternative Hypothesis (H₁): At least one of the means of the wait assessment in one borough is significantly different than that of the other boroughs, and there is a difference between peak and off peak adjusted values.

The F-statistic was calculated by comparing the means of the different groups and factors.

The Hypothesis Test results were as follows:

F-statistic for the main effect of borough was 1132.22

F-statistic for the main effect of period was 19453.18

F-statistic for the interaction between borough and period was 313.29

The degrees of freedom for the main effect of borough was 4

The degrees of freedom for the main effect of period was 1

The degrees of freedom for the interaction effect was 4

The p-value for the main effect of seasonality is 0.00, and as it is less than the alpha of 0.05, we reject the null hypothesis that there is no statistically significant difference in means.

The p-value for the main effect of countries is 0.00, and as it is less than the alpha of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in means.

The p-value for the interaction effect between seasonality and country is 1.09E-268, and as it is less than the alpha of 0.05, we reject the null hypothesis that there is no interaction effect between factors.

Table 4: Two-way ANOVA results

|  | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| --- | --- | --- | --- | --- | --- |
| **borough** | 4 | 30.43194096 | 7.607985241 | 1132.21655 | 0 |
| **period** | 1 | 130.7165874 | 130.7165874 | 19453.17701 | 0 |
| **borough:period** | 4 | 8.420739689 | 2.105184922 | 313.2925649 | 1.09E-268 |
| **Residual** | 123990 | 833.1569521 | 0.006719549577 |  |  |

An interaction plot showed how several of the borough variables overlapped each other. This shows the presence of the interaction effect, which indicates that another variable is influencing the relationship between an independent and dependent variable.

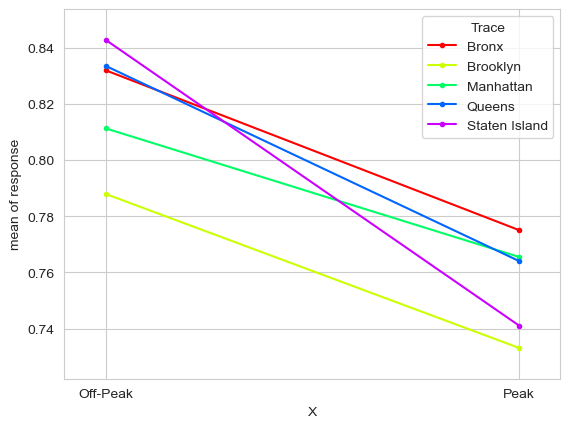


Figure 8: Wait Assessment Interaction Chart for Borough and Period

*Kruskall Wallus*

The Kruskal Wallus test was chosen due to the non-parametric nature of the data. The first group to be compared was the period categories, peak and off peak. As with the borough crossed box plots previously seen, off peak has a higher wait assessment mean when plotted.

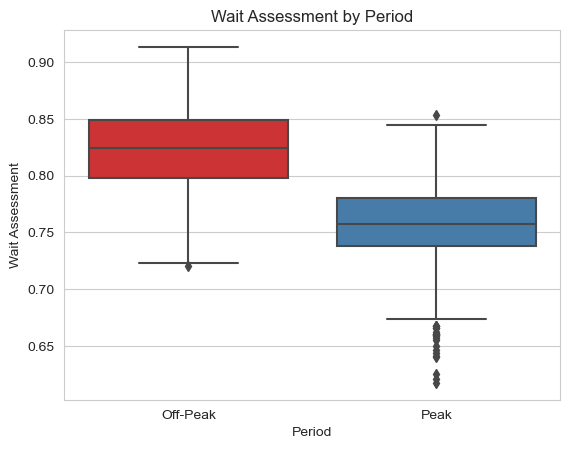


Figure 9: Boxplot of wait assessment values for period

The Hypothesis Test is as follows:

Null Hypothesis (H₀): The distributions of the wait assessment values are the same across periods.

Alternative Hypothesis (H₁): The distributions of the wait assessment values differ significantly across periods.

With an alpha (Significance level)

The test statistic for this test was calculated by looking at the differences between the distributions of the groups.

The Hypothesis Test results were as follows:

The t-statistic came to 495.72.

The p-value associated with the t-statistic is 0.0000, and as it is less than the alpha of 0.05, we reject the null hypothesis that there was no statistically significant difference between the distributions.

The second grouping was by borough, with the values split across the five boroughs that comprise NYC.

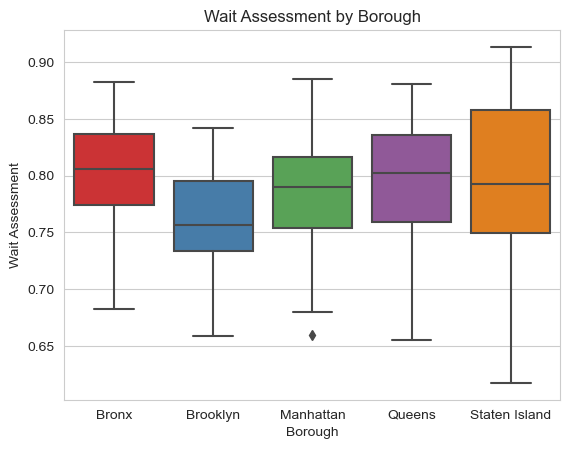


Figure 10: Boxplot of wait assessment values for borough

Staten Island can be seen to have the largest interquartile range (IQR), indicating an even spread over a larger possible wait assessment period. Brooklyn has the lowest median wait assessment and the Bronx has the highest, with 0.76 and 0.81 respectively. Manhattan, Queens and Staten Island all have a close median range, lying between 0.79 and 0.80.

The Hypothesis Test is as follows:

Null Hypothesis (H₀): The distributions of the wait assessment values are the same across boroughs.

Alternative Hypothesis (H₁): The distributions of the wait assessment values differ significantly across boroughs.

With an alpha (Significance level)

The test statistic for this test was calculated by looking at the differences between the distributions of the groups.

Post Hoc tests were also performed to further examine pairwise comparisons between the boroughs; this was accomplished using Dunn's test.

The Hypothesis Test results were as follows:

The t-statistic came to 106.88.

The p-value associated with the t-statistic is 0.0000, and as it is less than the alpha of 0.05, we reject the null hypothesis that there was no statistically significant difference between the distributions.

The Dunn’s post hoc tests showed that there were significant differences in the distribution of Brooklyn which was different to all other boroughs. The Bronx was also significantly different to Manhattan.

Table 5: Dunn’s post hoc test results

|  | **Bronx** | **Brooklyn** | **Manhattan** | **Queens** | **Staten Island** |
| --- | --- | --- | --- | --- | --- |
| **Bronx** | 1.00 | 0.00 | 0.00 | 0.48 | 0.12 |
| **Brooklyn** | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| **Manhattan** | 0.00 | 0.00 | 1.00 | 0.09 | 0.48 |
| **Queens** | 0.48 | 0.00 | 0.09 | 1.00 | 0.48 |
| **Staten Island** | 0.12 | 0.00 | 0.48 | 0.48 | 1.00 |

## 5.4 Time Series Analysis

*Data Processing*

With a combined dataset having a datetime index, wait assessment as the target variable and three exogenous variables in average temperature, average precipitation and vehicle count the first step was to seasonally decompose the wait assessment values. This looks at the values in terms of a combination of level, trend seasonality and noise components (Brownlee, 2017). In this case an additive model was used, which is linear, where changes over time are consistently made by the same amount.

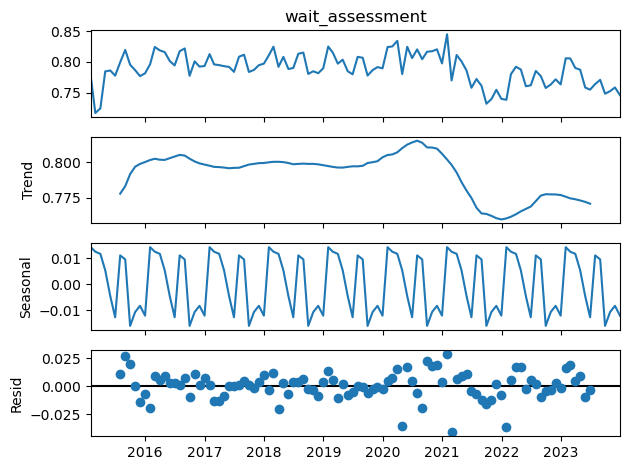


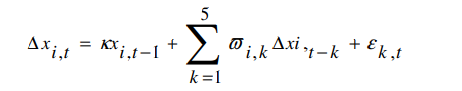
Figure 11: Decomposition of wait assessment

Wait assessment dropped significantly from its peak in mid 2020 to its lowest point at the end of 2021. The seasonality looks both regular and reasonable and the residuals show lower variability from 2016 until 2020 at which point the residuals spread significantly.

An Augmented Dickey-Fuller (ADF) test was performed to check the stationarity of the data. For a time series to be considered stationary the statistical properties of its value, “such as the mean, variance, auto correlation etc, are all constant over time” (Nau, 2014).

An ADF test is a generalised auto-regression model formulated in the following regression

equation

 (11)

The Hypothesis test is

Null Hypothesis (H₀): The data is non Stationary

Alternative Hypothesis (H₁): The data is Stationary

The p-value of the initial ADF test was 0.45, and therefore the null hypothesis was accepted, and the data was found to be non stationary.

A technique using rolling means is employed to try to make the data stationary. “A rolling mean is the mean of the previous x number of observations in the series” (Mitrani, 2020), in the case of the delay data 12 is the number of previous observations as it is of monthly intervals. A column named ‘rolling\_mean\_diff’ was created by subtracting the rolling mean shifted by one month, from the rolling mean value. When the ADF test was performed on this set the p-value was 0.01, and therefore the null hypothesis was rejected, and the data was found to be stationary, where rolling\_mean\_diff was used as the target variable.

*Modelling and Parameter Tuning*

After splitting the data into training and testing sets, the Sarimax model was fitted on the training set. The exogenous variables of average precipitation, average temperature and vehicle counts were also included and the frequency was set to monthly.

When the model was used to predict the test set it resulted in a Mean Absolute Error (MAE) of 0.00 and a Root Mean Squared Error (RMSE) of 0.00.

The data was then standardised using StandardScaler, which rescales the distribution of values, so that the mean of observed values is 0 and the standard deviation is 1 (Brownlee, 2020). After this the S ARIMAX model was run again and the MAE was found to be 0.59 with a RSME of 0.78.

In an attempt to improve the accuracy of the model a grid search was performed where the p, d, q and seasonal order parameters were stored in each possible combination and then loop tested to return the best performing combination. This resulted in a combination of 1,0,0 for the p, d, q values respectively and a seasonal order of 0, 0, 1

When the parameters were applied to the model and it was again run on the training set it resulted in a MAE of 0.34 and a RMSE of 0.41. Plotted it can be seen that the predicted values do follow the peaks and valleys of the testing set, although the alignment seems less accurate as time goes by.

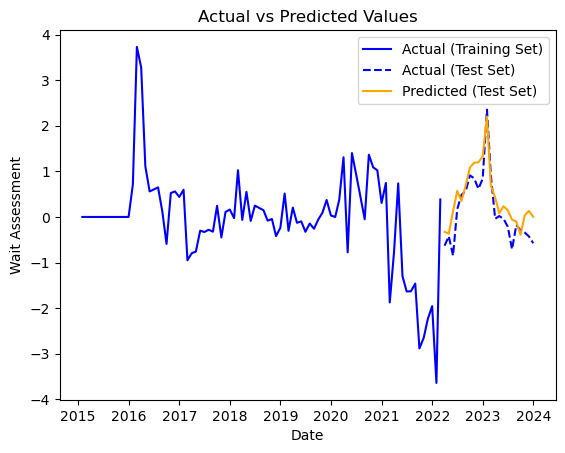


Figure 11: Forecasted SARIMAX values

## 5.5 Random Forest

*Data Processing*

The Random Forest model would have the same date, wait assessment, average precipitation and temperature, and vehicle count variables. In the case of the Random Forest model, it can be helpful to augment the data, transforming the existing variable in some way which can provide more predictive power to the model. The first method of doing so was by creating categorical representations of the level of the traffic and weather variables. Each variable was examined with the 75th and 90th percentile calculated. Values below the 75th, between the 75th and 90th and above the 90th were then associated with a number from 1 to 3 indicating the performance of the group they belong to. This resulted in three new columns, ‘vc\_rating’, ‘ap\_rating’, and ‘at\_rating’ for the vehicle count, average precipitation and average temperature respectively. The date column is then split into a number of columns, again in an attempt to augment the dataset. This removes the date column and replaces it with ‘Year’, ‘Month’, ‘Is\_quarter\_start’, ‘Is\_year\_start’ and ‘Elapsed’.

*Modelling and Parameter Tuning*

If the only data available is the values for the corresponding date series it can be treated as a time series problem, which does not fit well with machine learning techniques as they ignore the temporal components inherent in the problem (Brownlee, 2016). In this case the data can be transformed from a time series dataset into a supervised learning dataset by shifting the values “using previous time steps as input variables and the next time step as the output variable”. We are setting the sliding window size to twelve, as our observations are monthly. Walk forward validation is then performed to test the performance of the models. This gives a sliding window of twelve months of data to create a model from, with this repeating with each step forward. The performance of the folds is then averaged, which in our case resulted in a MAE of 0.014. When the data was scaled using Standard Scaler, it returned a MAE of 0.58 and a RMSE of 0.61.

Since the data collected has a number of features including traffic and weather impacts, and the date specifically has been augmented, it can be approached as a supervised learning problem, with the X data being the features and y being the ‘wait assessment’ target. In such an approach the X and y data is split into training and testing sets and the random forest model fit on the training data. Predictions are then made on the X testing set and then compared to the y testing set to calculate the MAE and RMSE which are both 0.02. When the actuals versus predicted are plotted via a scatter plot, it is seen that the lower values and highest values appear more accurate, with more variance around the upper half of the distribution.

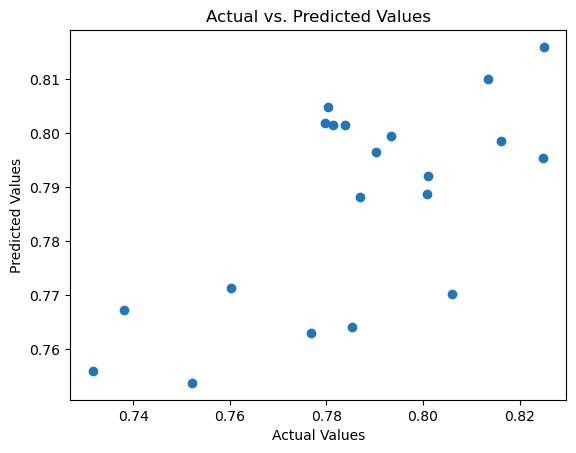


Figure 12: Initial RF predicted versus actual values scatter plot

When the feature importance is plotted, ‘Elapsed’ is the predominant feature with more than twice the importance of the second place feature, the ‘vehicle\_count’.

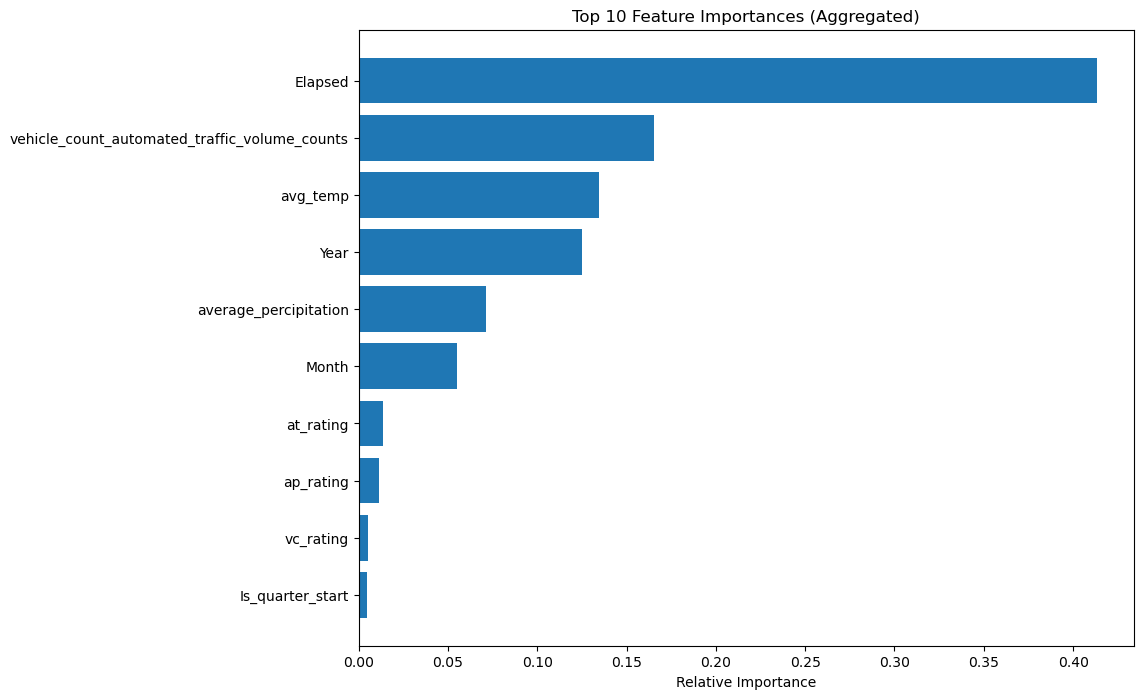


Figure 13: RF feature importance

The data was then scaled using Standard Scaler and a new Random Forest model was run. Predictions using the new model had a MAE of 0.64 and a RMSE of 0.75. There is no clear pattern from plotting the residuals against the actual values.

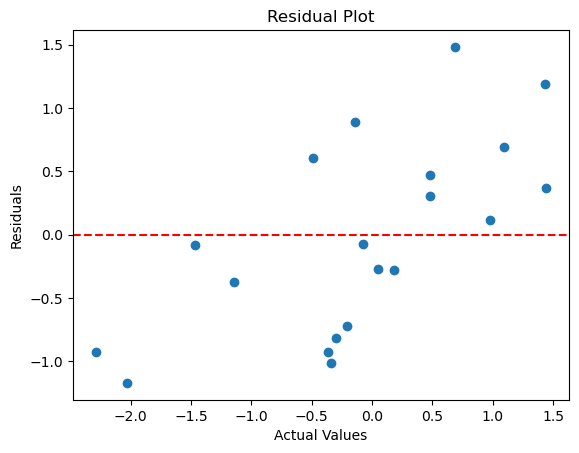


Figure 14: Scaled RF residuals versus actual values scatterplot

The feature importances still had ‘Elapsed’ as the most important feature, but less so than last time with slightly over 0.3, as compared to second place ‘Year’ which had slightly less than 0.25.

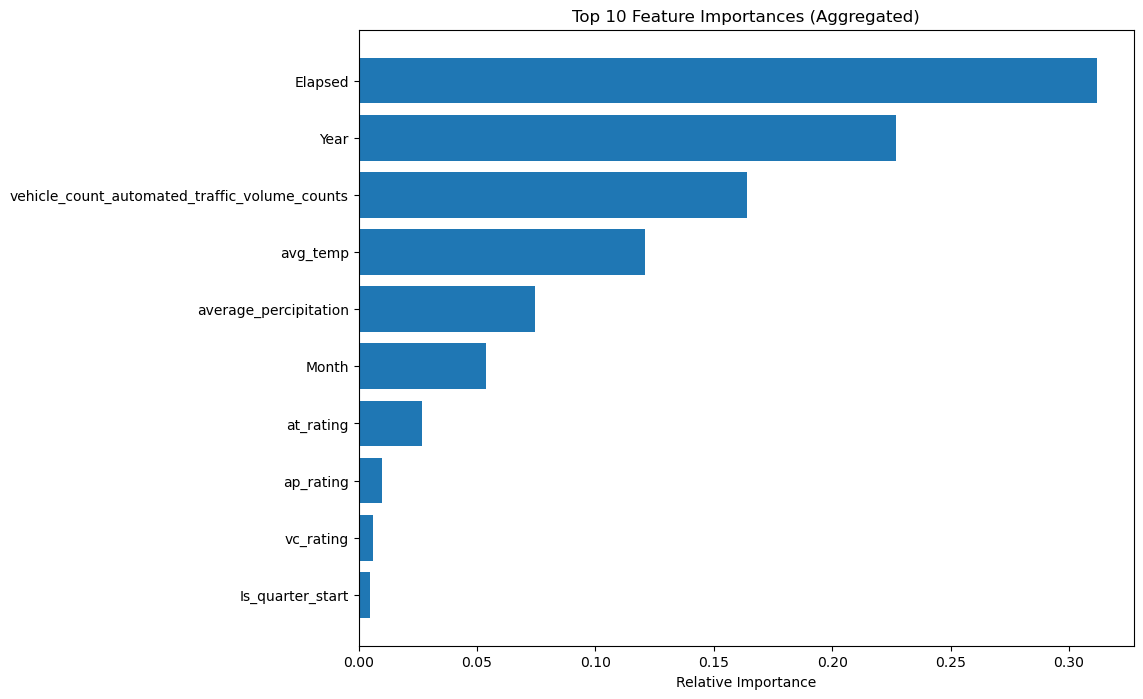


Figure 15: Scaled RF feature importance

In order to tune the parameters of the model, a parameter grid is created, containing differing values for the parameters of: number of estimators, maximum depth, minimum sample split, minimum samples leaf, and maximum features. GridSearchCV is then used to cycle through the different permutations of the parameter grid, and selecting the optimal combination. This ran for 3 minutes and 12 seconds and ran through 1215 different models. For the five different parameters included in the parameter grid, an optimal value was given for each. These parameters were then applied to the next model and it was again trained and tested.

This returns a MAE of 0.49 and a RMSE of 0.67. From the predicted versus actual values scatter plot it can be seen that the lower values have a higher accuracy with larger values having a greater spread.

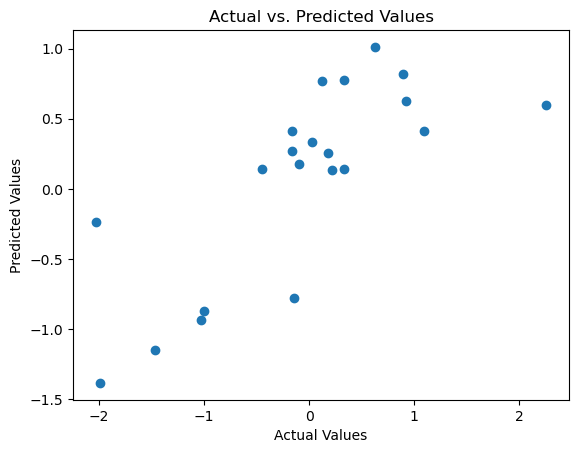


Figure 16: Tuned RF predicted versus actual values scatterplot

## 5.6 Hybrid Model

*Data Processing*

The hybrid model was designed to be a combination of the SARIMAX time series analysis and the Random Forest machine learning model. Previously explored by Xu et al. (2019) this combination of the linear ARIMA model and the ML model takes advantage of the two models for predicting time series. The SARIMAX model is fitted to the original values, and residuals are calculated using the difference between the predicted and original values.

The residuals were taken from the full SARIMAX time series previously calculated and added to the dataframe containing the wait assessment variable and the other features. The data is then scaled using StandardScaler.

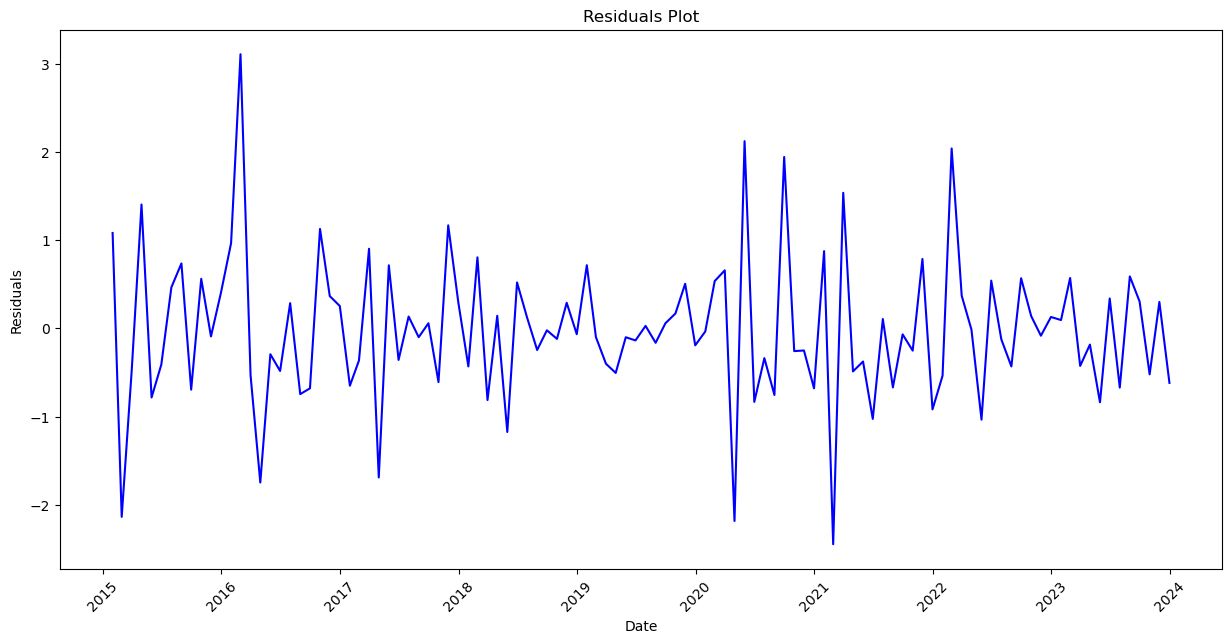


Figure 17: SARIMAX residuals

*Modelling and Parameter Tuning*

The data is split into testing and training sets and the model is fit on the training data. After then using the model to predict the test set a MAE of 0.56 and a RSME of 0.71 is obtained.

The actual versus predicted values show a relatively tight spacing throughout the range of values, with the accuracy diminished by a handful of strong outliers.

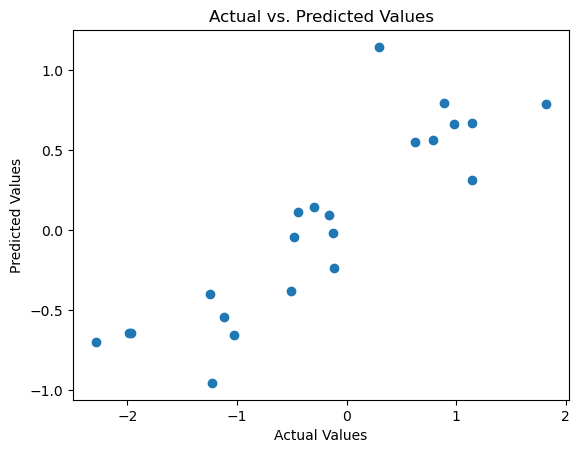


Figure 18: SARIMAX-RF predicted versus actual values

When the feature importance is plotted, it can be seen that the sarimax\_residuals are in fourth place and carry a significant contribution to the model.

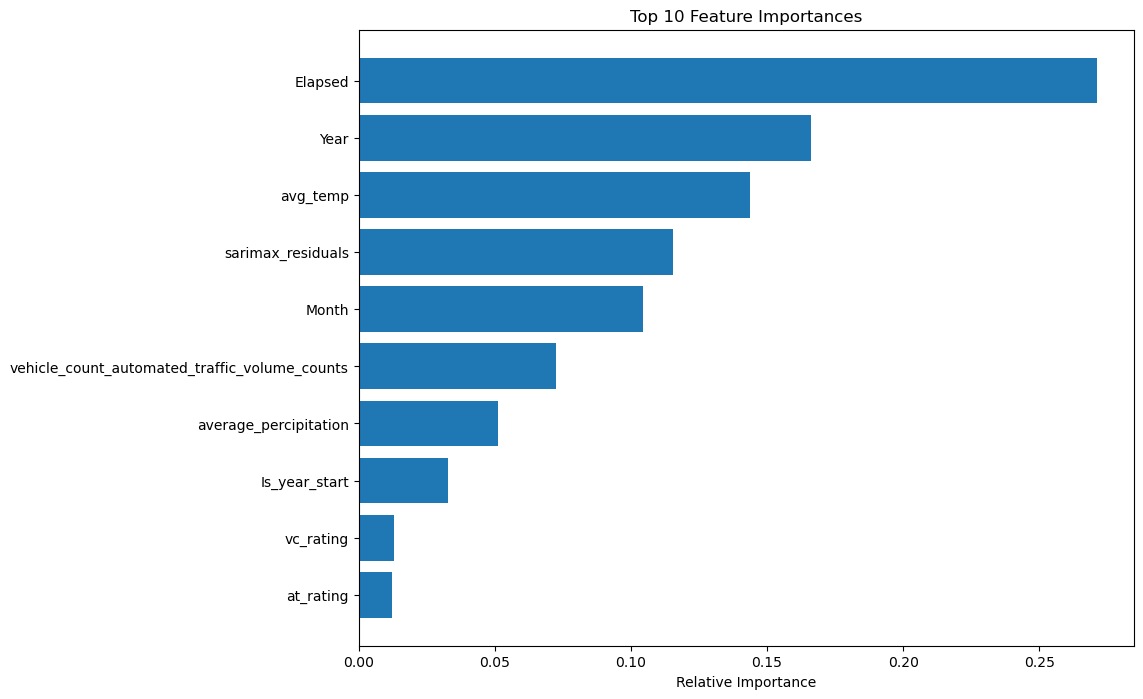


Figure 19: SARIMAX-RF feature importance

After this a parameter optimisation is performed, again using a grid search, to understand the best parameter for the updated dataset. This ran for three minutes and 10 seconds and returned optimum parameters of; ‘max depth’: 10, ‘max features’: sqrt, ‘min samples leaf’: 1, ‘min samples split’: 5, and ‘n estimators’: 100.

These parameters are selected for a new model which is then fit and tested in the same manner as the previous one. It resulted in a MAE of 0.56 and a RMSE of 0.70

## 5.7 Interactive Dashboard

*Data Preparation*

To create an interactive dashboard that would show a time series that is specific to a certain route, the data being used needed to be in a certain format. The coding used for the time series creation would be using for loops, and need the length of each series to be the same. This involved grouping the date by the route id column, and then ensuring that the length of values for each value within the route id column was the same length. Any routes which did not contain the full length needed were removed from the dataset. This left 200 unique routes which each had monthly stepped data from January 1st 2015 to December 31st 2023. The values were scaled using StandardScaler. Although the time series data for each route was not stationary, the process to make it stationary using a rolling mean transformation does not occur here, but rather as part of the loop which creates each route model and stores its predictions.

*Multiple Model and Predictions*

An empty dictionary is created to store the models, predictions and route ids. A for loop is then run which does the following for every set of values for each route id:

* Makes it stationary using a rolling mean
* Ensures the index is set to the correct datetime and frequency
* Defines the endogenous and exogenous variables
* Splits the data into training and test sets
* Fits the model, then makes predictions
* Saves the model, predictions, and route id to the empty dictionary previously created

*Interactive Dashboarding*

An interactive dashboard was created using Dash, an open source library designed for “creating reactive, analytical Web-based applications” (Plotly, 2017). A list is created from the sarimax dictionary holding the models, routes and predictions. The list contains dictionaries with label and value, where the route id is both, with it being stored as a string for the label. This is to be used for the dropdown selection box which will filter the data. The dash app is created, this is what our dashboard will run on. The layout of the dashboard is defined, with a title, the dropdown filter with the options set the the labels and values of the previously created list, and then the graph which will display the times series plot. An app callback is used, this gives an output, the time series graph, which is what is to be changed by the following function. It also gives an input, which is what will trigger the function, changes in the ‘value’ property of the ‘route-id-dropdown’. The function then extracts the relevant test and training values from the sarimax results dictionary, and plots them, with separate lines for the actual training values, actual testing values, and predicted testing values. Finally the app is run locally with a <http://127.0.0.1:8050/> link allowing for interaction in the browser.

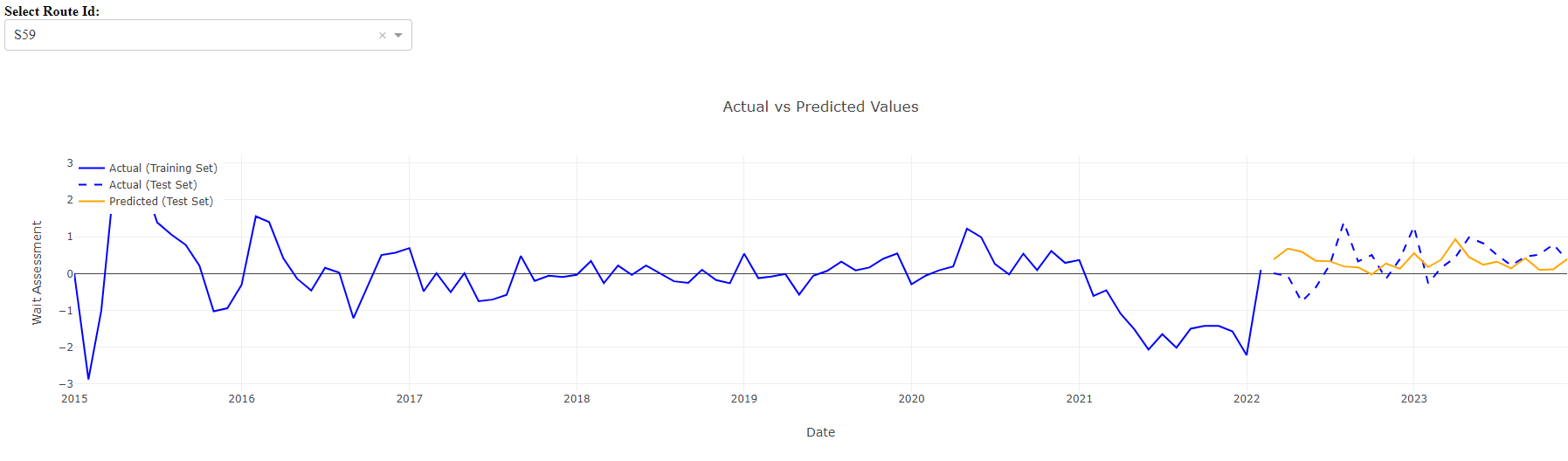


Figure 20: Interactive Time Series Dashboard

## 5.8 Table of Results

The various forecasting models have their results summarised in table 6. The columns cover the model type, the chosen hyperparameters for the model, what, if any, transformation the data has undergone, the variables included in the model, as well as accuracy metrics, in this case the RMSE and the MAE. The first Sarimax result was 0.00 for both RMSE and MAE. This was the best performing, with no transformation or tuning, but is unlikely to be a true reading which will be looked at further in the discussion section. The sliding window RF model with only the wait\_assessment target variable, and the initial untuned RF model with all the features both scored the same RMSE and MAE with 0.17 and 0.14 respectively. This is also unusually low for untuned models when compared to the rest of the results and will again be discussed further. The sliding window RF model using scaled data achieved a MAE of 0.58 and RMSE of 0.61 which put it behind only the tuned SARIMAX in terms of performance. For the remaining models, which also all used scaled data, there was a spread in results with performance seemingly depending on the tuning and refinement of the model as much as the model type itself. The untuned SARIMAX has a RMSE of 0.78, higher than the untuned RF’s of 0.72. After hyperparameter tuning, the RMSE of the SARIMAX model is reduced to 0.41, whereas the RF model showed a worse performance after tuning, with both RMSE and MAE increasing to 0.72 and 0.57 respectively. The SARIMAX-RF showed the best untuned result for the scaled data with a RMSE of 0.63 and a MAE of 0.50. Hyperparameter tuning improved its performance slightly with a RMSE and MAE of 0.62 and 0.49 respectively. The table indicates that while the tuned SARIMAX model is the best performing of the scaled datasets, no one model dominates in terms of performance, with the hybrid SARIMAX-RF having the best untuned performance.

Table 6: Table of results for various models and methods

| **Model Type** | **Hyperparameters** | **Transformation** | **Variables** | **RMSE** | **MAE** |
| --- | --- | --- | --- | --- | --- |
| SARIMAX | (1,0,0) (0,0,0,12) | none | rolling\_mean\_diff, 'average\_percipitation', 'avg\_temp', 'vehicle\_count\_automated\_traffic\_volume\_counts' | 0.00 | 0.00 |
| SARIMAX | (1,0,0) (0,0,0,12) | Scaled | rolling\_mean\_diff, 'average\_percipitation', 'avg\_temp', 'vehicle\_count\_automated\_traffic\_volume\_counts' | 0.78 | 0.59 |
| SARIMAX | (1,0,0) (0,0,1,12) | Scaled | rolling\_mean\_diff, 'average\_percipitation', 'avg\_temp', 'vehicle\_count\_automated\_traffic\_volume\_counts' | 0.41 | 0.34 |
| RF | n\_estimators = 100 | none | wait\_assessment | 0.02 | 0.01 |
| RF | n\_estimators = 100 | Scaled | wait\_assessment | 0.61 | 0.58 |
| RF | n\_estimators = 100 | none | wait\_assessment, 'vehicle\_count\_automated\_traffic\_volume\_counts','average\_percipitation','avg\_temp','vc\_rating', 'ap\_rating', 'at\_rating','Year','Month','Is\_quarter\_start','Is\_year\_start','Elapsed' | 0.02 | 0.01 |
| RF | n\_estimators = 100 | Scaled | wait\_assessment, 'vehicle\_count\_automated\_traffic\_volume\_counts','average\_percipitation','avg\_temp','vc\_rating', 'ap\_rating', 'at\_rating','Year','Month','Is\_quarter\_start','Is\_year\_start','Elapsed' | 0.70 | 0.56 |
| RF | max\_depth = None, max\_features = sqrt, min\_samples\_leaf = 1, min\_samples\_split = 5, n\_estimators = 100 | Scaled | wait\_assessment, 'vehicle\_count\_automated\_traffic\_volume\_counts','average\_percipitation','avg\_temp','vc\_rating', 'ap\_rating', 'at\_rating','Year','Month','Is\_quarter\_start','Is\_year\_start','Elapsed' | 0.72 | 0.57 |
| SARIMAX-RF | n\_estimators = 100 | Scaled | wait\_assessment, 'vehicle\_count\_automated\_traffic\_volume\_counts','average\_percipitation','avg\_temp','vc\_rating', 'ap\_rating', 'at\_rating','Year','Month','Is\_quarter\_start','Is\_year\_start','Elapsed', 'sarimax\_residuals | 0.63 | 0.50 |
| SARIMAX-RF | max\_depth = None, max\_features = None, min\_samples\_leaf = 2, min\_samples\_split = 2, n\_estimators = 100 | Scaled | wait\_assessment, 'vehicle\_count\_automated\_traffic\_volume\_counts','average\_percipitation','avg\_temp','vc\_rating', 'ap\_rating', 'at\_rating','Year','Month','Is\_quarter\_start','Is\_year\_start','Elapsed', 'sarimax\_residuals | 0.62 | 0.49 |

# 

# **Discussion**

Part of this study was to compare the use of linear, nonlinear and hybrid models when it came to forecasting public bus wait assessment values. It was to examine which method offered the greatest accuracy, with attention being given to exogenous variables and a sliding window method also tested. All of the models which were proposed to be implemented were created, and the results ultimately showed that a tuned SARIMAX model using weather and traffic exogenous variables was the model successful when it came to forecasting values. This was despite the inclusion of a more sophisticated hybrid model, which in the literature review, generally performed better than its single model counterparts.

What was not successful was the initial SARIMAX and RF models which contained unscaled data, the results here are recommended to be disregarded as false for a number of reasons. Upon running the initial SARIMAX, a warning regarding the covariance matrix being singular is displayed, with the advice that standard errors may be unstable. Looking at the research this may be an indication that the data is not regular, something which scaling it fixes. A ridge regression or a least absolute shrinkage and selection operator (LASSO) could be applied to attempt to resolve this. While a RF does not involve a covariance matrix, a similar issue of overfitting is assumed for the unscaled data models created there.

The hybrid model performed better than the standard RF series, so it would seem that the inclusion of the SARIMAX residuals did have a positive effect, just not to the extent that was anticipated. In the literature review a large number of studies focused on the creation of hybrid models using ARIMA models and deep learning or SVM type models, and it would seem that the latter two may offer better performance than the RF, although further research would be needed to accurately compare the two. The study cannot recommend using a hybrid SARIMAX-RF model as the results obtained cannot be said to be worth the process involved in creating it, especially when simpler models have proven more effective.

The tuned SARIMAX model was clearly the best performing and so the recommendation would be to use this as the model of choice for predicting wait assessment values. Interestingly the sliding window model performed well on the scaled data and was second only to the tuned SARIMAX. There are a few reasons to take note of this as, even though the accuracy is not particularly close to that of the SARIMAX, if it could be refined there would be good reason to select it. Firstly the SARIMAX model was trained using exogenous variables, which adds another level of complexity to maintaining the model. The sliding window approach uses only the past wait assessment values so not additional data gathering is required. Secondly, the sliding window method means that as the size of the dataset grows, the computational expense for the model will not, as will intake the most recent value, expelling the least recent and maintaining the window size. Ultimately without improvements being made the accuracy of the SARIMAX model means it is still the recommended method at this time.

There were, of course, limitations to this study, in a number of different areas. Firstly as can be seen in the above discussion, a number of models were disregarded due to what was considered inconclusive results. While such results can occur in a study of this kind, with more experience it may have been avoided, and with more time it may have been explored and explained further. The time series itself was supplied in monthly format which for some domains could be considered high-frequency, but for predicting public transport delays having a shorting time frequency would allow for more accurate short-term forecasting. In practice the transportation companies themselves are likely to have access to more recent and higher frequency data, this is just what is aggregated and released publicly. In other data supply limitations, having a more comprehensive traffic dataset could have added more predictive power, the factor chosen was ultimately the most complete dataset, were others available they may have held more relevance. Finally, while the primary data was useful in helping to identify the factors involved as well as the need for a study of this kind, interviews with experts working in the field would also have enhanced the understanding of the area. They were requested but remained unanswered at the time of publishing.

Future studies would benefit from a larger dataset size, and with the knowledge that the SARIMAX-RF model did not perform as expected could look at using neural networks or SVM’s as a replacement for the RF. Another area of interest would be refining the sliding window method to see if there is the possibility of boosting performance, as it is a promising model with good performance on relatively little input.

# 

# **Conclusion**

This study set out to investigate the best method of predicting the wait assessment value, comparing a number of forecasting models. This broke down to a comparison of SARIMAX, RF, and SARIMAX-RF hybrid models looking at which trained model performed best on the testing data. To inform the study, a literature review was conducted, which validated the necessity for the study and the features involved. The literature review also gave an overview of the current state of the art as regards to the models and methods we were looking to employ. Specifically it helped form the process for implementing the sliding window method, as well as creating the SARIMAX and SARIMAX-RF models by looking at comparable efforts and the results.

The data which the study is based on came from primary and secondary sources. Primary sources were public transportation users who are directly impacted by the subject matter and could benefit from the improvement of services. It provided insight into the factors that they felt influenced delays, identifying the extra variables to be included. Secondary sources were datasets containing time series data for the wait assessment values, as well as a number of different weather and traffic conditions.

EDA was performed on all datasets and included standard checks such as visualising histograms and correlation heatmaps. The statistical properties of the data were explored and verified using tests such as the ADF and Shapiro-Wilk. The analysis helps to make the data more suitable for the modelling process by exposing areas where assumptions or requirements did not fit and transformation was required. This transformation was done through methods such as a rolling mean difference and a StandardScaler. In addition, PCA was performed on the traffic and weather datasets to inform which features were most relevant to the successful forecasting of the wait assessment values. Comparisons of the performance of boroughs by peak and off-peak periods was conducted using a two-way ANOVA and Kruskal-Walis tests. Time series charts were created and an interactive dashboard was created to display the forecast predictions for each route, selectable in a drop down filter, and compare them to the actual values.

Several models were created to produce a forecast of the wait assessment data, with SARIMAX and RF being the two main model types. A hybrid SARIMAX-RF model was also created to attempt to capture both linear and nonlinear patterns. These were optimised using hyperparameter tuning to find the ideal combination of parameters and an alternative RF model trained on a sliding window method was also created in an attempt to create a model efficient model.

With the question of the study being which forecasting model for bus wait assessment values is the best performing, the results of the models described above showed that the tuned SARIMAX model using vehicle volume counts, average temperature, and average precipitation as exogenous variables was the best performing option. The SARIMAX-RF model performed better than the individual RF models and untrained SARIMAX but was not as effective as had been hoped. The sliding window method applied to the RF model showed promising results in terms of the simplicity of the input variables and the self limiting model size, with results behind the tuned SARIMAX but ahead of all other configurations.

The study was limited by a lack of access to primary data sources regarding the current best practices in an active transportation company, further research which had an understanding of modelling techniques in the field could adapt to optimise the methods as opposed to finding the best based on similar domain research. Hybrid modelling could be further explored with different model selection, as it is such a large field, but the sliding window method offered the results that seemed most appropriate for followup study, applying the method to different models with a greater understanding of the parameter tuning.

Concluding the study, we can take that model optimisation and data transformation methods offer strong forecasting ability even when compared to more sophisticated approaches. Currently a SARIMAX model such as the one described would be recommended for helping to predict wait assessment values, but following some of the recommendations could lead to increased performance and a benefit to public transportation companies and users.

# **A Model Plots**

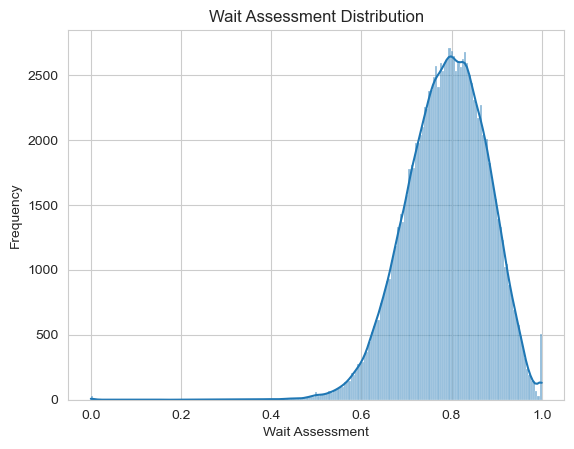


Figure 1: Wait Assessment kdf distribution

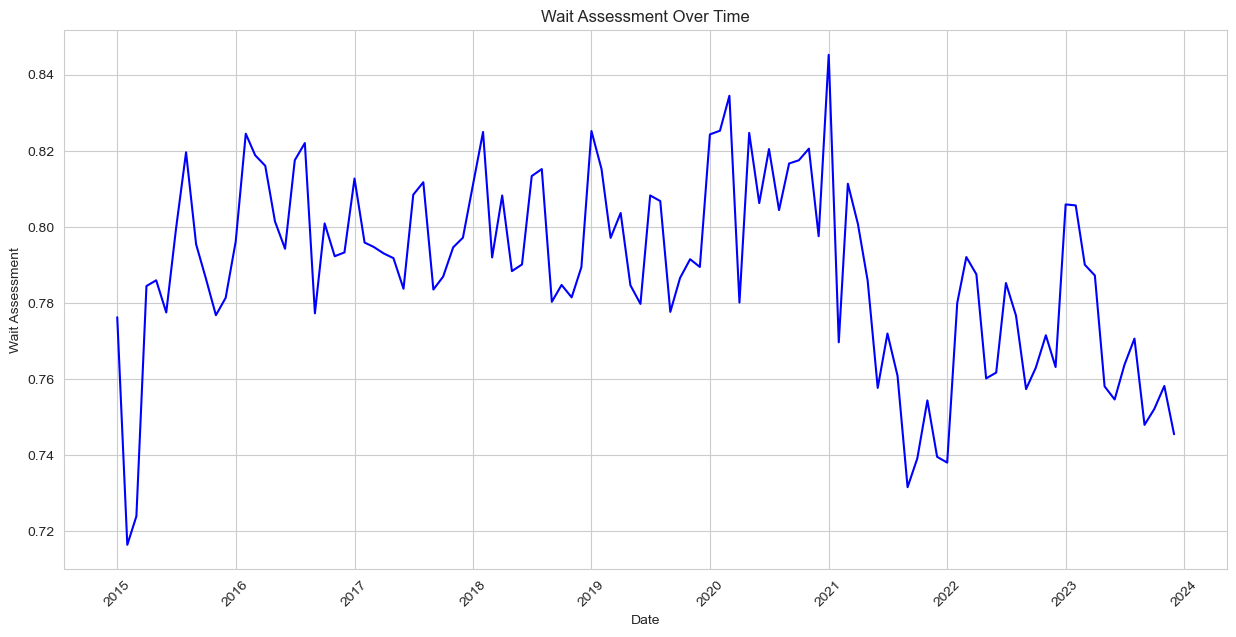


Figure 2: Wait Assessment over time

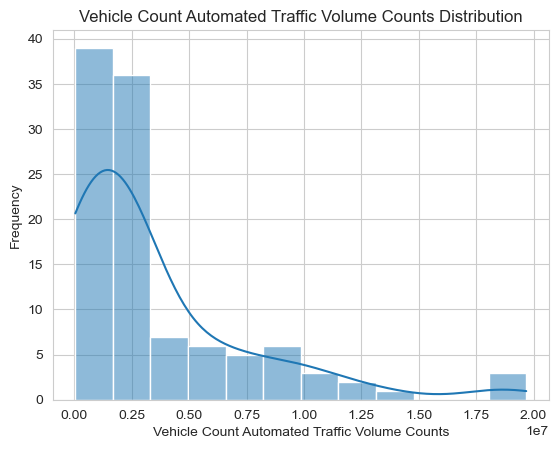


Figure 3: Vehicle Count distribution

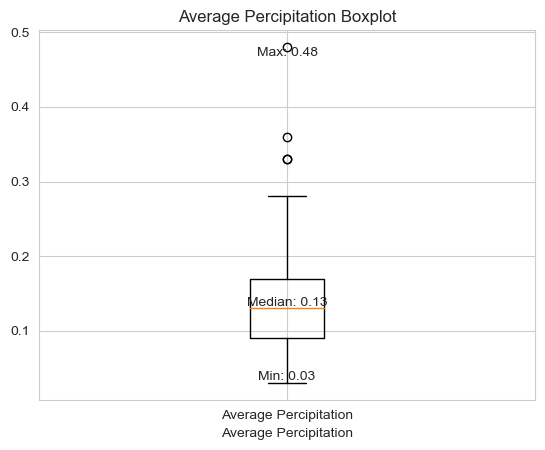


Figure 4: Average precipitation boxplot

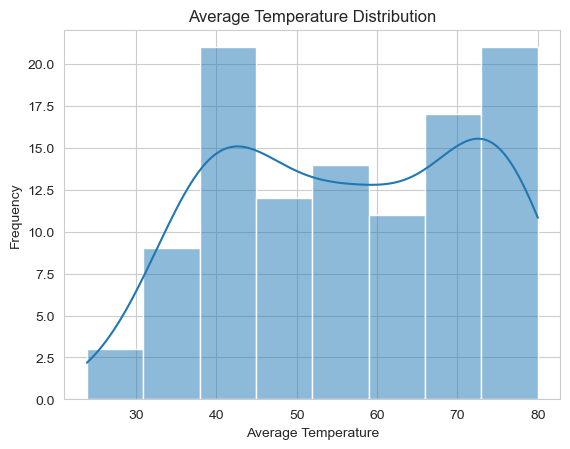


Figure 5: Bimodal Average temperature distribution

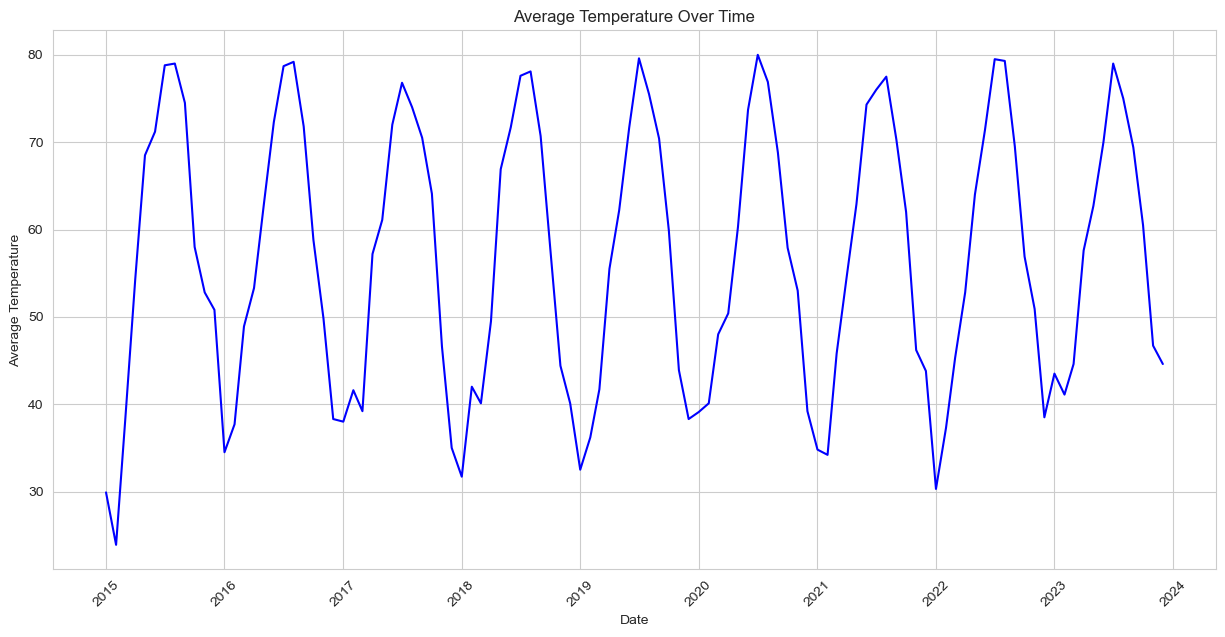


Figure 6: Average temperature over time

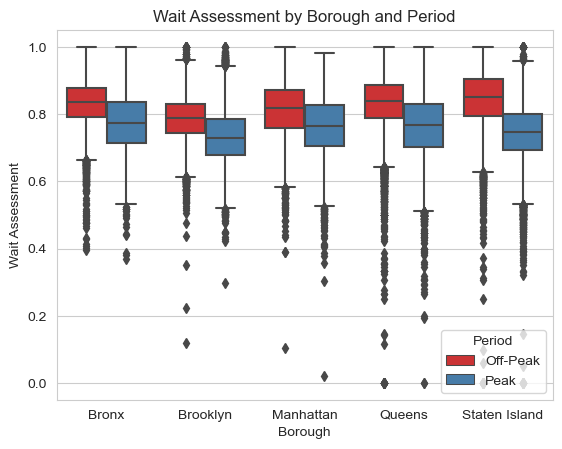


Figure 7: Wait assessment boxplot for borough and period

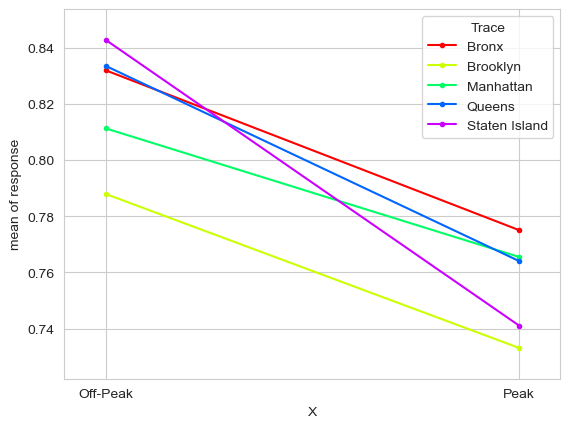


Figure 8: Wait Assessment Interaction Chart for Borough and Period

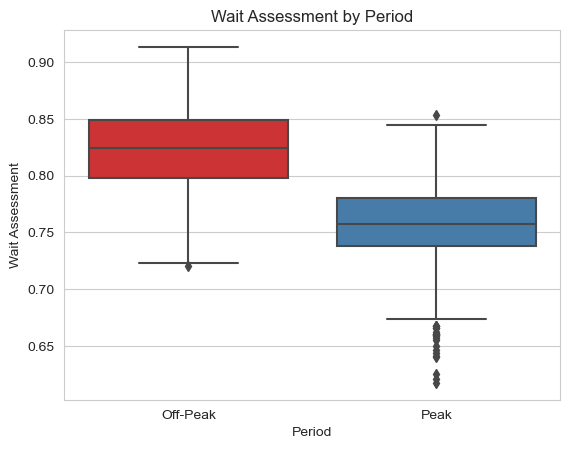


Figure 9: Boxplot of wait assessment values for period

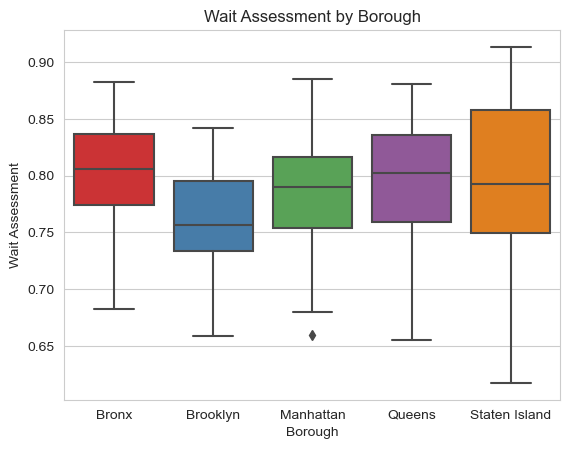


Figure 10: Boxplot of wait assessment values for borough

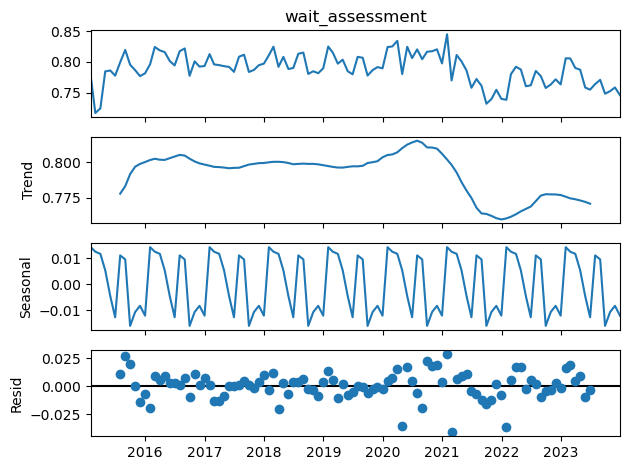


Figure 11: Decomposition of wait assessment

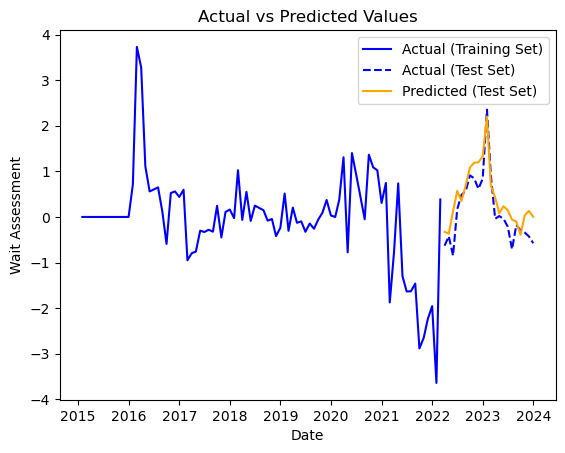


Figure 11: Forecasted SARIMAX values

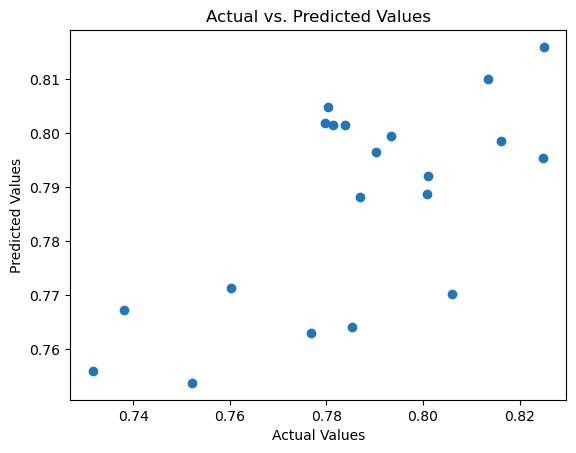


Figure 12: Initial RF predicted versus actual values scatter plot

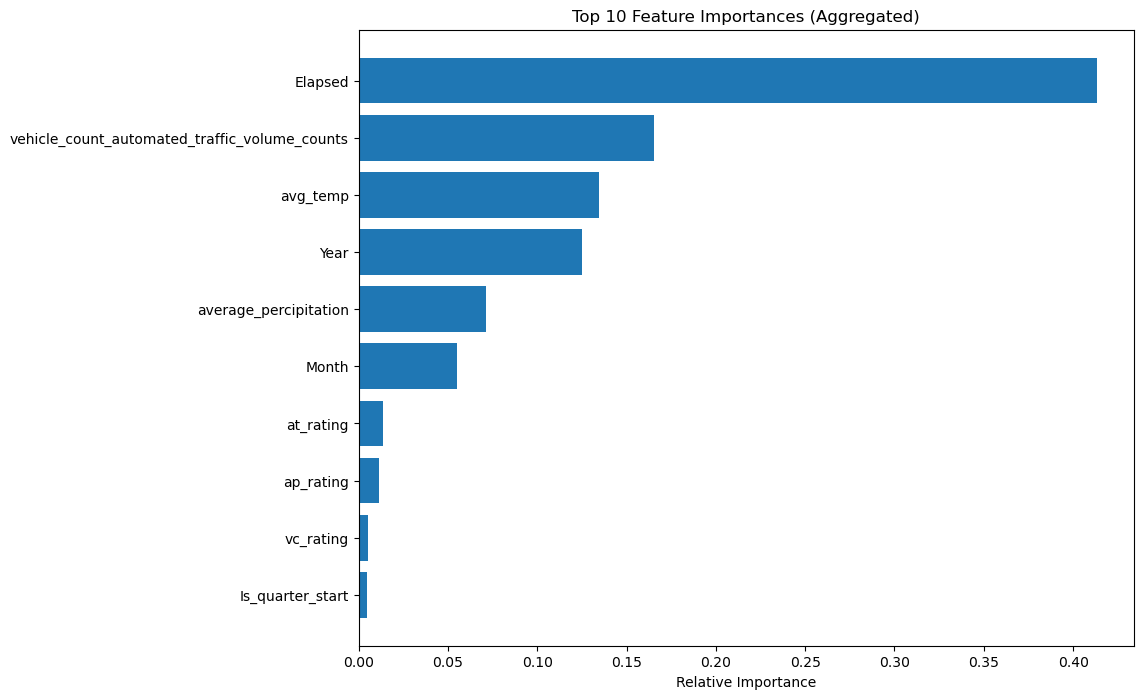


Figure 13: Initial RF feature importance

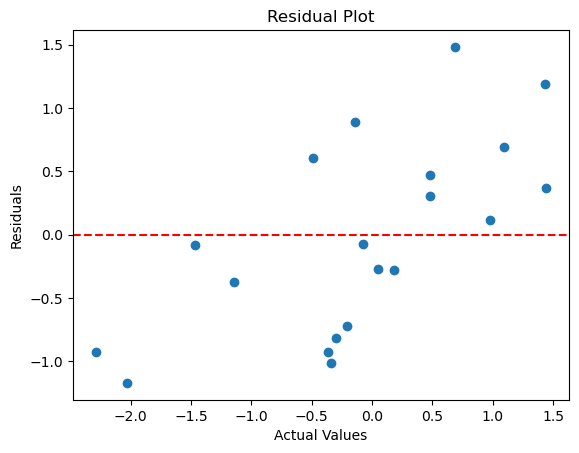


Figure 14: Scaled RF residuals versus actual values scatterplot

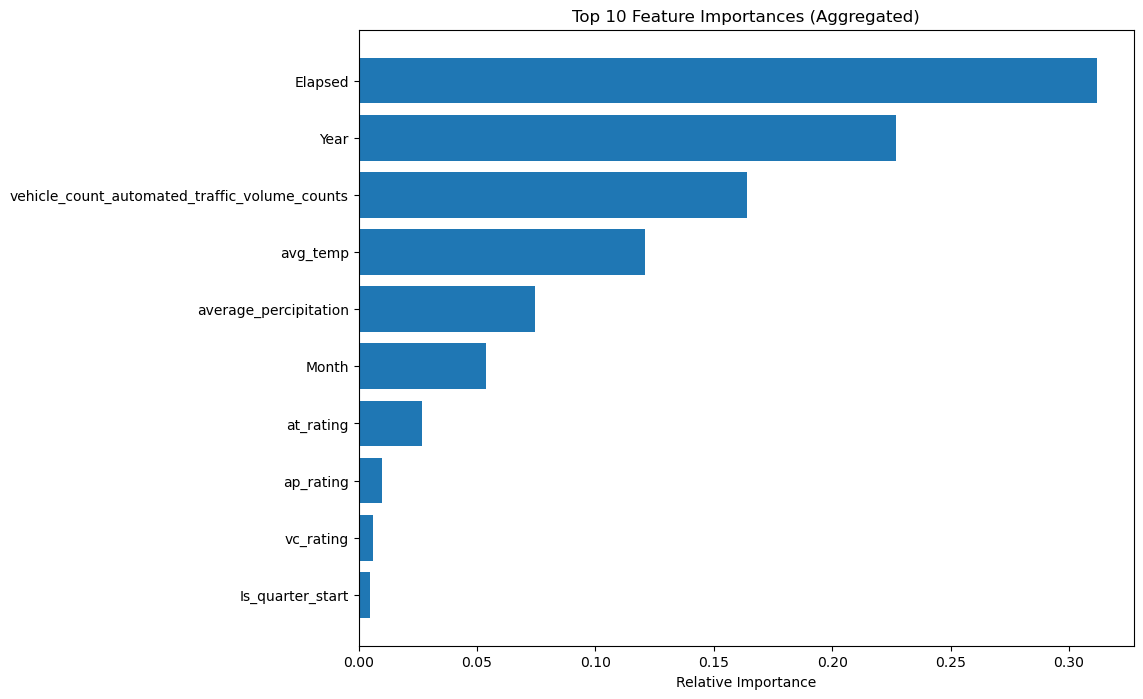


Figure 15: Scaled RF feature importance

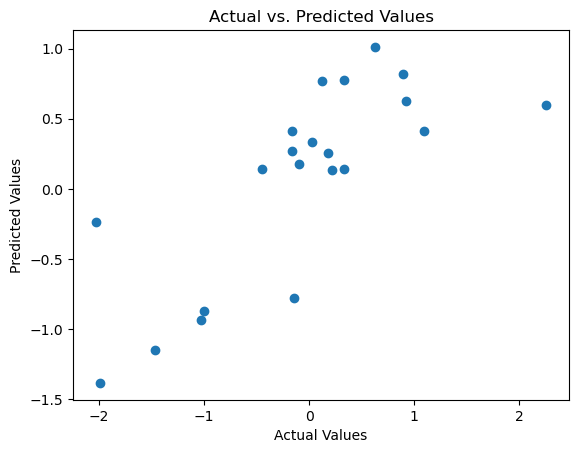


Figure 16: Tuned RF predicted versus actual values scatterplot

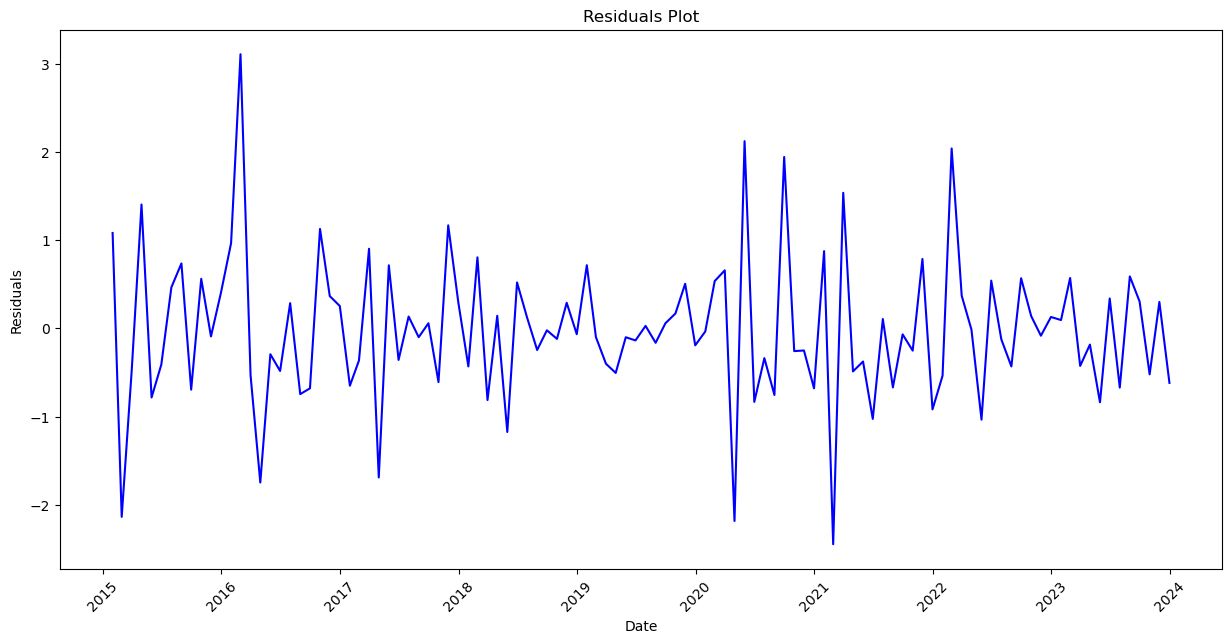


Figure 17: SARIMAX residuals

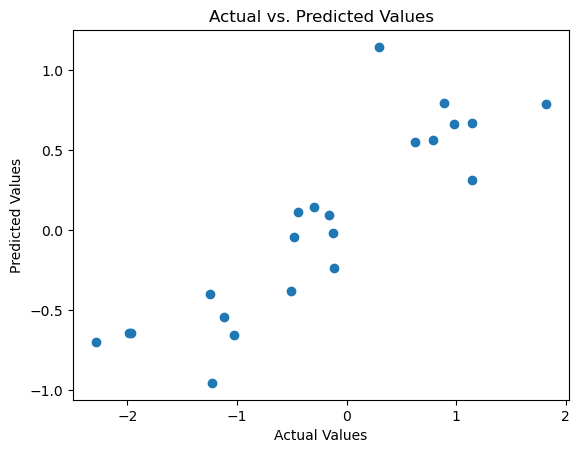


Figure 18: SARIMAX-RF predicted versus actual values

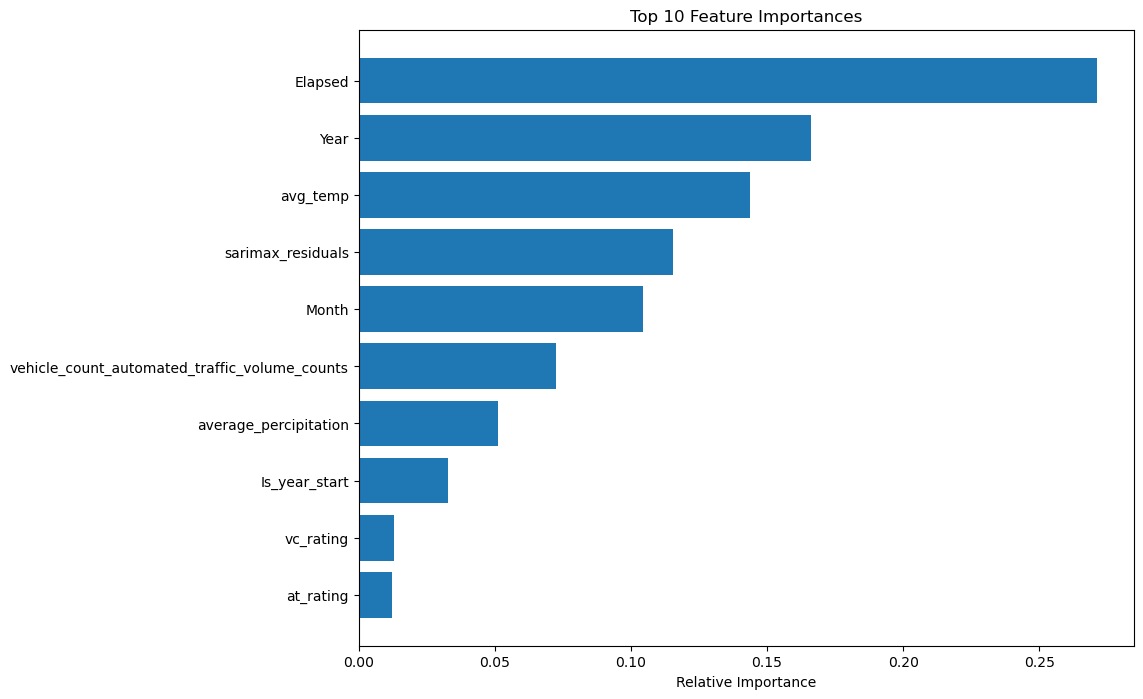


Figure 19: SARIMAX-RF feature importance

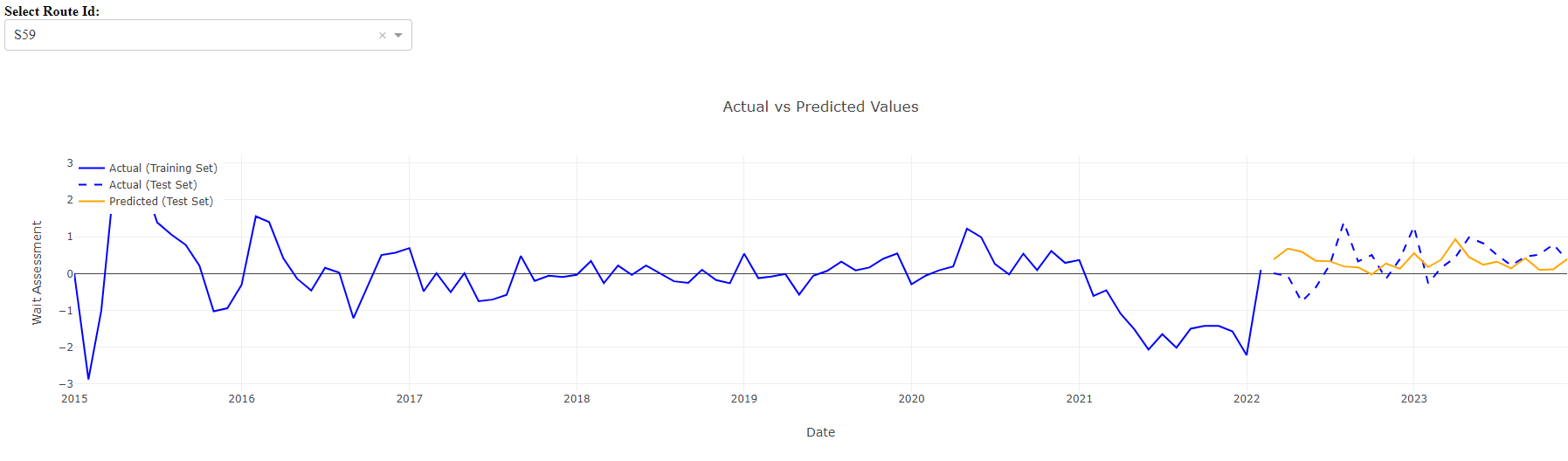


Figure 20: Interactive Time Series Dashboard

# **B Transcripts**

**Transcript A**

**Speaker 1:**

How often would you say you use the bus, if you had to categorise it? Daily, weekly, monthly and so on.

**Speaker 2:**

I would say I use it weekly, like every week you mean?

**Speaker 1:**

Yes, I mean do you use it multiple times per week for example?

**Speaker 2:**

I would use it on weekdays and maybe four days a week.

**Speaker 1:**

Okay and what is the reason for using the bus, in terms of is it for work, is it for school?

**Speaker 2:**

I use it for work, I use it to go to some of the houses that I clean.

**Speaker 1:**

So you use it to get to work and is it the only way you get to work or do you have other ways?

**Speaker 2:**

No, I use the bus, sometimes my boyfriend drops me, and I also have a scooter if it's close to where I live.

**Speaker 1:**

And what kind of what determines whether you use the bus or not,what goes into that decision?

**Speaker 2:**

Well I use it [the bus] normally for jobs where I've been doing it for a while, it's regular and it's on the route. My boyfriend will drop me to places that are way outside of the way if he can, and I normally just use the scooter for places closeby when I coming back home again afterwards.

**Speaker 1:**

Ah okay so it kind of depends on the location?

**Speaker 2:**

Yeah the location and also how much I know where I'm going. I wouldn't normally take the bus in the first journey because I'm not sure where it's going to drop me.

**Speaker 1:**

And what are some issues you would find with the buses, are there things that you see that are negative in the experience of using them.

**Speaker 2:**

Well there's some antisocial behaviour.

**Speaker 1:**

Okay.

**Speaker 2:**

And then it's hard sometimes to know when a bus is going to come.

**Speaker 1:**

So reliability.

**Speaker 2:**

Yes.

**Speaker 1:**

So would you say that delays in bus schedules affect you?

**Speaker 2:**

Oh yeah it can be very frustrating.

**Speaker 1:**

And how does that impact you, is it in your work, just that it makes you late?

**Speaker 2:**

Yes, it can make me late. I have to call the house owners and tell them I'm running late normally it means I have to get to work done in less time because I already have another appointment scheduled for afterwards, it can make things quite complicated it's why I generally don't use the bus unless I know it's a fairly reliable route.

**Speaker 1:**

I see, so would having an advanced warning of the delays be beneficial to you?

**Speaker 2:**

You mean knowing if there's how long it's going to take, like on the bus stop sign?

**Speaker 1:**

Yes, I suppose more so knowing that he on a certain day or at a certain period it's likely that the bus is going to be delayed.

**Speaker 2:**

Maybe, it's just the things are so unreliable you can never really be sure if it's telling you what is really going to happen, it'll say that the buses coming and then five minutes later it'll say the bus won't be here for another 30 minutes.

**Speaker 1:**

I see okay and what are the causes of the delays in your view, why are the buses late?

**Speaker 2:**

I don't know, it depends. Normally it's when there's really bad weather that I feel it the most. When it's snowing it seems like all of the buses just stop working. The same for when it's raining, buses can be full and pass you by without stopping or just not come, until 2 or 3 come all at the same time.

**Speaker 1:**

If there were less delays in the scheduling and more reliability do you feel like you would use the bus service more?

**Speaker 2:**

I suppose so but I suppose so but I'd have to really trust it. If I’m using it to get to work and I'm late, it really messes up my work and the rest of my day as well. As I said I don't like to use it unless I'm pretty confident that it's gonna get me there on time.

**Speaker 1:**

Thank you very much for your time.

**Transcript B**

**Speaker 1:**

Thanks for talking to me so I'm gonna start off asking you how often, in general, you would say you use the bus on a daily, weekly, monthly basis?

**Speaker 2:**

I use it daily. I use it to get into work so I would use it in the mornings and then coming out of work in the evening time so twice a day. I go to the office three days a week so I guess roughly six times per week

**Speaker 1:**

Great so that kind of covers my next question around the reasons for your journey, so you use it mainly for work?

**Speaker 2:**

Yes I don't generally use it outside of that it's mainly for getting to and from work in Swords and I live in Donabate so it's a short enough trip.

**Speaker 1:**

Okay and do you have alternative transportation?

**Speaker 2:**

My wife and I share a car, she generally uses it for work since she works in Blanchardstown which is further away and I don't mind taking the bus, I don't particularly like driving.

**Speaker 1:**

So when you are taking your bus journeys do you feel the impact of scheduling and bus delays at all? Is it ever an issue for you?

**Speaker 2:**

Yeah it definitely can be, it's not normally as bad as people say. You know you always want to get to the bus stop 5 minutes before the bus says it's going to be there and assume it's gonna get there 10 minutes after it's gonna be there so there is generally a window around when you think it's going to appear, and the real time scheduling, I find, can can help but it's not always accurate so as I said sometimes it'll tell you the bus is 5 minutes away and then it'll be just one minute away and you'll have missed it. I tend to get there early and with the knowledge that I might be waiting then for a little bit, but I do also, just in case, I also get the bus that will get me there at 8:40 as opposed to a bus that will get me there at five to nine just because I know with that window that I can't really rely on it. I need to take the earlier bus.

**Speaker 1:**

So you kind of factor delays into your day as you know that there might be some?

**Speaker 2:**

Yeah.

**Speaker 1:**

And if there is, is that the full extent of them?

**Speaker 2:**

Well it's generally a lot worse in the winter time. Sometimes I'll have to get my wife to drop me down because it's just, there's no bus, or they're just delayed to the point where the real time information is useless. It doesn't really tell you anything, you might as well guess when the bus is going to arrive.

**Speaker 1:**

So winter weather has a noticeable effect.

**Speaker 2:**

Yeah it does.

**Speaker 1:**

It sounds like you use the bus and you're fairly happy with the service, enough to use it, for work. But you said you didn't use it in your personal life really?

**Speaker 2:**

Not really no, generally for going somewhere my wife or I will drive, it's just we go places where to get a bus would just take some much longer and be so much less convenient.

**Speaker 1:**

So it's a convenience factor?

**Speaker 2:**

Yeah

**Speaker 1:**

If you thought the buses were more reliable, if there was less delays, would that encourage you to use it more?

**Speaker 2:**

I suppose so, going in and out of the city centre. I would tend to rely more on a train and Luas than a bus just because I find them to be slightly more reliable so maybe then. Unless it's a regular journey you know, I might take the bus for something I'm doing more than more than just once, just not necessarily for one off things, you know, on the weekend with my family kind of thing

**Speaker 1:**

Okay thank you very much

**Transcript C**

**Speaker 1:**

How often would you use the bus, daily, weekly, monthly?

**Speaker 2:**

A few times a year two times a year maybe

**Speaker 1:**

Once a month?

**Speaker 2:**

Or less than that even, maybe once a month or every two months

**Speaker 1:**

What is the main purpose of your journey?

**Speaker 2:**

To go into the city normally, on weekends

**Speaker 1:**

So for personal use, shopping, restaurants, things like that.

**Speaker 2:**

Yes.

**Speaker 1:**

Do you have alternative transportation?

**Speaker 2:**

Yes I have a car.

**Speaker 1:**

And so on occasions where you could drive, why do you choose to use the bus

**Speaker 2:**

Oh for pleasure, for not having to drive during the weekend. I drive all week to work.

**Speaker 1:**

What are the causes of delays in your opinion?

**Speaker 2:**

I believe it's because of the lack of staff.

**Speaker 1:**

Okay

**Speaker 2:**

That's what I hear from colleagues who need to take the bus to work

**Speaker 1:**

If you knew about delays that were going to happen ahead of time would that be helpful to you?

**Speaker 2:**

I'm still be late or I would have to go earlier I suppose

**Speaker 1:**

And then if you found buses to be more reliable would you use them more or is it not really a factor?

**Speaker 2:**

I don’t think it would change how I use them right now, it might just make it easier.

**Speaker 1:**

Thank you.

**References**

Alberg, D. and Last, M. (2018). Short-term load forecasting in smart meters with sliding window-based ARIMA algorithms. *Vietnam Journal of Computer Science*, [online] 5(3-4), pp.241–249. doi:https://doi.org/10.1007/s40595-018-0119-7.

Alharbi, F.R. and Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. *Inventions*, 7(4), p.94. doi:https://doi.org/10.3390/inventions7040094.

Allen, P.G. and Fildes, R. (2001). Econometric Forecasting. *International Series in Operations Research & Management Science*, pp.303–362. doi:https://doi.org/10.1007/978-0-306-47630-3\_15.

Ashoke Kumar Biswas, Sina Ibne Ahmed, Temitope Bankefa, Ranganathan, P. and Hossein Salehfar (2021). Performance Analysis of Short and Mid-Term Wind Power Prediction using ARIMA and Hybrid Models. doi:https://doi.org/10.1109/peci51586.2021.9435209.

Athanasopoulos, G., Hyndman, R.J., Song, H. and Wu, D.C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, 27(3), pp.822–844. doi:https://doi.org/10.1016/j.ijforecast.2010.04.009.

Baeli, V. (2021). *Public transport disruptions: How to keep passengers’ trust*. [online] PTV Blog. Available at: https://blog.ptvgroup.com/en/city-and-mobility/public-transport-disruptions-how-to-keep-passengers-trust/ [Accessed 12 Nov. 2023].

Boata, R.St. and Gravila, P. (2012). Functional fuzzy approach for forecasting daily global solar irradiation. *Atmospheric Research*, 112, pp.79–88. doi:https://doi.org/10.1016/j.atmosres.2012.04.011.

Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Brownlee, J. (2016). *How To Backtest Machine Learning Models for Time Series Forecasting*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/ [Accessed 13 Feb. 2024].

Brownlee, J. (2017). *How to Decompose Time Series Data into Trend and Seasonality*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/ [Accessed 11 Feb. 2024].

Brownlee, J. (2018). *Taxonomy of Time Series Forecasting Problems*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/taxonomy-of-time-series-forecasting-problems/ [Accessed 18 Feb. 2024].

Brownlee, J. (2020). *How to Use StandardScaler and MinMaxScaler Transforms in Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/ [Accessed 11 Feb. 2024].

Cao, J.C. and Cao, S.H. (2006). Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis. *Energy*, 31(15), pp.3435–3445. doi:https://doi.org/10.1016/j.energy.2006.04.001.

Cha Zhang and Ma, Y. (2012). *Ensemble Machine Learning*. New York: Springer.

Chen, C., Zhang, Q., Kashani, M.H., Jun, C., Bateni, S.M., Band, S.S., Dash, S.S. and Chau, K.-W. (2022). Forecast of rainfall distribution based on fixed sliding window long short-term memory. *Engineering Applications of Computational Fluid Mechanics*, 16(1), pp.248–261. doi:https://doi.org/10.1080/19942060.2021.2009374.

Cowpertwait, P.S.P. and Metcalfe, A.V. (2009). *Introductory Time Series with R*. New York, Ny: Springer.

Date, S. (2022). *A Guide To Exogenous And Endogenous Variables*. [online] Time Series Analysis, Regression, and Forecasting. Available at: https://timeseriesreasoning.com/contents/exogenous-and-endogenous-variables/ [Accessed 18 Feb. 2024].

Davtyan, A., Rodin, A., Muchnik, I. and Alexey Romashkin (2020). Oil production forecast models based on sliding window regression. *Journal of Petroleum Science and Engineering*, 195, pp.107916–107916. doi:https://doi.org/10.1016/j.petrol.2020.107916.

Deo, R.C., Wen, X. and Qi, F. (2016). A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Applied Energy*, 168, pp.568–593. doi:https://doi.org/10.1016/j.apenergy.2016.01.130.

Dong, H., Guo, X., Reichgelt, H. and Hu, R. (2020). Predictive power of ARIMA models in forecasting equity returns: a sliding window method. *Journal of Asset Management*, 21(6), pp.549–566. doi:https://doi.org/10.1057/s41260-020-00184-z.

Feldman, K. (2022). *Homogeneity of Variance and Statistical Inference: What You Need to Know*. [online] isixsigma.com. Available at: https://www.isixsigma.com/dictionary/homegeneity-of-variance/ [Accessed 11 Feb. 2024].

Frost, J. (2018). *Central Limit Theorem Explained*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/basics/central-limit-theorem/ [Accessed 10 Feb. 2024].

Frost, J. (2020a). *Time Series Analysis Introduction*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/time-series/time-series-analysis-introduction/ [Accessed 17 Feb. 2024].

Frost, J. (2020b). *Using Moving Averages to Smooth Time Series Data*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/time-series/moving-averages-smoothing/ [Accessed 17 Feb. 2024].

Geert van den Berg, R. (2023). *SPSS Shapiro-Wilk Test - Quick Tutorial with Example*. [online] SPSS-Tutorials. Available at: https://www.spss-tutorials.com/spss-shapiro-wilk-test-for-normality/#shapiro-wilk-test-interpretation [Accessed 10 Feb. 2024].

Guermoui, M., Melgani, F., Gairaa, K. and Mekhalfi, M.L. (2020). A comprehensive review of hybrid models for solar radiation forecasting. *Journal of Cleaner Production*, 258, p.120357. doi:https://doi.org/10.1016/j.jclepro.2020.120357.

Hofmann, M.J. and O’Mahony, M.J. (2005). The impact of adverse weather conditions on urban bus performance measures. *IEEE Intelligent Transportation Systems*. [online] doi:https://doi.org/10.1109/itsc.2005.1520087.

Hyndman, R.J. and Athanasopoulos, G. (2018). *Forecasting : Principles and Practice*. 2nd ed. Heathmont, Vic.: Otexts.

IBM (2023). *What is principal component analysis? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/principal-component-analysis [Accessed 10 Feb. 2024].

Ji, W. and Chee, K.C. (2011). Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. *Solar Energy*, 85(5), pp.808–817. doi:https://doi.org/10.1016/j.solener.2011.01.013.

Jiménez-Pérez, P.F. and Llanos Mora-López (2014). Modeling Daily Profiles of Solar Global Radiation Using Statistical and Data Mining Techniques. *Lecture Notes in Computer Science*, pp.155–166. doi:https://doi.org/10.1007/978-3-319-12571-8\_14.

Kapoor, P. and Bedi, S.S. (2013). Weather Forecasting Using Sliding Window Algorithm. *ISRN Signal Processing*, 2013, pp.1–5. doi:https://doi.org/10.1155/2013/156540.

Kumar, M. and Thenmozhi, M. (2014). Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting and Finance*, 5(3), p.284. doi:https://doi.org/10.1504/ijbaaf.2014.064307.

Li, J., Shao, X. and Zhao, H. (2018). An Online Method Based on Random Forest for Air Pollutant Concentration Forecasting. doi:https://doi.org/10.23919/chicc.2018.8483621.

Mellit, A., Benghanem, M., Arab, A.H. and Guessoum, A. (2005). A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach. *Solar Energy*, 79(5), pp.469–482. doi:https://doi.org/10.1016/j.solener.2004.12.006.

Metropolitan Transportation Authority (2022). *Metropolitan Transportation Authority Mission Statement, Measurements, and Performance Indicators Report Covering Fiscal Year 2021 In Compliance with New York State Public Authorities Law §1269-f and §2824-a Submitted as Part of the MTA 2021 Annual Report to the Governor*. [online] Available at: https://new.mta.info/document/80026 [Accessed 23 Feb. 2024].

Mitrani, A. (2020). *Achieving Stationarity With Time Series Data*. [online] Medium. Available at: https://towardsdatascience.com/achieving-stationarity-with-time-series-data-abd59fd8d5a0 [Accessed 11 Feb. 2024].

National Institute of Standards and Technology (2003). *Engineering Statistics Handbook*. [online] itl.nist.gov. Available at: https://itl.nist.gov/div898/handbook/eda/section3/eda35a.htm [Accessed 11 Feb. 2024].

Nau, R. (2014). *Stationarity and differencing of time series data*. [online] Duke.edu. Available at: https://people.duke.edu/~rnau/411diff.htm [Accessed 11 Feb. 2024].

Nayak, S.C. (2017). Development and Performance Evaluation of Adaptive Hybrid Higher Order Neural Networks for Exchange Rate Prediction. *International Journal of Intelligent Systems and Applications*, 9(8), pp.71–85. doi:https://doi.org/10.5815/ijisa.2017.08.08.

Plotly (2017). *Introducing Dash*. [online] Medium. Available at: https://medium.com/plotly/introducing-dash-5ecf7191b503 [Accessed 15 Feb. 2024].

Probst, P., Wright, M.N. and Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, [online] 9(3). doi:https://doi.org/10.1002/widm.1301.

Reddy Beeram, S. and Kuchibhotla, S. (2021). *Communication Software and Networks*. *Lecture Notes in Networks and Systems*. Singapore: Springer Singapore. doi:https://doi.org/10.1007/978-981-15-5397-4.

Rissel, C., Curac, N., Greenaway, M. and Bauman, A. (2012). Physical Activity Associated with Public Transport Use—A Review and Modelling of Potential Benefits. *International Journal of Environmental Research and Public Health*, 9(7), pp.2454–2478. doi:https://doi.org/10.3390/ijerph9072454.

Romanowska, A. and Budzyński, M. (2022). Investigating the Impact of Weather Conditions and Time of Day on Traffic Flow Characteristics. *Weather, Climate, and Society*, 14(3), pp.823–833. doi:https://doi.org/10.1175/wcas-d-22-0012.1.

Rout, A.K. and Dash, P.K. (2016). Forecasting foreign exchange rates using hybrid functional link RBF neural network and Levenberg-Marquardt learning algorithm. *Intelligent Decision Technologies*, 10(3), pp.299–313. doi:https://doi.org/10.3233/idt-160257.

Sah, S., Surendiran, B., Dhanalakshmi, R. and Yamin, M. (2022). Covid‐19 cases prediction using SARIMAX Model by tuning hyperparameter through grid search cross‐validation approach. *Expert Systems*. doi:https://doi.org/10.1111/exsy.13086.

Schonlau, M. and Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, [online] 20(1), pp.3–29. doi:https://doi.org/10.1177/1536867x20909688.

Seltman, H.J. (2018). *Experimental Design and Analysis*. [online] Available at: https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf [Accessed 11 Feb. 2024].

Song, H. and Witt, S.F. (2003). Tourism Forecasting: The General-to-Specific Approach. *Journal of Travel Research*, 42(1), pp.65–74. doi:https://doi.org/10.1177/0047287503253939.

Su, Y., Weng, K., Lin, C. and Zheng, Z. (2021). An Improved Random Forest Model for the Prediction of Dam Displacement. *IEEE Access*, 9, pp.9142–9153. doi:https://doi.org/10.1109/access.2021.3049578.

Vagropoulos, S.I., Chouliaras, G.I., Kardakos, E.G., Simoglou, C.K. and Bakirtzis, A.G. (2016). Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. *2016 IEEE International Energy Conference (ENERGYCON)*. doi:https://doi.org/10.1109/energycon.2016.7514029.

Verma, Y. (2021). *Complete Guide To SARIMAX in Python for Time Series Modeling*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/ [Accessed 17 Feb. 2024].

Voyant, C., Muselli, M., Paoli, C. and Nivet, M.-L. (2013). Hybrid methodology for hourly global radiation forecasting in Mediterranean area. *Renewable Energy*, 53, pp.1–11. doi:https://doi.org/10.1016/j.renene.2012.10.049.

Wang, Z., Bai, Y., Zhu, R., Wang, Y., Wu, B. and Wang, Y. (2019). Impact Analysis of Extra Traffic Induced by Project Construction during Planned Special Events. *Transportation Research Record*, 2673(7), pp.402–412. doi:https://doi.org/10.1177/0361198119840346.

Xiao, Y., Xiao, J. and Wang, S. (2012). A hybrid model for time series forecasting. *Human Systems Management*, 31(2), pp.133–143. doi:https://doi.org/10.3233/hsm-2012-0763.

Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X. and Peng, X. (2019). A hybrid modelling method for time series forecasting based on a linear regression model and deep learning. *Applied Intelligence*, 49(8), pp.3002–3015. doi:https://doi.org/10.1007/s10489-019-01426-3.

Yaiza Montero-Lamas, Novales, M., Orro, A. and Currie, G. (2023). A New Big Data Approach to Understanding General Traffic Impacts on Bus Passenger Delays. *Journal of Advanced Transportation*, 2023, pp.1–15. doi:https://doi.org/10.1155/2023/4082587.

Zhang, L., Aggarwal, C. and Qi, G.-J. (2017). Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:https://doi.org/10.1145/3097983.3098117.

Zhu, Y., Ye, Y., Zhao, X. and James J.Q. Yu (2023). Enhancing Traffic Prediction with Learnable Filter Module. *arXiv (Cornell University)*. [online] doi:https://doi.org/10.48550/arxiv.2310.16063.