SEISMIC MONITORING FOR TRAFFIC AND ENVIRONMENTAL MANAGEMENT IN GREATER MANCHESTER

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE FACULTY OF SCIENCE AND ENGINEERING

2025

By

Shamili Govindaraj

Department of Computing and Mathematics

Contents

List of Tables

List of Figures

Abstract

Urban road traffic is a major contributor to air pollution and greenhouse gas emissions in cities, with Greater Manchester's transport sector alone accounting for over 38% of CO_2 output (*Greenhouse gas reporting: conversion factors 2019*, 2020). On Oxford Road, one of the busiest routes in Manchester, congestion and emissions have a direct impact on both public health and environmental sustainability. Traditional monitoring methods such as roadside cameras and sparse air-quality stations provide valuable insights but are limited by cost, coverage gaps, and privacy concerns. Recent studies have demonstrated that seismic signals can act as a proxy for anthropogenic activity, suggesting a scalable and cost-effective alternative for traffic monitoring (Lecocq *et al.*, 2020) .

This project investigates the feasibility of seismic data as an indicator of road traffic activity by integrating heterogeneous datasets. A Raspberry Shake seismometer was combined with Drakewell camera counts, DEFRA air-quality readings, and Open-Meteo weather data between December 2024 and March 2025. A baseline Linear Regression model was compared against Random Forest regression models, trained on seismic features to estimate quarter-hourly traffic counts. The Random Forest approach, particularly when incorporating air-quality and weather covariates, demonstrated stronger predictive capability than the baseline. Evaluation using standard metrics such as R^2, MAE, RMSE, and MSE confirmed the robustness of the model, including analyses with stratified vehicle categories.

This report covers the preprocessing of datasets, the construction and evaluation of predictive models, and the analysis of traffic–pollution relationships. The findings demonstrate that seismic data provide a viable signal for estimating traffic intensity, with improved reliability when combined with air-quality and weather data. Limitations include the winter-only dataset and the need for seasonal and multi-site validation, which future work should address.

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures and has received ethical approval number: 81222.

Signed: Shamili Govindaraj

Date:

Acknowledgements

I would like to express my deepest gratitude to Dr. Rochelle Taylor, my project supervisor, for her invaluable guidance and continuous support throughout this dissertation. Her encouragement in shaping the timeline, provision of resources, and willingness to share her own research activities were instrumental in helping me refine my methodology and literature review. Her insightful feedback, constructive suggestions, and patient mentoring have not only strengthened this project but also greatly enhanced my academic development.

Abbreviations

AI	Artificial Intelligence
ANPR	Automatic Number Plate Recognition
API	Application Programming Interface
ARIMA	AutoRegressive Integrated Moving Average
AURN	Automatic Urban and Rural Network (DEFRA)
CCTV	Closed-Circuit Television
CNN(s)	Convolutional Neural Network(s)
CO_2	Carbon dioxide
COVID-19	Coronavirus Disease 2019
DEFRA	Department for Environment, Food & Rural Affairs
EDA	Exploratory Data Analysis
ffill	Forward fill (method shorthand)
GNN(s)	Graph Neural Network(s)
HGV(s)	Heavy Goods Vehicle(s)
Hz	Hertz
ID	Identifier (e.g., site ID)
IoT / IOT	Internet of Things
LSTM	Long Short-Term Memory (network)
LOCF	Last Observation Carried Forward
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NO	Nitric Oxide
NO_2	Nitrogen Dioxide
NOx	Nitrogen Oxides
PCA	Principal Components Analysis
PC1 / PC2	First/Second Principal Component
PII	Personally Identifiable Information
PM	Particulate Matter
R^2	Coefficient of Determination (R-squared)
RF	Random Forest

RNN(s)	Recurrent Neural Network(s)
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
RShake	Raspberry Shake (sensor shorthand)
SARIMA	Seasonal ARIMA
SHAP	SHapley Additive exPlanations
SVR	Support Vector Regression
TfGM	Transport for Greater Manchester
TPE	Tree-structured Parzen Estimator
UK	United Kingdom
µg/m	Micrograms per cubic metre

Chapter 1- Introduction

1.1 Project Overview

Air pollution and greenhouse gas emissions from road transport remain among the most pressing environmental and health challenges worldwide. Globally, traffic is a major source of NO_2, PM, and CO_2, which contribute to respiratory illness, cardiovascular disease, and climate change. In the UK, the transport sector is accountable for nearly 38% of Greater Manchester's CO_2 output (Greenhouse gas reporting: conversion factors 2019, 2020). During the COVID-19 lockdowns, reduced vehicle usage was found to significantly improve air quality in cities across the world (Adam, Tran and Balasubramanian, 2021; Barua and Nath, 2021) further reinforcing the strong link between traffic density and pollutant concentrations. Oxford Road, one of the busiest routes in Manchester, illustrates this problem acutely, as heavy traffic density and congestion lead to persistent air-quality concerns. Monitoring traffic and its impact on air quality is therefore a critical challenge for urban planning, climate action, and policy development.

Existing monitoring methods, such as roadside cameras provide valuable data but are constrained by privacy concerns. Recent advances in urban seismology have demonstrated that traffic activity can be captured by ground vibrations measured with seismometers. The global seismic quieting observed during COVID-19 lockdowns revealed that traffic reductions could be detected seismically (Lecocq *et al.*, 2020).

Oxford Road hosts a Raspberry Shake seismometer (Healy, 2023) that continuously records ground vibrations, providing an opportunity to evaluate this approach in a real-world, high-traffic urban corridor. This project builds on that potential by incorporating multiple heterogeneous datasets: seismic vibrations from a Raspberry Shake seismometer, traffic counts from Transport for Greater Manchester's Drakewell cameras and Telraam sensors, pollutant concentrations from the DEFRA Oxford Road AURN station, and weather data from the Open-Meteo API.

Aggregating of all these sources, the project develops a data-driven framework for understanding the correlation between traffic intensity, seismic signals, air quality, and

weather conditions. The project specifically focuses on developing a machine learning model to estimate traffic volumes from bandpass-filtered seismic features, with predictions compared against ground-truth observations from Drakewell camera data. A secondary analysis investigates the sensitivity of seismic-derived estimates to different vehicle classes by stratifying traffic into light and heavy categories. In addition, the project explores correlations between traffic activity, pollutant concentrations, and meteorological variables to evaluate how environmental conditions influence both traffic emissions and seismic monitoring accuracy. The findings have direct implications for smart-city initiatives, low-emission zone strategies, and sustainable urban planning.

1.2 Problem Statement

Road transport is a major contributor to UK's urban air pollution and greenhouse gas emissions, which directly impact public health and global heating. In cities such as Manchester, traffic is responsible for high levels of nitrogen oxides, particulate matter, and carbon dioxide, with Oxford Road identified as a corridor of persistent congestion and poor air quality. The problems do not only pose threat to environmental sustainability but also to the rise of respiratory and cardiovascular disease in the urban population.

Monitoring traffic intensity and its influence on air quality is critical. However, existing monitoring methods, such as roadside cameras and fixed air-quality stations, are limited by high installation and maintenance costs, restricted spatial coverage, and concerns regarding data privacy (Ahmad and Tsuji, 2021). While these systems provide valuable insights, their constraints reduce their suitability for scalable, fine-grained, and real-time monitoring in complex urban environments.

Recent studies on urban seismology has highlighted the potential of using ground vibrations to detect anthropogenic activity such as traffic (Lecocq *et al.*, 2020; Ahmad and Tsuji, 2021). However, the application of seismic monitoring to estimate vehicle counts and link them with air-quality fluctuations remains underexplored, particularly in the UK context. This project addresses this gap by developing a machine learning

framework that integrates seismic, traffic, pollution, and weather data to evaluate the feasibility of seismic monitoring.

1.3 Aim and Objectives

The aim of this project is to quantitatively compare 15-minute seismic features with Drakewell traffic counts on Oxford Road and to investigate the relationship between nowcasted traffic intensity and short-term air-quality variation.

The following Objectives are set out to achieve the goal:

- Develop a feature extraction pipeline from Raspberry Shake seismic data at 15-minute resolution.
- Build and evaluate baseline Linear Regression and tuned Random Forest models for traffic nowcasting using time-ordered train/validation/test splits.
- Assess model performance using standard metrics against Drakewell camera ground truth.
- Analyse the added value of air-quality and weather covariates through incremental model sets.
- Interpret model behaviour using feature importance and SHAP, and discuss limitations, privacy, and ethical considerations.

1.4 Tools and Techniques Used in this Study

This project employed a combination of open-source software, programming libraries, and analytical platforms to ensure reproducibility, scalability, and efficiency. The primary environment was Python (Jupyter Notebook)**,** which provided an interactive workflow for data preprocessing, modelling, and evaluation. Core machine learning tasks were implemented using the scikit-learn library, particularly the RandomForestRegressor for developing the nowcasting models.

For hyperparameter optimisation, the study adopted Optuna (Akiba et al., 2019), a state-of-the-art framework that leverages the Tree-structured Parzen Estimator (TPE) to

efficiently explore parameter spaces. Data analysis and visualisation were supported by pandas**,** NumPy for numerical computation, and Matplotlib/Seaborn for generating statistical plots. For Interpretability SHAP was employed to quantify feature contributions, while PCA was used for dimensionality reduction and exploratory insight into high-dimensional seismic and environmental data.

All packages were installed and managed through pip, ensuring reproducibility. Together, these tools provided a coherent, accessible framework for integrating heterogeneous datasets and developing reliable machine learning models for traffic–environment nowcasting.

1.5 Report Structure

This report is organised into various chapters.

Chapter 1 introduces the project background, problem statement, aim, and objectives.

Chapter 2 presents a literature survey on traffic monitoring, air pollution, seismic sensing, and machine learning approaches, identifying the research gap.

Chapter 3 describes the datasets used and outlines the EDI steps.

Chapter 4 explains the methodology, focusing on Random Forest regression and a comparative baseline with Linear Regressor model.

Chapter 5 reports the results and evaluates model performance

Chapter 6 reviews critically and proposes recommendations for future work.

Chapter 7 concludes the study, reflecting on the aim and objectives.

Chapter 2 - Literature Survey

2.1. Introduction

Before proceeding to the design and implementation of the project, it is important to establish a strong academic foundation through a review of relevant literature. The purpose of this chapter is to analyze and critically evaluate earlier research on traffic monitoring, air pollution, seismic sensing, and the application of machine learning techniques to environmental data.

The literature review is structured into thematic sections. It first explores the established links between urban traffic, air pollution, and public health, followed by an assessment of conventional monitoring approaches and their limitations. The discussion then shifts towards the emerging field of urban seismology, where seismic signals are increasingly recognised as proxies for anthropogenic activity. Subsequently, machine learning methods for traffic and pollution prediction are examined, with a particular focus on Random Forest (Breiman, 2001) models. Finally, the chapter combines these strands to identify the research gap addressed by this project: the lack of an integrated, cost-effective, and privacy-preserving framework that combines seismic, traffic, air quality, and weather data.

2.2. Smart Urban Systems and Data-Driven Monitoring

The smart cities concept emerged from the integration of digital technologies into urban infrastructure, which improves efficiency, sustainability, and quality of life. Early foundational work in the field emphasized the role of the IOT in enabling interconnected networks of devices for real-time monitoring and control (Al-Fuqaha *et al.*, 2015). Subsequently (Masek *et al.*, 2016) demonstrated how IoT-driven environments could redefine urban traffic modelling, providing an early vision for intelligent transportation management in smart cities. These foundation studies set the stage for modern urban observatories, where vast sensor networks capture mobility, pollution, and environmental conditions (Ahmad and Tsuji, 2021). More recently, (Ouallane *et al.*, 2022) extended this line of inquiry by critically evaluating the

intersection between IoT and AI and how adaptable data-driven solutions can revolutionize traffic management.

Importantly, (Ouallane et al., 2022) highlights that while technical feasibility has improved, challenges remain in scaling systems across entire cities, particularly with issues of cost, privacy, and sensor coverage. Critically, much of the smart-city discourse has focused on optimising traffic flow rather than addressing the wider environmental and health consequences of traffic emissions. Although the IoT-enabled systems have enhanced efficiency in managing traffic, they tend to make pollution just another effect rather than a primary monitoring objective. This leaves a research gap that the technology-based traffic solutions have not yet been wholly integrated into frameworks of environmental monitoring thereby neglecting air quality impact. Consequently, the next section shifts focus from the technological foundations of smart urban monitoring towards the direct impacts of traffic on air pollution and public health, which form the broader societal motivation for this project.

2.3. Urban Traffic, Air Pollution, and Public Health

Urban traffic is widely recognised as a main source of air pollutants such as NO_2, PM, and CO_2, which contribute significantly to both environmental degradation and human health risks. Studies such as (Adam, Tran and Balasubramanian, 2021; Barua and Nath, 2021) shows that traffic emissions are strongly correlated with respiratory and cardiovascular illnesses, making traffic intensity a major determinant of urban air quality and public health outcomes. Societal costs of congestion thus go beyond delays in mobility to chronic disease burden and climate conditions.

Traditional monitoring methods, such as air-quality stations and vehicle-count cameras, have provided valuable evidence of this relationship. Nevertheless, cost, low density, and privacy issues restricted the potential of these monitoring systems leading to a lack of information on the spatial and temporal fine-grained picture regarding air pollution in urban areas (Lecocq *et al.*, 2020). The COVID-19 lockdowns offered an unprecedented natural experiment to study this link. Global evidence showed that traffic pollution has decreased substantially during mobility restrictions (Barua and Nath, 2021) while in Manchester, Oxford Road specifically experienced measurable

improvements in air quality alongside reduced traffic counts (Healy, 2023). These findings critically reinforce the causal connection between traffic intensity and urban air quality, while also highlighting the potential for data-driven approaches to assess and mitigate risks.

Despite this evidence, current approaches often treat traffic and air pollution monitoring as separate domains, rather than part of an integrated urban health framework. This gap highlights the need for innovative, scalable monitoring solutions that connect traffic dynamics directly with pollution and health impacts. The following section explores how seismic sensing can contribute to filling this gap by providing a continuous, low-cost, and user-data privacy preserving imitation of traffic activity.

2.4. Traditional Traffic Monitoring Methods

Traffic monitoring has relied on conventional sensing technologies such as inductive loops, roadside cameras, and pneumatic counters. These systems remain central to transport planning, but their limitations have increasingly been recognised in both practice and research. Inductive loop detectors embedded in road surfaces, are among the earliest and most widely adopted technologies. They provide reliable vehicle counts and speed measurements, but their installation and maintenance costs are high, and performance degrades with road wear (Williams and Hoel, 2003).

Roadside cameras and computer vision techniques have become popular for real-time traffic analysis, as they can detect vehicles, classify modes, and even capture behavioural features (Buch, Velastin and Orwell, 2011). However, their dependence on clear visibility makes them vulnerable to weather and lighting conditions. Privacy concerns also present major barriers to widespread deployment, especially in dense urban areas where cameras capture identifiable data beyond traffic flow (Ahmad and Tsuji, 2021).

In recent years, Telraam have emerged as alternatives. Telraam devices use household-mounted cameras to monitor local traffic volumes and stratify vehicles by type. These sensors are affordable and extend coverage into residential areas often ignored by official monitoring (Healy, 2023). However, their accuracy depends on volunteer

participation and proper calibration, raising questions about long-term reliability and consistency.

While traditional monitoring methods have produced useful datasets, they face persistent limitations. High installation and maintenance costs restrict coverage to a few sites, creating gaps in spatial representation (Ahmad and Tsuji, 2021). Data is often aggregated, limiting temporal detail and missing dynamic changes. Moreover, cameras raise privacy concerns that hinder wider acceptance and deployment in urban areas (Ahmad and Tsuji, 2021).

Critically, these gaps mean that above monitoring methods cannot always capture the fine-grained relationship between traffic, pollution, and urban health risks. These shortcomings justify exploring alternative approaches such as seismic sensing, which offer the potential for continuous, low-cost, and private monitoring across wider zones.

2.5. Seismic Monitoring for Urban Traffic

Seismic monitoring, traditionally developed for earthquake detection and subsurface imaging, has more recently been adapted to capture anthropogenic activity such as road traffic, railways, and industrial vibrations. Vehicles generate ground motions in low-frequency bands that can be detected by seismometers, enabling traffic estimation without visual or intrusive sensing technologies (Díaz *et al.*, 2017). The COVID-19 lockdowns provided strong empirical validation of this approach. Global seismic networks observed substantial reductions in high-frequency anthropogenic noise during periods of reduced mobility, directly correlating with traffic decreases (Lecocq *et al.*, 2020). This underscored the sensitivity of seismic data to changes in transport activity and its potential use as a traffic proxy. While privacy-preserving, seismic data quality is highly site-dependent and requires careful validation.

More recent work has expanded these findings with applied monitoring frameworks. (Healy, 2023) focusing on Oxford Road in Manchester, showed that Raspberry Shake seismometers can detect traffic intensity using bandpass filtering, providing a direct local precedent for this project. The advantages of seismic monitoring are notable as it offers continuous coverage, low-cost deployment, and inherent privacy protection,

since no images or personal identifiers are collected. However, limitations remain seismic data can be contaminated by construction, pedestrian movement, or industrial noise, and requires careful feature engineering and modelling to achieve reliable traffic estimates.

However, (Díaz *et al.*, 2017; Healy, 2023) have validated seismic noise as a proxy for traffic, relatively few have sought to integrate seismic, traffic, pollution, and weather data into a unified machine learning framework. Addressing this gap, the present project employs Random Forest regression to evaluate the relationship between seismic signals and traffic counts on Oxford Road, alongside secondary analyses correlating seismic-derived volumes with air quality fluctuations. This builds directly on the urban seismology evidence base while extending it into practical, data-driven urban monitoring for sustainable city planning.

2.6. Processing Anthropogenic Seismic Signals

Urban seismic records generated by human activity require systematic processing before they can be used for traffic and environmental monitoring.

2.6.1. Characterising Seismic Signal:

Understanding the nature of anthropogenic seismic signals is the first step in their analysis. These signals typically occupy frequency bands distinct from natural seismicity, reflecting repetitive patterns of human mobility such as vehicle movements and train operations (bse *et al.*, 2015; Díaz *et al.*, 2017). For example, (Boese *et al.*, 2015) used borehole seismometers in Auckland to reveal recurring vibration signatures tied to urban transport. (Díaz *et al.*, 2017) further demonstrated that urban seismicity could be distinguished through its unique frequency characteristics, offering valuable baselines for traffic-related monitoring. However, overlapping sources in dense cities complicate the clarity of these descriptions.

2.6.2 Anthropogenic Noise:

The seismic background in cities is dominated by vibrations from daily human activity, often termed anthropogenic noise. While traditionally regarded as a disturbance, recent

work has reframed this noise as a proxy for urban activity. (Lecocq *et al.*, 2020) showed how global lockdowns during COVID-19 led to significant reductions in high-frequency noise, demonstrating the strong link between anthropogenic activity and seismic energy. This finding suggest traffic-related seismic noise holds value for mobility studies, though its variability makes consistent interpretation challenging.

2.6.3. Noise Reduction Technique:

Filtering techniques are employed to isolate relevant anthropogenic signals from background seismic noise. Approaches such as spectral filtering, cross-correlation, and machine learning based denoising have been explored. (Birnie *et al.*, 2021) proposed self-supervised neural networks to suppress random seismic noise, significantly improving signal clarity without requiring large, labelled datasets. Similarly (Dou *et al.*, 2017) used distributed acoustic sensing to enhance anthropogenic traffic signatures, demonstrating the potential of innovative filtering approaches. Nonetheless, filtering requires careful calibration, as excessive denoising may distort key features of traffic-related signals.

2.6.4. Wave Attenuation in Urban Environments

Seismic wave attenuation plays a critical role in determining the sensitivity of traffic monitoring systems, as signals weaken while propagating through heterogeneous urban substrates. (Fuchs *et al.*, 2018) demonstrated how high-frequency seismic waves rapidly decay with distance in rock-like media, significantly constraining detectability and the spatial footprint of anthropogenic signals. This underlines the importance of accounting for geology, sensor placement, and distance from traffic corridors when designing seismic monitoring frameworks (Fuchs *et al.*, 2018). Neglecting attenuation effects may lead to underestimation of vehicle intensity or misinterpretation of traffic patterns, thereby introducing systematic bias into urban monitoring studies.

2.7. Integration of Heterogeneous Data Sources

Effective urban traffic and air quality monitoring increasingly relies on the integration of diverse data streams to overcome the limitations of single-source approaches. Traditional roadside monitoring stations provide highly reliable information but are

often restricted in spatial coverage and entail high installation and maintenance costs (Buch, Velastin and Orwell, 2011). More recently, citizen-science initiatives such as Telraam (*Telraam - Smart traffic counters for all transport modes*, 2025) have enabled distributed, low-cost collection of traffic data through community-installed sensors, offering broader coverage but with challenges in data consistency and calibration (Ouallane et al., 2022).

In Manchester, Transport for Greater Manchester's Drakewell system provides high-resolution traffic counts from roadside cameras, serving as a trusted ground-truth dataset for model validation. Drakewell data capture vehicle intensity along major corridors such as Oxford Road, making them indispensable for benchmarking other sources. In parallel, Telraam complements this official data by offering a more participatory monitoring approach that enhances spatial diversity and engages the public in traffic observation.

Beyond traffic counts, air quality monitoring stations (info@airqualityengland.co.uk, 2025) provide high-resolution pollutant concentration measurements, which can be directly linked to traffic intensity. Coupling traffic and pollution datasets enables the exploration of how changes in mobility patterns translate into measurable air-quality outcomes (Barua and Nath, 2021). Similarly, meteorological data from platforms such as the Open-Meteo API (*Open-Meteo.com*, 2025) capture weather influences, which play a crucial role in dispersing or concentrating pollutants (Williams and Hoel, 2003).

The integration of traffic (Drakewell and Telraam), environmental, and meteorological datasets provides a multidimensional view of urban mobility and its environmental consequences. Critically, such integration supports nowcasting applications, where real-time estimation of traffic activity is needed to inform urban planning and public health interventions.

2.8. Machine Learning in Traffic and Environmental Prediction

Machine learning has become central to modern traffic monitoring because it can detect complex patterns in large and dynamic datasets that traditional models often miss. Machine learning algorithms are well suited to capture these relationships by learning

directly from historical and real-time data, without requiring strict assumptions about linearity or stationarity. This project builds on these developments by applying machine learning to integrate seismic and traffic datasets, aiming to evaluate its effectiveness in estimating road traffic intensity.

2.8.1. Time-Series and Statistical Models

Early traffic forecasting relied heavily on time-series approaches such as ARIMA and SARIMA. These approaches have been widely used because they can effectively capture trend and seasonality in traffic flows and air quality, while remaining computationally efficient and interpretable. (Williams and Hoel, 2003) applied SARIMA to urban traffic forecasting and demonstrated that these models performed well under stable and recurring traffic conditions, making them reliable baseline predictors.

However, the limitations of such statistical approaches have become increasingly evident. They rely on assumptions of linearity and stationarity, which restrict their ability to handle the nonlinear interactions between traffic, weather, and pollution. More recent evaluations, such (Yu, Markos and Zhang, 2022), show that while ARIMA and SARIMA remains useful for benchmarking, it is consistently outperformed by modern machine learning and deep learning methods in capturing complex, multimodal temporal dependencies. Thus, time-series models provide valuable historical context and benchmarks but are inadequate for integrated, real-world urban monitoring which justifies the move towards machine learning frameworks in this project.

2.8.2. Regression Based Models

Regression models have long been central to traffic forecasting and air pollution analysis, offering a statistically grounded method for identifying relationships between explanatory variables and outcomes such as vehicle counts. For example, Linear regression has been widely applied in urban studies to estimate pollutant concentrations from traffic density, owing to its interpretability and relatively low computational cost (Williams and Hoel, 2003). More recent studies have demonstrated its continued relevance; for instance, (Kumar and Vanajakshi, 2015) successfully applied Seasonal

ARIMA and regression-based models to short-term traffic prediction, illustrating their value when data availability is limited.

Despite these strengths, regression-based approaches struggle with the non-linear and dynamic nature of traffic systems. (Lv *et al.*, 2015) highlighted how regression often underperforms when compared with machine learning in highly variable environments. These constraints make regression valuable as a transparent benchmark, providing interpretability and a baseline for comparison, but insufficient for high-dimensional, multi-source data such as seismic, pollution, and weather signals.

In this project, Linear Regression is therefore employed as a baseline model, serving as a benchmark against which the performance of more advanced methods can be evaluated. While not expected to capture the full complexity of multimodal datasets, it provides an essential reference point for assessing the benefits of more flexible machine learning approaches.

2.8.3. Machine Learning Models

Machine learning approaches have gained prominence in traffic and air quality prediction because they offer the flexibility to model complex, non-linear interactions that regression methods cannot fully capture. Support Vector Regression (SVR) has been widely applied in transport forecasting, with studies demonstrating its ability to incorporate noise reduction and clustering strategies for more reliable predictions (Tang *et al.*, 2019) . These methods illustrate how ML can outperform linear models in handling variability, though they often require careful parameter tuning and can be computationally intensive.

Ensemble-based methods such as Random Forests provide a more robust alternative by averaging across multiple decision trees, thereby reducing overfitting and enhancing generalisability. (Breiman, 2001) seminal work established Random Forest as a versatile model for complex datasets, balancing predictive power with interpretability. Recent applications in traffic forecasting have highlighted its strengths in capturing multi-variable interactions, especially when integrating traffic counts, meteorology, and environmental indicators (Essien *et al.*, 2019). Similarly, (García-Sigüenza *et al.*,

2023) emphasised the role of explainability in ML-based traffic forecasting, noting that ensemble methods such as Random Forest are particularly valuable because they provide interpretable feature importance while maintaining high predictive accuracy.

Within this project, Random Forest regression is adopted as the primary modelling framework because it aligns well with the heterogeneous nature of the datasets. This project focuses on nowcasting rather than long-term forecasting, aiming to monitor current traffic intensity by fusing seismic, traffic, air quality, and weather data. For such applications, Random Forest regression (Breiman, 2001) is particularly suitable, as it provides robust and interpretable estimates without requiring the large temporal datasets typically needed for forecasting models.

2.8.4. Deep Learning Models

Deep learning has gained prominence in traffic forecasting and environmental prediction because of its ability to model complex non-linear relationships and capture long-term temporal dependencies. Deep learning models such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) have been widely explored for traffic analysis. CNNs are effective at learning spatial features in traffic images and grid-based sensor data, making them useful for short-term congestion detection (Chen, Yu and Liu, 2018).GNNs extend this by capturing complex spatial–temporal dependencies in road networks, treating junctions and roads as nodes and edges to improve prediction accuracy (Yu, Markos and Zhang, 2022).

While Recurrent Neural Networks (RNNs) and their advanced variant, Long Short-Term Memory (LSTM) networks, are particularly suited for sequential data such as traffic counts or pollutant concentrations, as they retain information across time steps and identify temporal patterns (Essien *et al.*, 2019). LSTM has been successfully applied in urban traffic prediction tasks, often outperforming traditional statistical models when sufficient historical data is available (Essien et al., 2019). Despite these advantages, deep learning models face challenges when applied to heterogeneous and noisy datasets such as those integrating seismic, traffic, and environmental data. They are computationally intensive, require large amounts of training data, and often behave as black boxes with limited interpretability (García-Sigüenza *et al.*, 2023).

In this project, the available dataset is insufficient as deep learning architectures rely on long histories and seasonal variations. Moreover, the primary aim of this project is nowcasting rather than long-horizon forecasting, where the flexibility and robustness of ensemble methods such as Random Forest provide a better trade-off between predictive accuracy, interpretability, and data availability.

2.9. Integrated Multimodal Approaches

While individual machine learning and statistical models have demonstrated promise in predicting traffic flow or air pollution independently, urban environments are inherently shaped by multiple interacting factors. For example, traffic intensity influences pollutant concentrations, while meteorological variables such as wind speed, humidity, and temperature determine how those pollutants disperse or accumulate (Williams and Hoel, 2003). Relying on a single data stream often fails to capture these dependencies, limiting the robustness of forecasts and real-time monitoring systems.

Recent studies have therefore adopted integrated multimodal approaches, combining traffic, meteorological, and environmental data to improve predictive accuracy and interpretability.(Barua and Nath, 2021) highlighted that COVID-19 mobility restrictions caused dramatic reductions in air pollution globally, reinforcing the need to account for interactions between traffic demand and environmental outcomes. Similarly,(Ouallane *et al.*, 2022) argued that IoT-enabled systems integrating traffic sensors and environmental monitors provide more scalable and adaptive solutions for smart-city applications. In practice, these multimodal frameworks improve nowcasting, enabling real-time assessment of traffic-related emissions and their public health impacts, rather than focusing solely on long-term forecasting.

From a methodological standpoint, integrated datasets allow for the application of ensemble and hybrid models, such as Random Forests or boosting methods, which are well-suited to handle heterogeneous predictors and uncover non-linear interactions (Breiman, 2001). Compared to univariate models such as ARIMA, multimodal frameworks provide richer representations of urban processes and demonstrate better resilience under varying conditions, including weather anomalies and unexpected mobility disruptions.

For this project, the integration of Drakewell and Telraam traffic counts, DEFRA air-quality monitoring, and Open-Meteo weather data reflects this multimodal perspective. The approach ensures that traffic-derived seismic signals are not interpreted in isolation but are instead contextualised against real-world environmental and meteorological conditions. This enhances the credibility of the nowcasting framework and directly supports smart-city initiatives by offering cost-effective, privacy-preserving, and holistic monitoring solutions.

2.10. Research Gap and summary

The literature shows that while traditional monitoring methods such as cameras and inductive loops provide useful insights into traffic and pollution, they remain limited by cost, spatial coverage, and privacy (Buch, Velastin and Orwell, 2011).Recent advances in urban seismology suggest that seismic data can capture anthropogenic activity, including road traffic (Díaz *et al.*, 2017; Lecocq *et al.*, 2020).

Machine learning models, particularly regression and ensemble methods, have proven effective for traffic and air-quality prediction (Breiman, 2001; Kumar and Vanajakshi, 2015) yet limited research has integrated seismic, traffic, air quality, and weather data into a unified framework. Furthermore, most existing work emphasises forecasting, while the potential of nowcasting for real-time monitoring and policy support remains overlooked. This project addresses these gaps by developing a Random Forest–based nowcasting framework that integrates multiple environment data. The contribution lies in demonstrating how seismic data can be combined with multimodal sources to provide scalable, privacy-preserving solutions for smart-city traffic and air-quality monitoring.

Chapter 3 - Data Sources and Analysis

3.1. Introduction

The development of a reliable nowcasting framework depends heavily on the quality and preparation of the underlying data. For this project, a multi-source approach was adopted by integrating seismic, traffic, air quality, and weather datasets collected along Oxford Road, Manchester. To ensure consistency and reliability, raw datasets were first examined through exploratory data analysis and then subjected to a series of preprocessing steps. These included cleaning missing or inconsistent entries, resampling to a common quarter-hourly resolution, engineering relevant features, and merging datasets into a unified structure. The upcoming sections describe the data sources in detail, dataset structures and followed by the preprocessing methods applied to prepare them for subsequent analysis.

3.2. Data Sources

The Oxford Road, Manchester, is recognised as one of the busiest transport corridors in the city, characterised by heavy traffic density, a large proportion of bus and taxi services, and consistently elevated air pollution levels. This road hosts two of the Manchester's major universities which contribute significantly to daily commuter flows, student mobility, and traffic congestion. These factors make Oxford Road an ideal real-world case study for investigating the interactions between traffic, seismic activity, and environmental conditions.

To ensure the dataset captured both typical and atypical urban dynamics, the analysis period was defined from December 2024 to March 2025. This four-month window was chosen deliberately because it encompasses a diverse range of traffic and air-quality conditions. December represents the holiday season, when traffic patterns are disrupted by shopping and travel peaks, while January through March reflects a return to routine commuting and university activity, with typical weekday congestion and seasonal weather influences. This combination of holiday anomalies and baseline traffic conditions ensures that the trained models are exposed to both peak events and regular

cycles, improving the robustness of the framework. The following section explains the data sources from which each data set is obtained.

3.2.1. Raspberry Shake Data

The primary dataset for this project was obtained from the Raspberry Shake (RShake) seismometer installed along Oxford Road, Manchester through the Raspberry Shake platform shown in below Figure 3.1, (*Data View: Raspberry Shake Data Visualization Tool*, 2025). The instrument records continuous vertical ground vibrations at a sampling rate of 100 Hz, which were subsequently processed into spectral representations suitable for traffic-related signal analysis.

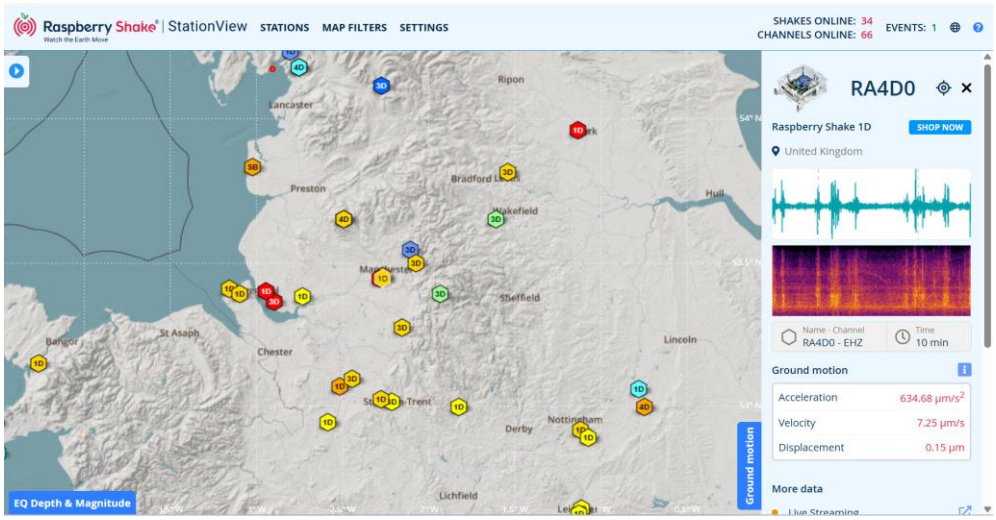


Figure 3.1: Raspberry Shake Data Visualization Tool

The Oxford Road location was chosen as it represents one of Manchester's busiest corridors, heavily influenced by buses, taxis, private vehicles, and pedestrian flows.

3.2.2. Traffic Cameras and Sensors Data

Traffic datasets provide direct measurements of vehicle flows and serve as the critical reference for validating seismic-derived nowcasting models. In this project, two sources of traffic data were employed. The Drakewell traffic dataset was obtained from TfGM's official traffic monitoring network, Figure 3.2. Cameras and inductive loops are strategically positioned along Oxford Road, especially in the city-centre section, to capture high-density traffic flows.

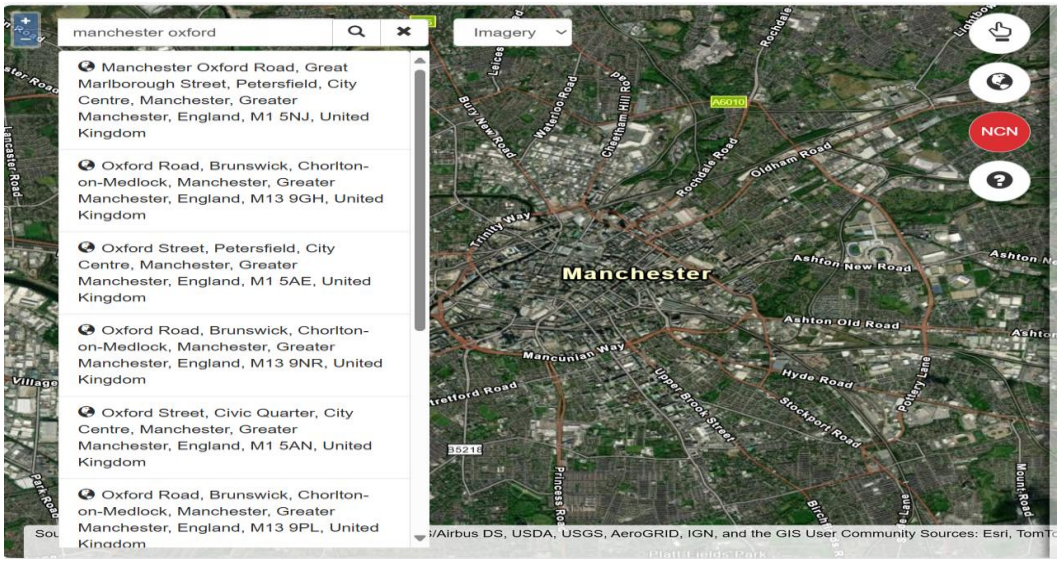


Figure 3.2: TfGM's official Drakewell Site

The second traffic source shown in Figure 3.3 was the Telraam sensor (ID: 9000005312) located on Oxford Road, available via the Telraam platform (Telraam, 2024). Unlike Drakewell's official infrastructure, Telraam sensors are citizen-installed, low-cost traffic monitoring devices that use a small camera to capture street-level mobility.

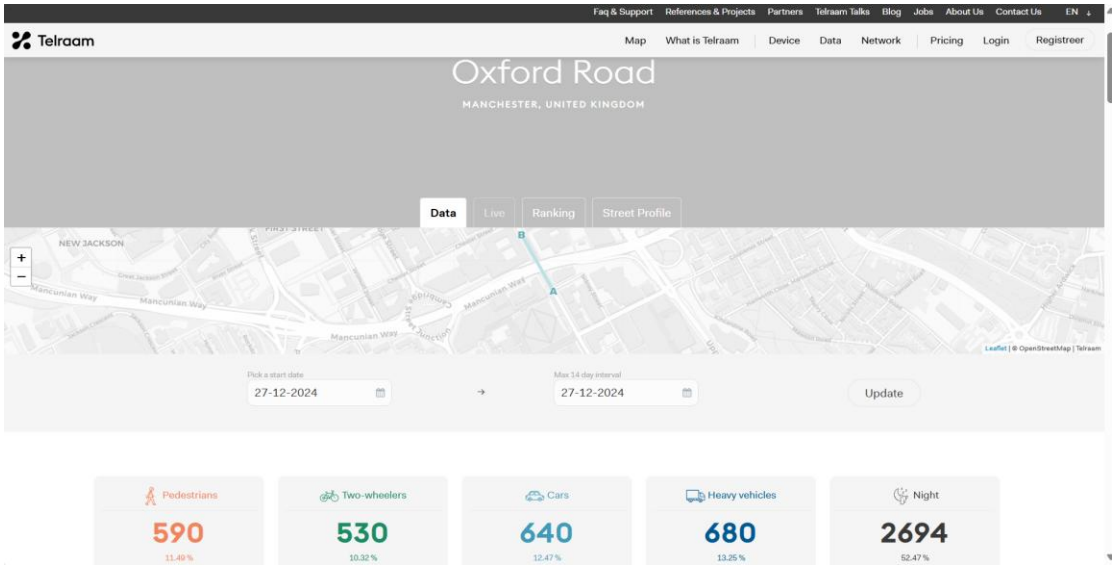


Figure 3.3: Telraam Sensor Site

The location of the Telraam device further along Oxford Road ensures complementary coverage to the Drakewell cameras, particularly in areas where official monitoring is

sparse. While Telraam data are less precise due to reliance on simple hardware and volunteer placement, they provide valuable continuous and fine-grained observations. In this project, Telraam data acted as a secondary validation source, allowing comparison of seismic-based predictions against both official and citizen-driven datasets.

3.2.3. DEFRA AURN Station Data:

Air quality data for this study were obtained from the DEFRA Automatic Urban and Rural Network (AURN) Oxford Road monitoring station (site ID: MAN1), accessed via the Air Quality England portal (Air Quality England, 2025) as shown in Figure 3.4. The Oxford Road AURN site is situated close to the University of Manchester campus, directly adjacent to the study corridor, and provides high-quality reference measurements of pollutant concentrations.

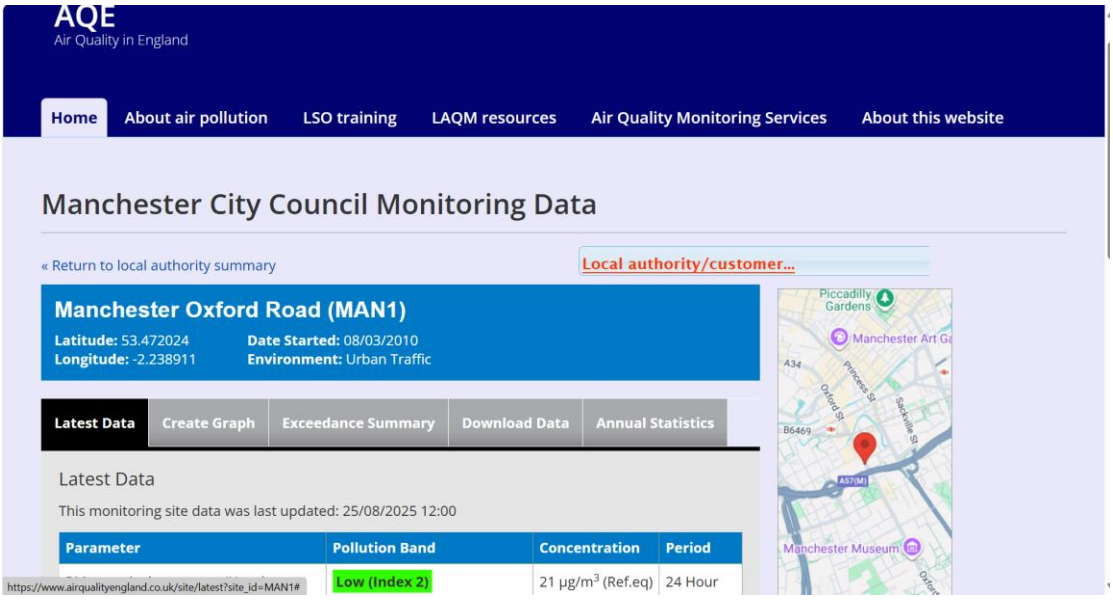


Figure 3.4: Air Quality Data source

3.2.4. Open-Meteo API Data

Weather data were incorporated as a extensive dataset to evaluate whether meteorological conditions influenced traffic dynamics or pollutant dispersion along Oxford Road. These data were obtained from the Open-Meteo Historical Weather API (*Open-Meteo.com*, 2025) see Figure 3.5, which provides free access to hourly weather

observations and reanalysis data for any specified geographic coordinates. To align the dataset with the Oxford Road case study, the latitude and longitude corresponding to the area were used when generating the dataset.

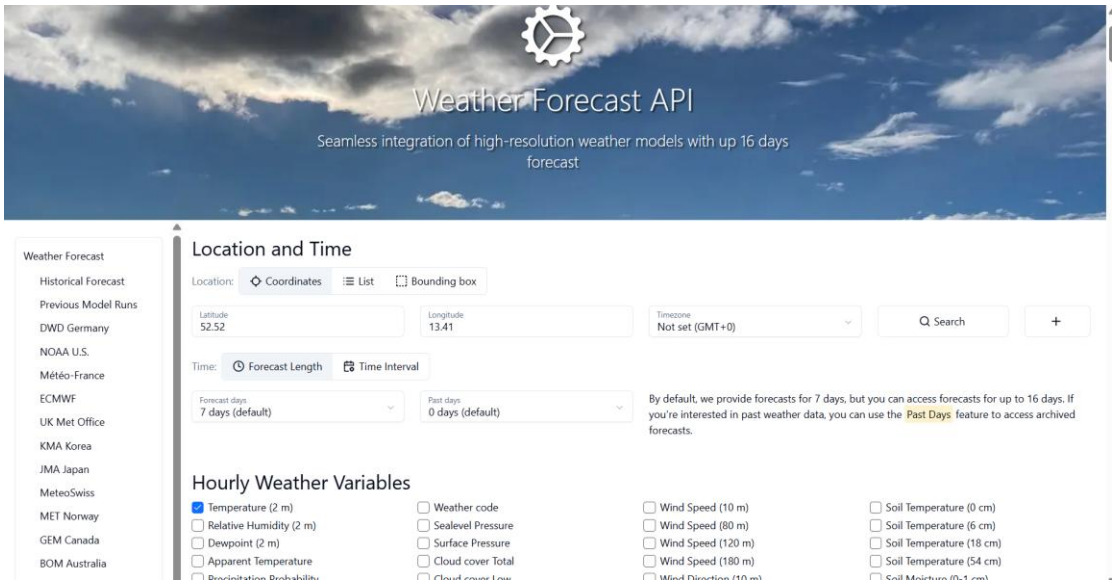


Figure 3.5: Open-Meteo API Data Source

3.3. Ethical Considerations

Ethical issues were considered throughout this project, ensuring that the research adhered to Manchester Metropolitan University's ethical guidelines (Approval Ref: 81222) and broader principles of responsible data science.

3.3.1. Data Privacy

The datasets used in this study were entirely open and aggregated, provided under public licences by Raspberry Shake, Drakewell, Telraam, DEFRA and Open-Meteo. None contained personally identifiable information (PII). Importantly, the use of seismic signals ensured that vehicle movements were inferred from ground vibrations rather than intrusive visual or registration plate data. This makes the framework inherently privacy-preserving, mitigating risks typically associated with CCTV or other vision-based monitoring systems.

3.3.2. Privacy Concerns in Data Collection

Although no direct human participants were involved, potential privacy risks were still considered. Seismic and traffic counts capture urban activity indirectly but remain non-intrusive and compliant with data protection standards. The Telraam dataset, contributed by citizen volunteers, was used cautiously: it was reported for transparency but excluded from final model evaluation due to limited field of view and under-representation of true traffic flows. This careful treatment reflects an ethical stance of avoiding misleading or biased use of volunteered data.

3.3.3. Balancing Data Utilisation with Ethical Responsibilities

The study sought to balance the benefits of multimodal data integration with responsibilities of fairness, transparency, and accountability. First, bias in representation was acknowledged: Telraam's restricted coverage and Oxford Road's bus-gate restrictions meant its outputs were not directly comparable with Drakewell or Raspberry Shake. This limitation was reported clearly rather than ignored, avoiding overclaiming results. Second, transparency and reproducibility were prioritised. All preprocessing steps were fully documented, and open-source tools were used to enable reproducibility. Finally, the project's design promotes social and environmental responsibility: by offering a scalable, privacy-preserving monitoring framework, it contributes to sustainable traffic and air-quality management without eroding public trust.

3.4. Dataset Structure and Features

The project adopts a multi-source approach by integrating heterogeneous datasets. Each dataset was obtained from publicly accessible sources and provides complementary insights into the dynamics of urban traffic and its environmental impacts. Before preprocessing, the raw datasets were examined to understand their structure, resolution, and potential limitations. The following subsections describe each dataset individually, outlining its structure, features, and relevance to the project.

3.4.1. Seismic Data

The RShake data are made available in daily comma-separated value (CSV) files, each representing a single day of observations. Within each file, ground motion is aggregated into 15-minute intervals, providing a fine-grained temporal resolution. The first column corresponds to the timestamp, while the subsequent 200 columns represent spectral amplitude values across frequency bins ranging from 0.0 Hz to 50.0 Hz at intervals of 0.25 Hz with 8639 entries see Table 3.1.

Property	Value
Index Type	DatetimeIndex
Entries	8639
Date Range Start	2024-12-24 00:00:00
Date Range End	2025-03-23 23:45:00
No. of Columns	200
Column Range	0.0 to 49.75
Dtypes	float64

Table 3.1: Summary of the seismic spectral dataset

An example is shown in Appendix A, where each row represents one 15-minute time window and each column corresponds to the energy within a specific frequency band. For the four-month study period, daily files were aggregated into a single data frame, producing a unified dataset with timestamped 15-minute entries suitable for merging with traffic, air quality, and weather datasets.

The seismic dataset forms a major input variable in this project, serving as a privacy-preserving proxy for traffic intensity. Unlike camera-based systems, seismic signals provide collective measures of road activity without recording identifiable vehicle information, making them well-suited for scalable smart city monitoring applications.

3.4.2. Traffic Data

The Drakewell dataset records cars, vans, buses, lorries and total traffic volumes at 15 mins intervals. This dataset serves as the primary ground-truth reference for validating Random Forest regression outputs. Its structure and variables are presented in the below

tables Table 3.2 and Table 3.3, which lists the available categories, temporal resolution, and coverage.

Property	Value
Entries	8,640
Start date	2024-12-24 00:00
End date	2025-03-23 23:45
Number of columns	14
Data types	float64×14

Table 3.2: Summary of Drakewell Dataset

Feature	Non-Null Count	Dtype
Vehicle count - Motorcycles southwest bound	8,633	float64
Vehicle count - Motorcycles northeast bound	8,633	float64
Vehicle count - Cars and Light Vans southwest bound	8,633	float64
Vehicle count - Cars and Light Vans northeast bound	8,633	float64
Vehicle count - Cars with Trailer southwest bound	8,633	float64
Vehicle count - Cars with Trailer northeast bound	8,633	float64
Vehicle count - Rigids, Heavy Vans or Mini Buses southwest bound	8,633	float64
Vehicle count - Rigids, Heavy Vans or Mini Buses northeast bound	8,633	float64
Vehicle count - Articulated HGVs southwest bound	8,633	float64
Vehicle count - Articulated HGVs northeast bound	8,633	float64
Vehicle count - Buses and Coaches southwest bound	8,633	float64
Vehicle count - Buses and Coaches northeast bound	8,633	float64
Vehicle speed - Vehicles Speed southwest bound (km/h)	8,633	float64
Vehicle speed - Vehicles Speed northeast bound (km/h)	8,633	float64

Table 3.3: Features Summary Table (Drakewell)

The Telraam sensor provides detailed 15 min interval counts of light vehicles, heavy vehicles, cyclists, and pedestrians, offering a more multi-modal perspective see Table 3.4.

Property	Value
Entries	16,171
Start date	2023-07-27 12:30
End date	2025-03-24 13:45
Number of columns	49
Data types	float64×32, int64×13, object×3, datetime64[ns]×1
Memory usage	8.44 MB

Table 3.4: Summary of Telraam Dataset

Together, these datasets capture both institutional and community perspectives on mobility within one of Manchester's busiest transport corridors. This Dataset is further filtered inorder to match the case study timeline.

3.4.3 Air Quality Data

The Air Quality dataset spans the same timeline of other datasets and is recorded at an hourly resolution. As shown in Table 3.5, Each record contains timestamped pollutant concentrations, expressed in micrograms per cubic metre (µg/m³). The main pollutants included are PM_{10}, Nitric Oxide (NO), Nitrogen Dioxide (NO_2), and Nitrogen Oxides (NO_x as NO_2). Each pollutant is accompanied by a status/unita column, which specifies measurement conditions and units in accordance with DEFRA's reporting standards.

End Date	End Time	PM10	Status/units	NO	Status/units.1	NO2	Status/units.2	NOXasNO2	Status/units.3
23-12-2024	01:00:00		nan	5.76	R ugm-3	10.19	R ugm-3	19.03	R ugm-3
23-12-2024	02:00:00	31.70	R ugm-3 (Ref.eq)	2.17	R ugm-3	7.20	R ugm-3	10.52	R ugm-3
23-12-2024	03:00:00	29.20	R ugm-3 (Ref.eq)	3.97	R ugm-3	9.55	R ugm-3	15.63	R ugm-3
23-12-2024	04:00:00	30.00	R ugm-3 (Ref.eq)	4.26	R ugm-3	13.88	R ugm-3	20.41	R ugm-3
23-12-2024	05:00:00	25.00	R ugm-3 (Ref.eq)	2.08	R ugm-3	10.88	R ugm-3	14.07	R ugm-3
23-12-2024	06:00:00	24.20	R ugm-3 (Ref.eq)	7.38	R ugm-3	18.04	R ugm-3	29.35	R ugm-3
23-12-2024	07:00:00	29.20	R ugm-3 (Ref.eq)	22.10	R ugm-3	40.69	R ugm-3	74.58	R ugm-3
23-12-2024	08:00:00	45.00	R ugm-3 (Ref.eq)	91.93	R ugm-3	96.87	R ugm-3	237.83	R ugm-3
23-12-2024	09:00:00	45.00	R ugm-3 (Ref.eq)	77.79	R ugm-3	96.69	R ugm-3	215.96	R ugm-3
23-12-2024	10:00:00	46.70	R ugm-3 (Ref.eq)	150.67	R ugm-3	110.84	R ugm-3	341.86	R ugm-3
23-12-2024	11:00:00	30.00	R ugm-3 (Ref.eq)	116.24	R ugm-3	100.15	R ugm-3	278.38	R ugm-3
23-12-2024	12:00:00	34.20	R ugm-3 (Ref.eq)	128.89	R ugm-3	101.32	R ugm-3	298.94	R ugm-3
23-12-2024	13:00:00	29.20	R ugm-3 (Ref.eq)	84.81	R ugm-3	86.83	R ugm-3	216.87	R ugm-3

Table 3.5: Sample Air Quality Dataset

The air quality dataset plays a dual role in this project: first, as an environmental outcome to correlate with seismic-derived traffic intensity, and second, as an explanatory feature in multi-modal models that combine traffic, seismic, and meteorological inputs.

3.4.4 Weather Dataset

The dataset spans the analysis period and is recorded at an hourly resolution, which needs to be down sampled, ensuring temporal compatibility with seismic and traffic data. The variables extracted include temperature (2 m), precipitation, wind speed (10 m), relative humidity (2 m), and cloud cover. These attributes were selected as they are directly relevant to both traffic behaviour and atmospheric conditions.

time	temperature_2m (°C)	precipitation (mm)	wind_speed_10m (km	lative_humidity_2m (	cloud_cover (%)
2024-12-23T00:00	3.4	0.0	14.2	85.0	100.0
2024-12-23T01:00	2.8	0.2	10.2	90.0	100.0
2024-12-23T02:00	2.1	0.9	8.1	94.0	100.0
2024-12-23T03:00	2.8	0.2	11.5	94.0	100.0
2024-12-23T04:00	3.0	0.0	13.4	93.0	91.0
2024-12-23T05:00	2.2	0.0	12.3	94.0	43.0
2024-12-23T06:00	1.7	0.0	13.1	94.0	100.0
2024-12-23T07:00	1.4	0.0	14.3	94.0	100.0
2024-12-23T08:00	1.6	0.0	16.6	92.0	100.0

Table 3.6: Sample Weather Dataset

An example of the raw dataset structure is shown in above Table 3.6, where each row represents an hourly record with corresponding weather attributes. For integration, the dataset was further resampled and cleaned to align timestamps with the seismic, traffic, and air quality datasets.

3.5 Exploratory Data Analysis

EDA was undertaken to examine temporal patterns, variability, and correlations across the datasets. Since the project aims at developing a nowcasting framework, understanding diurnal, weekly, and monthly cycles was essential before modelling. The following subsections highlight key findings, supported by visualisations.

3.5.1 Seismic data for Traffic

Seismic data provided a strong proxy for traffic activity. The monthly average profile Figure 3.6, revealed higher activity in January to March compared to December, reflecting post-holiday increases.

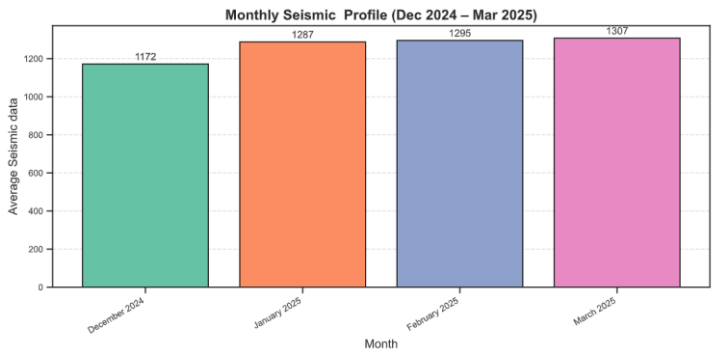


Figure 3.6: Monthly Seismic Profile

Heatmaps of hourly seismic activity Figure 3.7 confirmed alignment with traffic cycles: higher amplitudes during commuting hours and reduced signals overnight.

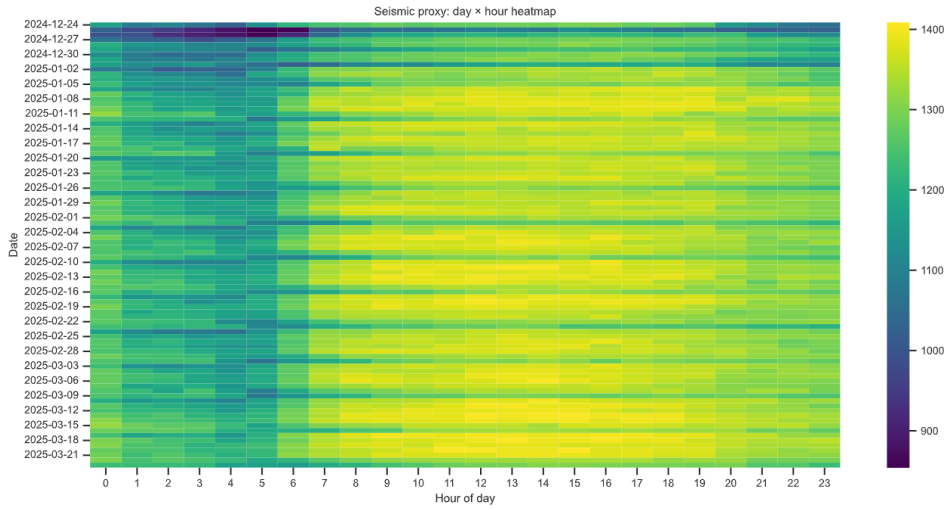


Figure 3.7 Seismic day and Hour Heatmap

Below Figure 3.8 displayed similar trends, with weekdays consistently recording higher seismic energy compared to weekends.

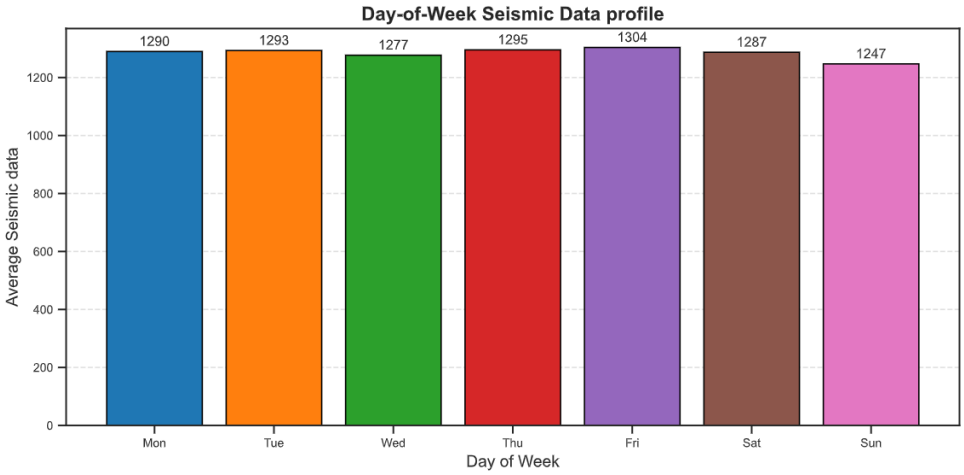


Figure 3.8 Day-of-week averages

These results validated the potential of seismic features as reliable traffic indicators.

3.5.2 Traffic Volume Characteristics

Traffic data from Drakewell roadside camera is examined to capture temporal fluctuations in vehicle activity. Average hourly traffic profiles Figure 3.9 reveals a typical urban dual-peak pattern: a sharp morning surge beginning at 7 AM, peaking around 8-9 AM, and an even stronger evening peak at 4-6 PM.

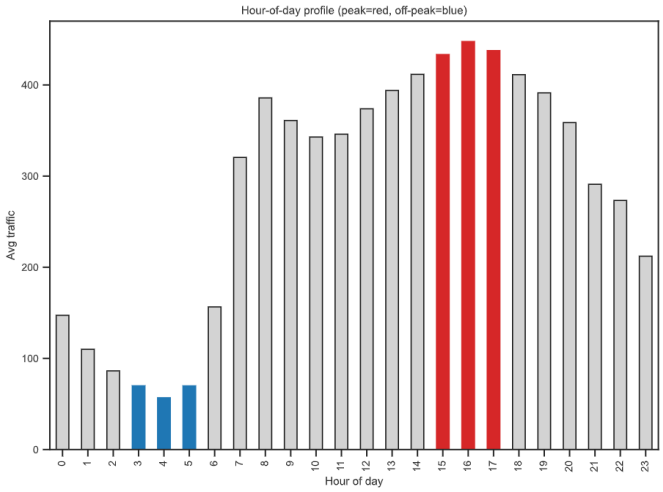


Figure 3.9: Traffic Volume in a day

Breaking down by vehicle category, light vehicles dominate counts across all hours, but heavy vehicles display a flatter trend, remaining consistent throughout the day with modest midday increases.

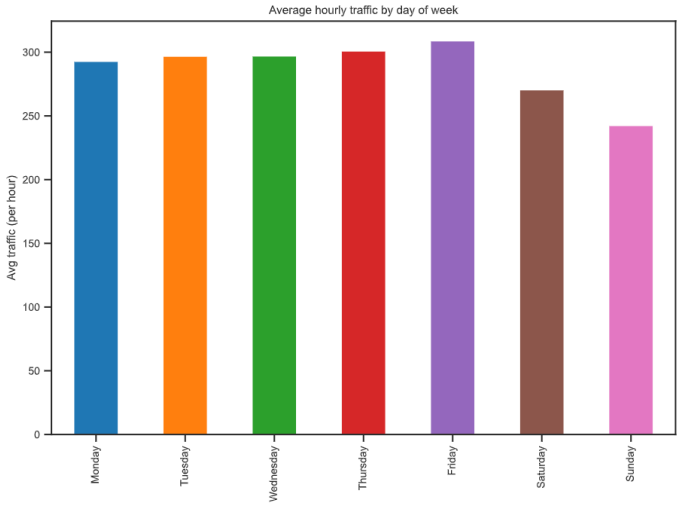


Figure 3.10: Average weekday–weekend contrasts plot

Weekly traffic averages Figure 3.10 show weekday–weekend contrasts. Volumes remain relatively stable Monday to Thursday, rise on Friday, and drop considerably on Sunday, when only essential transport dominates. Saturday traffic remains moderate, illustrating the role of leisure and shopping activities. These insights provide a strong rationale for modelling strategies that stratify by time-of-day and day-of-week.

3.5.3 Air Quality Dynamics

Hourly pollutant averages highlight the strong diurnal behaviour of air quality Figure 3.11. Peak concentrations are observed in the morning between 8–10 AM and again in the evening between 5–7 PM, coinciding with commuter rush hours and elevated vehicle use. The lowest levels typically occur during the early morning hours (2–5 AM), when both traffic activity and combustion emissions are minimal.

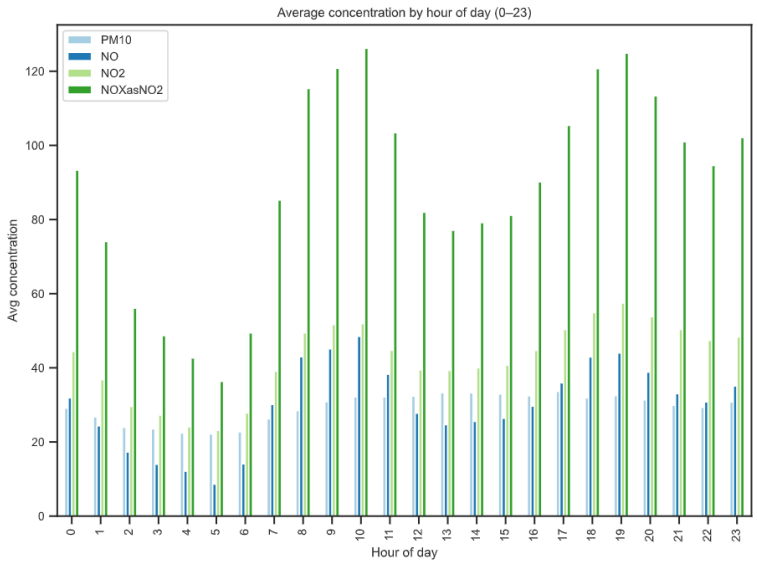


Figure 3.11: Air Pollutant Concentration Pattern

This pattern confirms that traffic is a dominant driver of pollution levels along Oxford Road. Importantly, the variation between pollutants also reveals specific behaviours: NO_2 and NOx show the sharpest increases during rush hours due to diesel-powered buses and heavy vehicles, while PM_{10} variations are less pronounced, reflecting additional contributions from non-traffic sources such as dust resuspension.

Boxplots Figure 3.12 further illustrated pollutant variability, highlighting NO and NOx as exhibiting the largest spread with frequent outliers, which is consistent with sporadic spikes from heavy vehicles or congestion events.

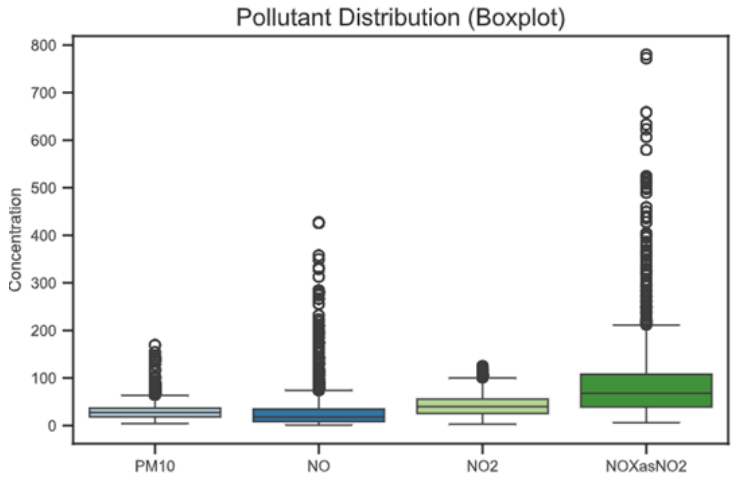


Figure 3.12: Pollution Distribution

This analysis confirmed that air pollutants closely follow traffic intensity, justifying their inclusion as explanatory features.

Overall, the EDA confirms that the multi-source datasets (traffic, seismic, and pollution) capture consistent and complementary information about mobility patterns on Oxford Road. The observed peak–off-peak cycles, weekday–weekend contrasts, and seasonal effects all align across datasets, providing strong justification for their integration into the nowcasting framework.

Chapter 4 - Experimental Methodology

4.1 Introduction and Design

The methodology of this study was structured to develop and evaluate a nowcasting framework for urban traffic and air quality using multimodal data collected along Oxford Road, Manchester. The approach followed a stepwise design beginning with the integration of seismic signals, traffic counts, air quality indicators, and weather variables into a unified structure suitable for machine learning. A pipeline was established that ensured consistency in temporal resolution, handled missing values, and generated derived features such as stratified traffic volumes and seismic spectral components.

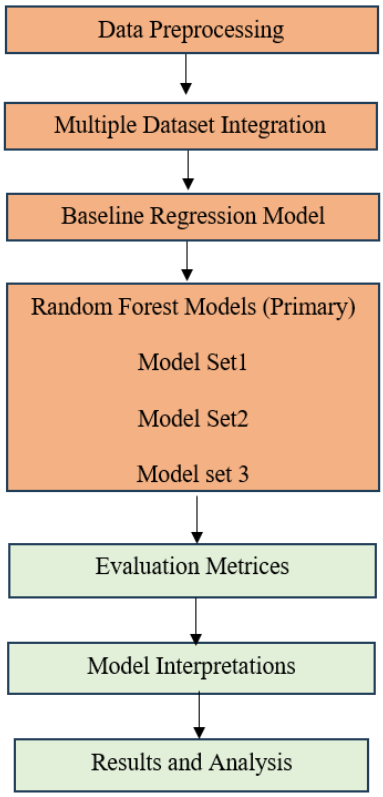


Figure 4.1 Experiment Design

The research design shown in above Figure 4.1 was organised into three main stages. First, preprocessing steps standardised all datasets to a uniform 15-minute interval and aligned them by timestamp, ensuring comparability across sources. Second, integrated

DataFrames were created to progressively test different combinations of modalities, beginning with seismic and traffic counts, and later incorporating air quality and weather. While Telraam data were initially trialled, their instability led to their exclusion from the final model set, with Drakewell data serving as the reliable traffic ground truth. Third, predictive modelling was conducted using Random Forest regression as the primary framework, supported by hyperparameter tuning.

Evaluation employed both accuracy metrics and interpretability tools (SHAP, PCA) to provide robust insights into model performance. By structuring the pipeline in this way, the methodology ensured that each stage the modelling was transparent, reproducible, and aligned with the project objectives.

4.2. Data Preprocessing

4.2.1 Handling Missing Values and Outliers

Since the datasets employed in this study originate from live monitoring systems (seismic sensors, roadside cameras, citizen-sensing devices, air quality stations, and weather APIs), they are designed to capture data continuously at fixed intervals. As a result, most of the datasets contained complete timestamp coverage with very few instances of genuinely missing values. In practice, when no activity was recorded these were encoded as zeros rather than as nulls, ensuring consistency across the time series.

A detailed inspection confirmed that the seismic, Telraam, air quality, and weather datasets exhibited minimal or no missing values, with all expected timestamps present and aligned. The only exception was in the Drakewell traffic dataset, which contained a small number of null entries. These gaps were attributed to temporary camera outages or data transmission errors.

Given the time series nature of the datasets, where observations are sequential and temporally dependent, missing values were handled using forward fill (ffill) method. Missing traffic observations were handled using forward fill given in Equation 3.2,

Let time series be:

$$X=\{x_1,x_2,\ldots,x_t,\ldots,x_T\} \quad (3.1)$$

where x_t the observed value at timestamp t.

If x_t is missing, forward fill replaces it with the most recent previous valid observation:

$$x_t = x_{t-k,} \quad where \;\; k = min\{ k > 0 \mid x_{t-k} \neq NaN\} \quad (3.2)$$

which replaces each missing entry with the most recent available value. This method, also referred to as Last Observation Carried Forward (LOCF), is widely applied in time-series contexts where short gaps can be assumed to continue the most recent trend (Little and Rubin, 2019). This ensured continuity by carrying the nearest valid observation forward to fill the gap, preserving the temporal structure of the series without introducing artificial trends. This approach is commonly applied in traffic and environmental monitoring contexts, where short gaps are assumed to follow the most recent trend in the data.

Thus each dataset retains consistent coverage across the study timeline, with filled values ensuring that all sources are aligned on the same index. This preprocessing step ensured that the integrated dataset was free of missing or inconsistent entries and suitable for use in the subsequent resampling, feature engineering, and modelling stages.

4.2.2 Temporal Resampling and Syncronization:

A key step in preparing the datasets for multimodal integration was ensuring that all sources shared a uniform temporal resolution. Since the seismic dataset (Raspberry Shake) was extracted into quarter-hourly (15-minute) spectral averages, this resolution was adopted as the common temporal index across all data sources.

- Drakewell traffic data is taken at sub-hourly granularity, which was aggregated directly into quarter-hourly counts.

- Telraam sensor data is orginally at quarter-hourly resolution, which was synchronized directly.
- Air quality data is at an hourly frequency; downscaled into quarter-hourly intervals using forward- and backward-fill imputation to avoid introducing artificial spikes.
- Weather data is also extracted hourly; interpolated linearly into quarter-hourly resolution to align with the other streams.

Formally, resampling can be expressed as below Equation 3.3

$$X_{t+k}^{(15m)} = f\left(X_t^{(1h)}\right), \quad k = 0, 1, 2, 3 \tag{3.3}$$

where $X_t^{(1h)}$ is the original hourly observation, and f() is an interpolation or forward/backward-fill function that generates quarter-hourly values for each sub-interval.

The quarter-hourly resolution reflects following considerations. Firstly, it preserves the natural granularity of seismic and traffic data, which capture rapid fluctuations in vehicle flows, for example congestion during peak hours. Secondly, it allows pollution and weather variables, originally coarser, to be aligned with traffic signals at a scale sensitive enough for nowcasting applications. This avoids the temporal smoothing inherent in hourly aggregation and ensures that short-term dynamics in urban traffic are retained. The resulting unified dataset, synchronised on a quarter-hourly grid, enabled direct comparisons and joint modelling across heterogeneous sources. This temporal alignment is critical for multimodal traffic–environment monitoring, a practice also emphasised in prior traffic forecasting studies (Kumar and Vanajakshi, 2015; García-Sigüenza *et al.*, 2023).

4.2.3 Feature Engineering

Feature engineering is a crucial step in preparing raw data for machine learning, as it transforms unprocessed variables into more informative representations that enhance model learning and predictive accuracy. In time-series applications such as traffic and air-quality analysis, raw measurements are often insufficient for capturing underlying patterns.

The raw traffic data collected from Drakewell and Telraam sensors contained a wide range of vehicle categories such as cars and light vans, rigid HGVs, articulated HGVs, buses, and motorcycles. While these categories provide granular insights into road usage, they are not directly suitable as model inputs due to their high dimensionality and potential sparsity across some categories.

To address this, a traffic stratification approach was applied. Stratification refers to the process of grouping detailed vehicle classes into broader, meaningful categories that better capture overall road usage patterns while reducing noise. Specifically, two aggregated traffic features were derived:

Total Traffic Count: The sum of all vehicles, providing an overall measure of traffic intensity at each timestamp.

Stratified Traffic Count: Disaggregated into two categories:

- Light Vehicle Traffic: Cars and light vans, representing the majority of commuter and passenger flows on Oxford Road.
- Heavy Vehicle Traffic: Rigid and articulated HGVs plus buses/coaches, which generate stronger seismic vibrations and higher pollutant emissions.

This stratification was essential for the nowcasting framework, as it enabled the models to distinguish between the impacts of different vehicle types on seismic vibrations and air quality. Light and heavy vehicles contribute differently to both vibration intensity and pollutant emissions and aggregating them into these categories preserved interpretability while avoiding overly complex feature sets.

The same methodology was applied to both the Drakewell dataset and the Telraam dataset. For Telraam, where the raw counts included categories such as car total, large vehicles, bikes, and night traffic, a similar aggregation strategy was adopted to extract light vehicles, heavy vehicles, and total traffic aligned to the Drakewell timeline. These derived features formed the foundation for subsequent machine learning models, allowing direct comparisons between seismic features, air quality data, and traffic volumes.

4.2.4 Seismic Feature Extraction

The raw seismic data from the Raspberry Shake sensor were provided in the form of spectral amplitudes across frequency bins ranging from 0.0 to 50 Hz, recorded at 15-minute intervals see Table 3.2. Each record therefore represented a time window with spectral energy distributed across frequency bands. While this raw spectral dataset (*spectra_df*) offered high resolution, its dimensionality and noise made it unsuitable for direct machine learning input.

To reduce dimensionality and highlight traffic-relevant features, a feature extractor pipeline (*SpectraFeatExtr*) was applied to the spectrum data. Following the approach of (Lecocq *et al.*, 2020; Healy, 2023), the continuous frequency domain was partitioned into bands (B1–B11) designed to capture traffic-related seismic energy. For each band, two key descriptors were extracted using below equation 3.4 and equation 3.5.

$$Power\ (B_i) = \int_{flow}^{fhigh} S(f)df \tag{3.4}$$

$$Freq\ (B_i) = arg \max_{f \in Bi} S(f) \tag{3.5}$$

where S(f) is the spectral amplitude at frequency f, and $[f_{low}, f_{high}]$ represents the limits of band B_i. In addition to band-based measures, higher-order descriptors were calculated to summarise the overall spectrum shape, including:

- Global descriptors such as peak frequency and peak amplitude were calculated to capture the strongest spectral contributions at each interval.
- Statistical descriptors: Higher-order measures, including skewness, kurtosis, and area under the curve (AUC), were derived to capture the distributional shape of the seismic spectrum.
- Dimensionality reduction: This transformation reduced the dataset from 200 raw frequency bins to 84 derived features, significantly improving interpretability and efficiency for downstream machine learning tasks.

In addition to numerical feature extraction, visual inspection of the spectrum was conducted to validate the presence of traffic signatures. Figure 4.2 shows an annotated

spectral plot with band cut points, highlighting how energy is distributed across low to mid frequencies. These bands align with traffic-induced ground vibrations, with heavy vehicles contributing more strongly at lower frequencies.

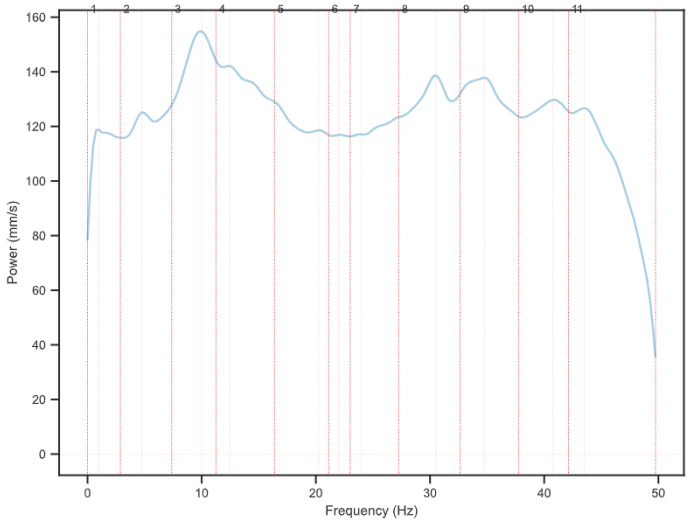


Figure 4.2 Spectrum Plot

This seismic feature extraction step was critical because it converted the raw, high-dimensional spectral data into structured, interpretable variables. These engineered features were later merged with traffic (Drakewell, Telraam) and air quality datasets to support the nowcasting framework, where the objective was to predict total traffic and stratified traffic volumes based on multi-source signals.

4.3 Data Integration

The process of Data Integration is essential because each dataset originated from independent sensors with distinct formats, sampling rates, and data coverage. To ensure consistency, all data streams were aligned on a common 15-minute timestamp index and merged using an inner join, so that only periods with complete observations across sources were retained.

To assess the relative contribution of different modalities, several integrated datasets were constructed. Each dataset progressively introduced additional feature sources, beginning with seismic spectral features alone and extending to combinations that included traffic counts, air quality pollutants, and weather variables. For each dataset,

Random Forest models were trained with target outcomes: total traffic, and startified vehicles traffic count.

Although Telraam sensor counts were initially incorporated in one set of experiments, these results were treated as exploratory only. The sensor's restricted coverage and instability meant that it did not reliably capture traffic along Oxford Road. Consequently, the primary analysis was conducted using Drakewell traffic counts, which provide a more representative and stable measure of road activity. Each integrated dataset followed a standardised structure see Table 4.1.

Rows	15-minute intervals
Columns	predictor features (seismic, air quality, weather)
Targets	Total Traffic counts and Stratified Traffic Counts

Table 4.1: Integrated Dataset

This design made it possible to evaluate the added value of including air quality and weather data, while retaining seismic-only and seismic-plus-traffic baselines for direct comparison. Thes integrated datasets are used in model builing process.

4.4 Train Validation and Test

To prevent temporal leakage and ensure fair model evaluation, the dataset covering December 2024 to March 2025 was partitioned chronologically. Following a nowcasting framework, the model was trained on historical sequences and evaluated on future unseen periods. Specifically, data from December–February was used for training, the first half of March for validation, and the second half of March for testing. This approach reflects real-world deployment conditions where models must generalise to future traffic states without prior knowledge (Essien *et al.*, 2019).

Hyperparameter optimisation was conducted exclusively within the training folds using Optuna (Akiba *et al.*, 2019), a modern optimisation framework that employs Bayesian search and pruning strategies. This ensured that tuning decisions were informed only by the training data and validation folds, without contaminating the final test set. By

embedding the tuning procedure into the evaluation protocol, the framework-maintained robustness and generalisability.

Unlike models sensitive to feature magnitudes, Random Forests are scale-invariant, as splits are based on thresholds rather than distances. Therefore, no explicit feature scaling was applied. Instead, all predictor datasets were pre-processed in Previous Chapter to ensure consistency in resolution (15-minute intervals) and completeness before entering the modelling stage.

Traffic data exhibits natural imbalance, with sparse night-time volumes and dense peak-hour flows. Instead of down sampling, which could discard meaningful information, the full distribution was preserved. This ensures that the models capture both high-volume congestion patterns and low-traffic conditions typical of early morning hours. The training process was applied consistently across three output targets. Each variant was evaluated under different feature sets, resulting in a comprehensive comparison of multimodal data integration for traffic nowcasting.

4.5 Baseline Model

4.5.1 Linear Regression Model

Supervised machine learning models are appropriate for this project as they establish explicit mappings between predictor variables and target outcomes, enabling the evaluation of how multimodal features can explain variations in road traffic intensity. While the role of regression models in traffic forecasting has already been introduced in the literature review, here Linear Regression (Lu, 2021) is applied as a baseline model to benchmark performance against more complex methods. Linear Regression assumes that the dependent variable can be expressed as a weighted linear combination of predictor features plus an error term as equation 3.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \quad (3.6)$$

where y represents the target variable, x_i are the predictor features (e.g., seismic spectral bands, traffic covariates), β_i are coefficients estimated from the training data, and ϵ is

the error term. The coefficients are obtained by minimising the residual sum of squares by below equation 3.7:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta o + \sum_{j=1}^{p} \beta i \, x_{ij}))^2 \quad (3.7)$$

This optimisation has a closed-form solution given by the normal equation 3.8:

$$\beta\hat{} = (X^T X)^{-1} X^T y \quad (3.8)$$

where X is the design matrix of predictors and y is the vector of observed responses.

Although Linear Regression provides interpretability and computational simplicity, its assumptions of linearity, independence of predictors, and homoscedastic errors are rarely satisfied in heterogeneous, multimodal datasets such as seismic–traffic–pollution streams. Previous studies has shown that regression methods can capture broad traffic trends but are limited in dealing with nonlinearities, high-dimensional interactions, and temporal dependencies (Williams and Hoel, 2003; Kumar and Vanajakshi, 2015).

In this project, the baseline model produced modest predictive accuracy, confirming its usefulness only as a comparative benchmark. These limitations motivate the use of a more flexible and robust approach. The next section introduces Random Forest regression, the primary model used in this study, which extends beyond linear assumptions to capture nonlinear relationships and complex feature interactions.

4.6 Primary Models Building

4.6.1 Random Forest Model

Decision trees form the basis of many supervised learning algorithms by recursively partitioning the predictor space into regions that minimise prediction error. While individual trees are intuitive and interpretable, they are prone to overfitting and instability, as small changes in the data can give very different splits. Random Forest by (Breiman, 2001), addresses these limitations through bootstrap aggregation and random feature selection.

In regression tasks, RF constructs an ensemble of decision trees, each trained on a bootstrap sample of the data with a random subset of predictor features considered at each split. Predictions are aggregated by averaging across all trees, which reduces variance and improves generalisation. The ensemble prediction for a given input x is given in below equation 3.9.

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \quad (3.9)$$

where $T_b(x)$ is the prediction of the b tree and B is the total number of trees. The variance of the ensemble decreases with the number of trees, as expressed by equation 3.10.

$$\mathrm{Var}(\hat{f}_{RF}) \approx \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \quad (3.10)$$

where ρ is the correlation between trees and σ^2 is the variance of an individual tree. By decorrelating trees through random feature sampling, RF achieves both variance reduction and robustness.

For this project, RF is particularly suited as the primary model because it can effectively handle heterogeneity. Seismic spectral features, pollution measures, and weather variables are recorded on different scales and distributions, yet RF does not require feature scaling and performs well with mixed data types. Moreover, it is capable of capturing non-linear relationships that are central to this study: traffic intensity and pollution levels are influenced by complex interactions between vehicle flow, meteorological conditions, and background seismic noise factors that linear models fail to represent adequately. Another advantage is its robustness to noise; seismic signals frequently contain anthropogenic and environmental interference, but by averaging across multiple decorrelated trees, RF mitigates overfitting and stabilises predictions.

4.6.2 Hyperparameters

The tuned hyperparameters that control RF performance are:

- n_estimators: Number of trees in the forest. Larger values reduce variance but increase computation.

- max_depth: Maximum depth of each tree, controlling complexity and overfitting.
- min_samples_split: Minimum number of samples required to split an internal node.
- min_samples_leaf: Minimum number of samples required at a leaf node, influencing smoothness of predictions.
- max_features: Number of features considered at each split, balancing correlation and variance reduction.
- bootstrap: Whether sampling with replacement is used to build each tree.

4.6.3 Hyperparameter Optimisation with Optuna

Hyperparameter tuning is essential to maximise RF performance, as suboptimal settings may result in either underfitting or overfitting. This study employed Optuna, a next-generation hyperparameter optimisation framework (Akiba *et al.*, 2019). Unlike grid search or random search, Optuna uses a Tree-structured Parzen Estimator (TPE) approach that adaptively samples promising hyperparameter configurations based on past evaluations. Formally, hyperparameter optimisation can be expressed as equation 3.11.

$$\theta^* = \arg\min_{\theta \in \Theta} L(y, \hat{y}\theta) \tag{3.11}$$

where θ is a set of hyperparameters, Θ the search space, and L the loss function (mean squared error in this case).

To implement Optuna, the optuna Python library was installed and integrated into the experimental pipeline alongside scikit-learn. A study object was also defined and the objective function was parameterised with the RF estimator. Within the objective function, candidate hyperparameters were sampled using Optuna's trial.suggest_int() and trial.suggest_float() methods. The search space includes all the hyperparameters, which are passed into RandomForestRegressor() from scikit-learn to evaluate performance on the validation fold.

Optuna further accelerated the optimisation process by using its pruner functionality to terminate unpromising trials early, thus allocating resources more effectively to stronger candidates. By leveraging Optuna, the tuning process was both computationally efficient and more adaptive than traditional methods, ensuring that each model variant was trained under optimised settings tailored to the complexity of its feature space.

4.6.4 Model Design

To evaluate how traffic volumes can be predicted from seismic, air quality, and weather data, three structured model sets were designed. Although the core learning algorithm remained the same, each set systematically expanded the predictor space while maintaining a consistent target variable design. Within each model set, three sub-models were generated corresponding to total traffic counts and stratified traffic counts, enabling a direct assessment of how different feature integrations influence prediction accuracy.

Model Set 1 focused on the integration of seismic spectral features derived from the *SpectraDF* with Drakewell ground-truth traffic counts. This set is to quantify the direct coupling between ground-motion energy and measured traffic volume and to establish a baseline for passive, sensor-side nowcasting without additional covariates.

Model Set 2 extended the predictor space by incorporating DEFRA air-quality measurements, alongside seismic features. Since vehicular emissions are closely coupled with traffic intensity, this configuration tested whether pollution covariates improve model accuracy and capture relationships not visible in seismic signals alone.

Model Set 3 introduced meteorological covariates from the Open-Meteo dataset in addition to seismic and air-quality inputs. This set aimed to capture the seasonal and weather-driven variability of traffic flows, particularly during holiday periods or adverse environmental conditions. By accounting for meteorological influences, this configuration evaluated the robustness of the nowcasting framework under more realistic, context-sensitive operating conditions.

Importantly, the design ensured that comparisons between sets were fair: all models targeted Drakewell counts as the ground-truth output, synchronised at 15-minute resolution, and applied the RF framework with hyperparameters optimised via Optuna. This systematic approach not only enhances reproducibility but also provides a clear rationale for adopting multimodal integration as the foundation of the nowcasting framework.

4.7 Model Evaluation Metrics

A rigorous evaluation protocol is central to ensuring the validity and reliability of predictive models in traffic nowcasting. As recommended in recent machine learning literature, multiple complementary metrics were employed rather than relying on a single measure of accuracy (Miller *et al.*, 2024). For this project, four commonly used regression metrics were applied: Coefficient of Determination (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

R^2 quantifies the proportion of variance in the dependent variable explained by the predictors (Chicco, Warrens and Jurman, 2021). It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3.12)$$

where y_i are the observed values, $\hat{y}_i$ the predicted values, and $\bar{y}$ the mean of the observed values. Values closer to 1 indicate stronger explanatory power, while negative values suggest that the model performs worse than a naive mean predictor.

MSE is a scale-dependent measure that penalises larger deviations more heavily, making it sensitive to outliers:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.13)$$

Although MSE is widely used, its squared units make it less interpretable in practical terms. To address this, RMSE is often reported alongside:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.14)$$

RMSE has the advantage of expressing errors in the same units as the target variable (traffic counts), allowing a more intuitive assessment of prediction magnitude.

The MAE provides a more robust alternative by averaging the absolute differences between observed and predicted values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.15}$$

Unlike MSE or RMSE, MAE is less sensitive to outliers, making it useful for evaluating overall model stability when the data contains occasional extreme values.

In this project, all metrics were computed on both training and testing sets to assess not only predictive accuracy but also the generalisation capacity of the models. Reporting performance on stratified traffic outputs further enabled the identification of systematic biases, such as underprediction of heavy-vehicle flows or overfitting to light-vehicle dynamics. Collectively, these metrics provide a rigorous basis for comparing the baseline Linear Regression with the Random Forest models across the experimental design sets.

4.8 Model Interpretation

In the context of urban traffic nowcasting, interpretability is essential because stakeholders such as city planners and environmental agencies require not only accurate predictions but also transparent explanations of which variables drive traffic and pollution dynamics. To address this, SHAP analysis, and Principal Component Analysis (PCA) were employed.

To obtain more granular interpretability, SHAP values were applied. SHAP builds on Shapley values from cooperative game theory (Lundberg and Lee, 2017), where the prediction is decomposed into additive contributions of each feature:

$$\text{f(x)} = \varphi_0 + \sum_{j=1}^{M} \varphi_j \tag{3.16}$$

Here, φ_j represents the contribution of feature j to the prediction, while φ_0 is the baseline expectation. Unlike feature importance, SHAP values provide **local explanations**, showing how each feature influenced individual predictions. Recent

studies emphasise SHAP's role in improving trust and transparency of complex models (García-Sigüenza *et al.*, 2023).

As seismic spectra and pollution features are high-dimensional, PCA was employed for exploratory interpretability and dimensionality reduction. PCA transforms correlated predictors into orthogonal principal components by maximising variance explained:

$$Z = XW \tag{3.17}$$

where X is the standardised data matrix, W is the eigen vector matrix of the covariance matrix of X, and Z contains the transformed principal components. The first two components (PCA1, PCA2) were visualised to assess whether seismic proxies and pollution covariates cluster differently during peak vs. off-peak traffic conditions. While PCA was not used directly in the predictive pipeline, it provided insights into the underlying structure of multimodal data.

4.9 Comparison of Models

The experimental model design culminates in a structured comparison of the three model sets to determine the added value of incorporating environmental covariates alongside seismic and traffic data. The staged comparison directly addresses the project's aim to quantitatively compare seismic signals with traffic counts and correlate those estimates with air-quality fluctuations. It also aligns with the objectives of assessing model sensitivity to stratified traffic counts and evaluating how multimodal integration strengthens real-time monitoring.

The following chapter presents the results of this evaluation. It compares the multiple Random Forest models, by reporting predictive accuracy, residual behaviours, and interpretability outcomes to demonstrate the robustness and practical significance of the proposed framework.

Chapter 5 - Results and Evaluation

5.1 Introduction

This chapter provides the outcomes of the modelling experiments, building on the data sources, preprocessing steps, and methodological design outlined. While the previous chapter detailed how the models were developed and trained, here the focus shifts to evaluating their performance and interpreting their outputs. The results presented in this chapter cover the outcomes of data preprocessing, the performance of the baseline Linear Regression model, and the Random Forest models across different experimental configurations. Evaluation metrics are reported to assess predictive accuracy, while SHAP and PCA are applied to interpret model behaviour. The chapter concludes with a discussion linking these findings to the overall project objectives.

5.2 Data Preprocessing Outcomes

5.2.1 Traffic Stratification:

The Drakewell dataset was stratified which ensured interpretability by grouping categories into meaningful predictors.

date_time	light_vehicle_traffic	heavy_vehicle_traffic	total_traffic
24-12-2024 00:00	130	4	134
24-12-2024 00:15	107	4	111
24-12-2024 00:30	113	2	115
24-12-2024 00:45	89	3	92
24-12-2024 01:00	104	1	105

Table 5.1: Sample of stratified traffic dataset.

The above Table 5.1 Shows how raw categories were transformed into aggregated traffic counts.

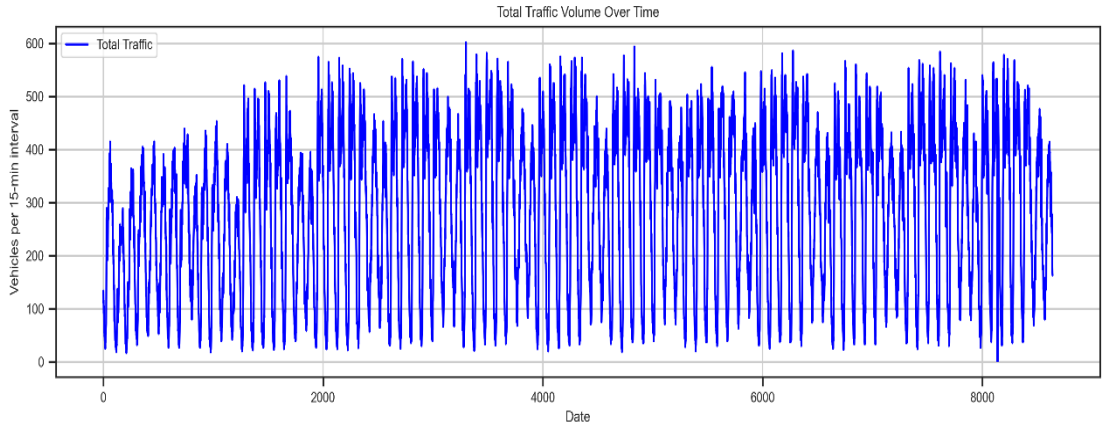


Figure 5.1: Total traffic over time (blue).

The above Figure 5.1 reveals strong diurnal cycles, with morning and evening peaks characteristic of commuter flows.

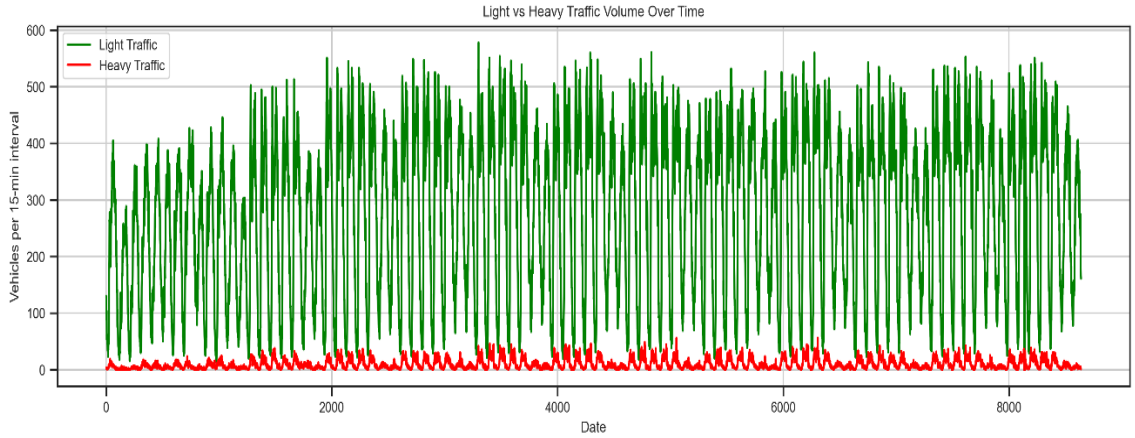


Figure 5.2: Light vs heavy traffic volumes.

The Figure 5.2 confirms that light vehicles dominate total counts, while heavy vehicles form a smaller but consistent baseline. This stratification ensured interpretability and provided the foundation for modelling total and stratified traffic volumes separately.

5.2.2 Spectral feature extraction

Seismic signals were converted into frequency-domain features using the SpectraDF pipeline.

date_time	b1_freq	b1_power	b2_freq	b2_power	b3_freq	b3_power	b4_freq	b4_power	b5_freq	b5_power
24-12-2024 00:00	2	74.688818	7.25	97.482164	9.75	126.664214	12.5	116.633878	16.5	102.023223
24-12-2024 00:15	2.75	61.600575	7.25	92.354089	10.25	124.565809	11.5	115.176477	16.5	100.647598
24-12-2024 00:30	2.75	61.841025	7.25	96.586479	10	128.679815	11.5	120.485942	16.5	99.476636
24-12-2024 00:45	2.75	57.997464	7.25	84.791977	10	105.677171	13	104.456358	16.5	92.656322
24-12-2024 01:00	2.75	58.029689	7.25	89.200325	9.75	113.477055	11.5	111.385969	16.75	101.755016
...	b11_skew	b11_kurt	b11_auc	peak_freq	peak_power	mean	std	skew	kurt	auc
...	2.24377	26.823349	21.344706	9.75	126.664214	23.554167	13.551597	307.223258	61052.348	136.665091
...	1.798848	26.078991	21.621221	10.25	124.565809	23.3941	13.484314	369.440912	59966.642	131.654275
...	2.47385	23.864685	19.823926	10	128.679815	23.656194	13.507882	309.82996	59507.669	132.706928
...	1.901474	25.723555	20.904743	10	105.677171	23.873962	13.540479	220.535851	60045.97	147.745152
...	1.816288	25.202303	21.210658	9.75	113.477055	23.513269	13.408602	310.524233	59292.346	144.272429

Table 5.2: Extracted seismic feature table (84 features).

The above Table 5.2 is the sample of extracted features which summarises reduced-dimensionality features, including band powers, peak frequency, skewness, and kurtosis, enabling traffic-relevant patterns to be represented compactly. The derived spectral dataset reduced dimensionality from hundreds of raw bins to 84 structured features, enabling efficient machine learning input.

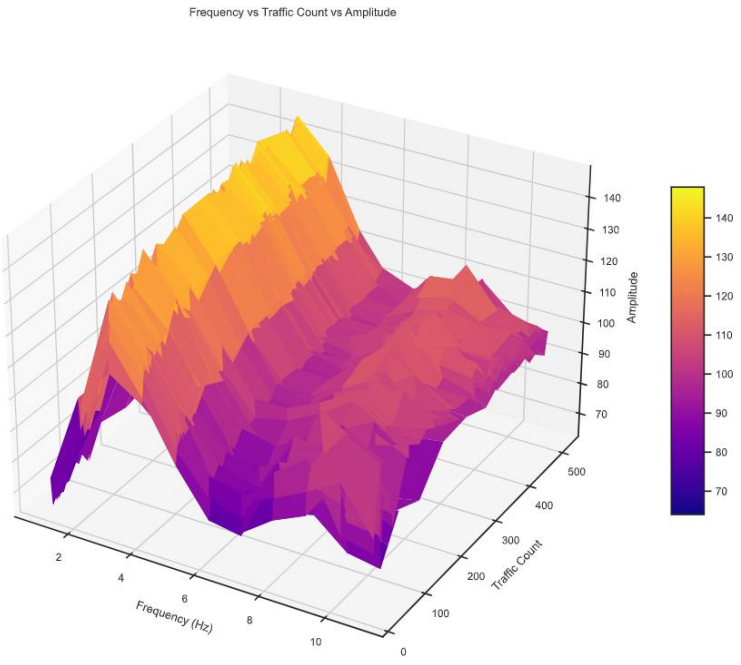


Figure 5.3: 3D seismic spectral Visualization

Figure 5.4 illustrates the relationship between seismic frequency bands, traffic counts, and amplitude. The 3D surface shows that lower-frequency bands (1–5 Hz) exhibit

stronger amplitude responses, which align with the passage of heavier vehicles generating low-frequency vibrations. Higher vehicle counts are consistently associated with elevated spectral energy, particularly in these bands, confirming that seismic signatures capture short-term traffic fluctuations. This validates the extraction of banded power features (B1–B11) as meaningful predictors for machine learning models.

5.2.3 Temporal resampling of external covariates

Air quality (DEFRA) and weather (Open-Meteo) datasets, originally at hourly resolution, were down sampled to 15-minute intervals via forward/backward fill and linear interpolation. This step aligned all modalities to the seismic resolution.

Before Resampling							
S.No	date_time	End Date	End Time	PM10	NO	NO2	NOXasNO2
2158	22-03-2025 23:00	22-03-2025	23:00:00	19.2	20.0545	49.97773	80.72753
2159	23-03-2025 00:00	23-03-2025	00:00:00	27.5	11.80628	40.20326	58.30597
After Resampling							
S.No	date_time	End Date	End Time	PM10	NO	NO2	NOXasNO2
8540	22-03-2025 23:00	22-03-2025	23:00:00	19.2	20.0545	49.97773	80.72753
8541	22-03-2025 23:15	22-03-2025	23:00:00	19.2	20.0545	49.97773	80.72753
8542	22-03-2025 23:30	22-03-2025	23:00:00	19.2	20.0545	49.97773	80.72753
8543	22-03-2025 23:45	22-03-2025	23:00:00	19.2	20.0545	49.97773	80.72753
8544	23-03-2025 00:00	23-03-2025	00:00:00	27.5	11.80628	40.20326	58.30597

Table 5.3: Air quality dataset after resampling.

Above Table 5.3 demonstrates continuity in air pollutants, avoiding gaps in the time series.

Before Resampling						
S.No	date_time	temperature_2m (°C)	precipitation (mm)	wind_speed_10m (km/h)	relative_humidity_2m (%)	cloud_cover (%)
0	24-03-2025 22:00	7.9	0	4.9	87	5
1	24-03-2025 23:00	7.1	0	5	89	10
After Resampling						
S.No	date_time	temperature_2m (°C)	precipitation (mm)	wind_speed_10m (km/h)	relative_humidity_2m (%)	cloud_cover (%)
0	24-03-2025 22:00	7.9	0	4.9	87	5
1	24-03-2025 22:15	7.7	0	4.925	87.5	6.25
2	24-03-2025 22:30	7.5	0	4.95	88	7.5
3	24-03-2025 22:45	7.3	0	4.975	88.5	8.75
4	24-03-2025 23:00	7.1	0	5	89	10

Table 5.4: Weather dataset after resampling.

Table 5.4 Shows how hourly observations were interpolated to produce smooth 15-minute profiles for temperature, wind, and precipitation. These outputs confirm that all datasets were synchronised on a quarter-hourly index, stratified where necessary, and feature-engineered for interpretability. The outcome is a unified multimodal dataset that is both temporally aligned and analytically robust, forming the foundation for the nowcasting framework evaluated in later sections.

5.3 Baseline Model Outcome

The Linear Regression model demonstrates a moderate predictive ability when applied to seismic features for estimating Drakewell traffic counts.

Group	Target	R^2	RMSE	MAE	MSE
Total Traffic Prediction	**Total**	0.7862	70.7465	55.5435	5005.0705
Stratified Traffic Prediction	**Light Vehicles**	0.7815	68.4239	53.7535	4681.8177
	Heavy Vehicles	0.5896	5.7519	4.2906	33.0845

Table 5.5 Evaluation Metrices Table-Linear Regression Model

The results Table 5.5 indicate that total traffic counts were predicted with reasonable accuracy, with an R^2 of 0.79, suggesting that nearly 79% of the variance in traffic volume is explained by the seismic predictors. For the stratified vehicle groups, light vehicles closely mirrored the total traffic patterns, confirming their dominant contribution to the overall volume. Heavy vehicles, however, showed weaker performance, reflecting their smaller presence in the dataset and higher variability compared to light vehicle flows.

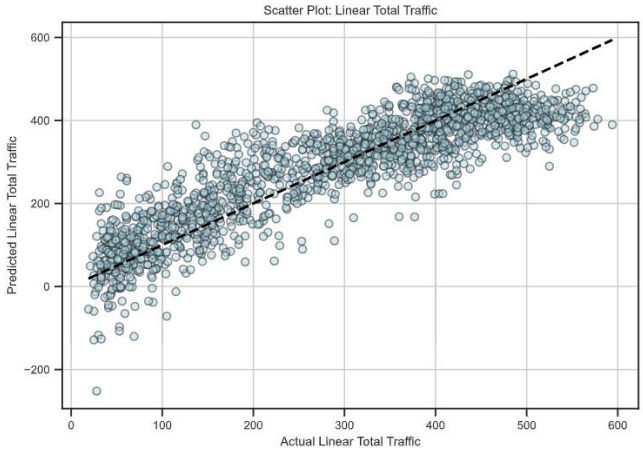


Figure 5.4: Scatter Plot- Linear Regression

The scatter plot Figure 5.5 illustrates a broadly linear alignment between predicted and observed total traffic counts, though deviations increase at higher volumes, indicating underfitting during peak hours. From the evaluation metrics, it indicates that the model achieves basic predictive precision but remains insufficient for practical nowcasting applications where higher accuracy is required. Overall, the baseline model confirms the feasibility of using seismic data to reflect traffic patterns but also reveals significant shortcomings in accuracy and robustness. These results provide a useful benchmark, justifying the transition to Random Forest regression in the next section, which is better equipped to handle non-linearities, noise, and feature interactions.

5.4 Primary Model Outcome

5.2.1 Initial Training

Following the baseline Linear Regression experiments, Random Forest was selected as the primary model to evaluate its ability to capture non-linear seismic–traffic relationships. The first exploratory trial was done with Telraam data to assess whether citizen-sensing could provide a viable alternative ground truth. While seismic features aligned moderately well with Telraam totals ($R^2 \approx 0.71$), the predictions were unstable.

Group	Target	R^2	RMSE	MAE	MSE
Total Traffic Prediction	Total	0.715366	21.552295	15.835980	464.501417
Stratified Traffic Prediction	Light Vehicles	0.693299	9.739507	6.256681	94.857990
	Heavy Vehicles	0.745877	8.083854	5.120447	65.348698

Table 5.6 Evaluation Metrices Table- RF model using Telraam Data

In particular, stratified categories were problematic. Heavy-vehicle counts were often recorded as zero across extended periods, while light-vehicle flows lacked the variability captured by Drakewell cameras see Figure 5.6. This is likely explained by Telraam placement on a permit-controlled stretch of Oxford Road, where restricted access reduces representativeness compared to Drakewell cameras covering the main corridor. Consequently, Random Forest predictions against Telraam data suffered from inflated error magnitudes and inconsistent stratified outputs see Table 5.6.

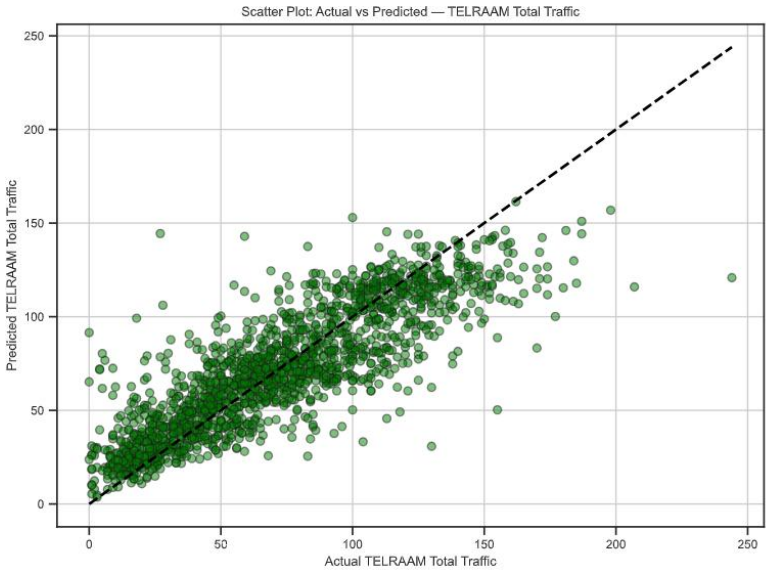


Figure 5.5: Scatter plot-Telraam

For these reasons, and consistent with the concerns highlighted in the methodology chapter, Telraam models are reported here for completeness but excluded from further evaluation.

Subsequent experiments shifted to Drakewell data, where Random Forest was first trained with scikit-learn default hyperparameters. These initial runs demonstrated stronger predictive performance, achieving $R^2 = 0.85$ for total traffic. Stratified outputs were more variable, with lower stability for heavy-vehicle flows. While this confirmed Random Forest's superiority over Linear Regression in capturing non-linear seismic–traffic dynamics, the results also highlighted the need for systematic hyperparameter tuning to further improve generalisation.

5.2.2 Hyperparameter Tuning with Optuna

To improve generalisation, Optuna was applied to optimise key hyperparameters of the Random Forest model. The optimisation objective was to minimise the Mean Squared Error (MSE) on the validation set, using the Tree-structured Parzen Estimator (TPE) algorithm for efficient exploration. Optuna's pruning functionality further accelerated the process by discarding unpromising trials. The final tuned parameters selected by Optuna are summarised in Table 5.7, which shows the optimal values applied consistently across all three model sets.

RandomForestRegressor

Parameters

n_estimators	400
criterion	'squared_error'
max_depth	30
min_samples_split	7
min_samples_leaf	2
min_weight_fraction_leaf	0.0
max_features	0.5
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	False
oob_score	False
n_jobs	None
random_state	42

Table 5.7 Optuna Hyperparameters

These tuned hyperparameters improved predictive performance across all configurations, which will be discussed further in below sections.

5.2.3 Model Set 1

The first model set evaluated the predictive relationship between seismic spectral features and Drakewell traffic counts. Using the tuned hyperparameters, the Random Forest was trained on synchronised 15-minute intervals, with separate outputs for total traffic and stratified vehicle counts.

Group	Target	R^2	RMSE	MAE	MSE
Total Traffic Prediction	Total	0.8652	56.8458	43.7896	3231.4469
Stratified Traffic Prediction	Light Vehicles	0.8570	55.3499	42.7814	3063.6157
	Heavy Vehicles	0.7283	4.6800	3.2273	21.9028

Table 5.8 Evaluation Metrices- Model Set 1

As Shown in Table 5.8, the model achieved an R^2 of 0.8652 on the test set for total traffic, with MAE ~45 vehicles and RMSE ~57 vehicles. Stratified outputs showed higher accuracy for light vehicles compared to heavy vehicles, confirming that light vehicles dominate seismic signatures while heavy vehicles, due to lower frequency, are harder to model.

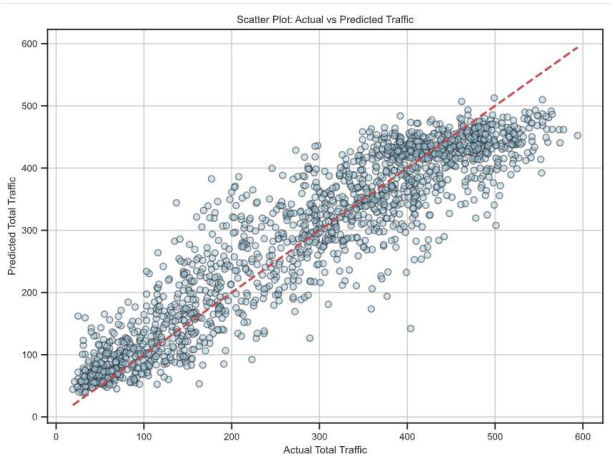


Figure 5.6: Scatter Plot-Model 1

Scatter plots Figure 5.7 showed strong alignment between actual and predicted counts. These findings establish seismic data as a strong driver of total and light-vehicle traffic

prediction, though further integration of pollution and weather data is expected to enhance robustness for more variable conditions.

5.2.4 Model Set 2

The second experimental configuration incorporated DEFRA air-quality measurements (NO_2, PM_{10}) alongside seismic spectral features and Drakewell traffic counts. The rationale for including air pollutants is that vehicular emissions are strongly coupled with traffic intensity, and their variability can act as an indirect proxy of vehicle flow dynamics.

Group	Target	R^2	RMSE	MAE	MSE
Total Traffic Prediction	Total	0.8700	56.9415	43.0904	3242.3326
Stratified Traffic Prediction	Light Vehicles	0.8623	55.7933	42.1547	3112.8902
	Heavy Vehicles	0.7525	4.8638	3.2568	24.6569

Table 5.9 Evaluation Metrices- Model Set 2

The results Table 5.9 show that the inclusion of air-quality covariates improved predictive performance compared to Model Set 1. Total traffic predictions reached an R^2 of 0.880, up from 0.865, with corresponding reductions in RMSE (57.4 vehicles/15 min) and MAE (43.9 vehicles/15 min). Stratified outputs confirm this trend: light-vehicle prediction, reflecting the dominant contribution of passenger cars to both seismic energy and pollutant emissions. Heavy-vehicle predictions also improved slightly, although their lower frequency in the dataset continues to limit precision.

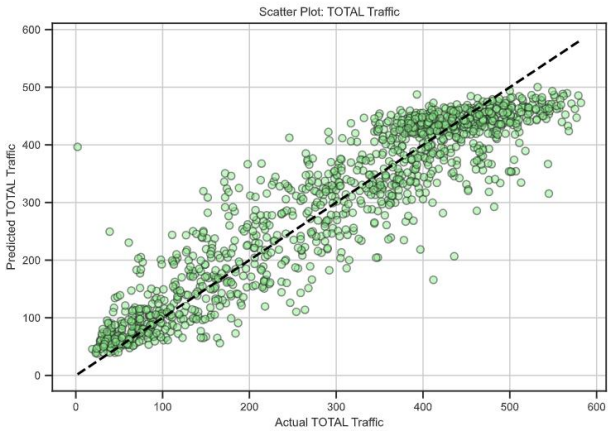


Figure 5.7: Scatter Plot-Model 2

The scatter plot Figure 5.8 demonstrates close alignment between predicted and observed counts. These findings support the hypothesis that traffic-related pollutants provide complementary information to seismic features. By capturing downstream effects of vehicle emissions, the model becomes more robust to fluctuations in flow intensity, thereby strengthening the overall nowcasting framework.

5.2.5 Model Set 3

The third model set incorporated meteorological covariates alongside seismic spectral features and DEFRA air-quality indicators. The rationale for this integration was to account for external environmental conditions that directly influence both traffic intensity and the detectability of seismic signals. For instance, adverse weather can suppress traffic volumes or alter driving patterns, while temperature and wind speed modulate pollutant dispersion, indirectly reflecting traffic density.

Group	Target	R^2	RMSE	MAE	MSE
Total Traffic Prediction	Total	0.8858	55.5065	41.8948	3080.9717
Stratified Traffic Prediction	Light Vehicles	0.8772	54.8017	41.2967	3003.2257
	Heavy Vehicles	0.7678	4.8117	3.2253	23.1526

Table 5.10: Evaluation Metrices- Model Set 3

The Random Forest trained on this multimodal dataset achieved the highest predictive accuracy across all model sets. For total traffic, the model produced an R^2 of 0.8858, with RMSE ≈ 55 vehicles/15 min and MAE ≈ 41 vehicles. Stratified outputs further demonstrated improvement, see Table 5.10, the best performance recorded for this category across all experiments. This suggests that meteorological covariates provided complementary information that enhanced the stability of predictions, particularly for the less frequent heavy-vehicle flows.

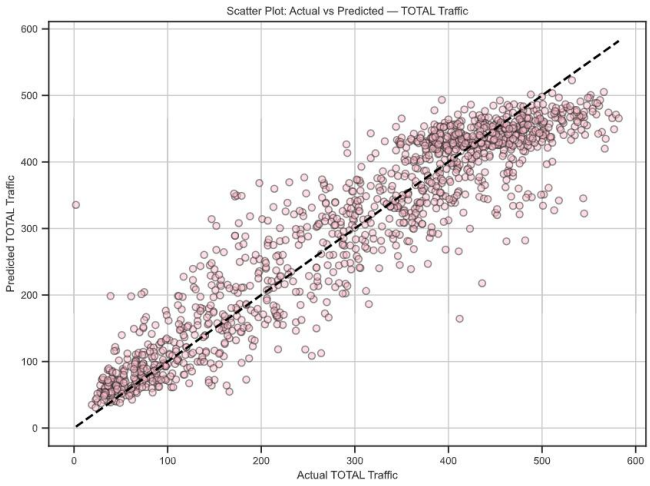


Figure 5.8: Scatter Plot-Model 3

Visual inspection of the scatter plot Figure 5.9 shows a tight clustering of predicted versus observed counts around the 1:1 reference line, confirming the improved alignment achieved through multimodal integration.

Overall, Model Set 3 demonstrates that incorporating air quality and weather covariates significantly strengthens the nowcasting framework, not only improving the accuracy of total-traffic prediction but also stabilising stratified outputs.

5.3 Model Interpretation Outcome

The interpretability analysis was conducted using SHAP values, and Principal Component Analysis (PCA). The SHAP beeswarm plot shown in Figure 5.10, provided a more granular explanation of feature impacts on individual predictions. B1_power and B2_power were shown to exert the largest influence, with higher SHAP values strongly associated with increased traffic volumes. The plot also revealed that other

mid-frequency bands provided supplementary predictive information, albeit at smaller magnitudes. This confirms the robustness of the RF model in capturing how specific spectral components translate into traffic activity.

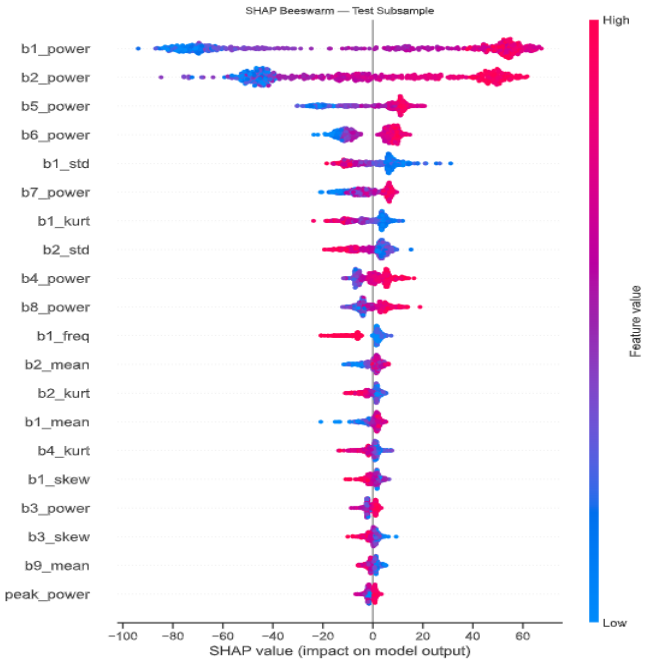


Figure 5.9: SHAP Summary Plot

Dimensionality reduction using PCA further validated these findings. The cumulative explained variance plot Figure 5.11 showed that ~95% of the variance could be explained with around 40 components, with the first two components already capturing significant structure. Examination of PC1 and PC2 loadings revealed that spectral power bands again dominated, particularly those in the low-to-mid frequency range. The PC1 vs PC2 scatter plot showed clustering patterns corresponding to higher traffic volumes, demonstrating how traffic-related seismic signatures emerge clearly from the spectral space.

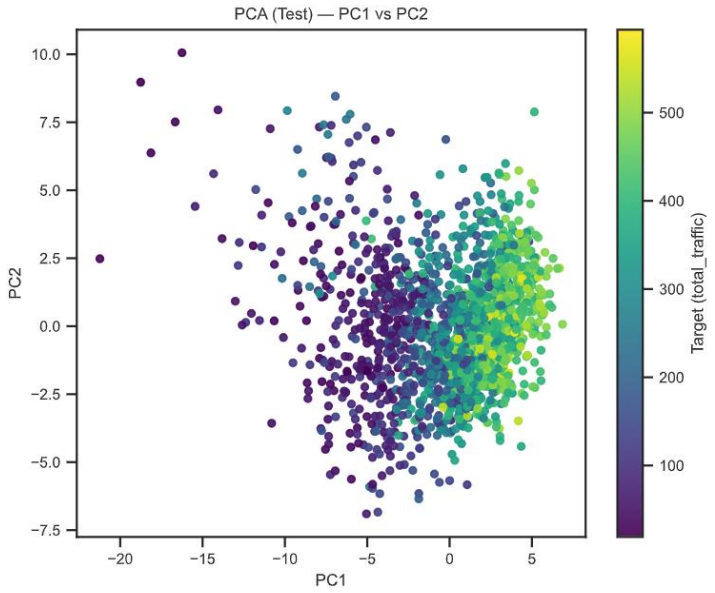


Figure 5.10: PC1 and PC2 Examination

Together, these interpretability outcomes provide strong evidence that seismic power features particularly in the 1-5 Hz range are reliable indicators of traffic activity. This confirms the theoretical expectation that low-frequency vibrations correspond to vehicle-induced ground motion, with light vehicles contributing to mid-frequency bands and heavy vehicles reinforcing the lowest bands.

5.4 Summary

This chapter has shown that while Linear Regression provided a useful benchmark, its performance was limited, particularly for stratified vehicle counts. Random Forest, enhanced with Optuna hyperparameter tuning, achieved much stronger accuracy and demonstrated the value of capturing non-linear, multimodal relationships. The staged design highlights the benefits of multimodal fusion for robust nowcasting.

Telraam trials confirmed data quality concerns, with sparse stratified counts leading to weaker models compared to Drakewell. Interpretability techniques revealed that seismic bands and NO_2 were dominant predictors, while weather added contextual value. Limitations remain seasonal restriction of the dataset, weaker performance for heavy vehicles, and citizen-sensing reliability but overall, the results meet the project objectives.

Chapter 6 - Critical Review and Future Work

6.1 Introduction

This chapter presents a critical review of the results, reflecting on how well the project objectives were met and discussing the strengths and limitations of the models. It also outlines directions for future work to improve and extend the framework. By combining critical reflection with forward-looking proposals, this chapter serves as a bridge between the empirical findings of this study and their potential development in real-world traffic and environmental monitoring systems.

6.2 Critical Review

6.2.1 Models Comparison

This section evaluates the comparative performance of the baseline Linear Regression and the Random Forest models and further analyses the incremental benefit of integrating additional data modalities. The below comparison (Table 6.1; Figure 6.1) demonstrates a clear performance improvement when moving from Linear Regression to Random Forest. For total traffic prediction, Random Forest achieved an R^2 of 0.8652 compared to 0.7862 for Linear Regression, with RMSE and MAE reduced by ~20%. This confirms that Random Forest is better suited for handling non-linear and heterogeneous interactions present in seismic and environmental data.

Target	Model	R^2	RMSE	MAE	MSE
Total Traffic Prediction	Linear Regressor	0.7862	70.7465	55.5435	5005.0705
	Random Forest Regressor	0.8652	56.8458	43.7896	3231.4469

Table 6.1: Comparison Evaluation Metrices

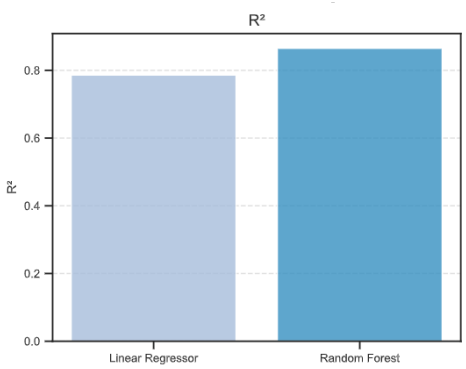


Figure 6.1: Comparison Plot

Within the Random Forest framework, progressively adding predictors from air quality and weather data further improved predictive accuracy. The below Table 6.2 demonstrating that seasonal and weather influences contribute additional predictive power.

Group	Target	R²	RMSE	MAE	MSE
Total Traffic Prediction using Randon Forest Regressor	Model Set1	0.8652	56.8458	43.7896	3231.4469
	Model Set2	0.8700	56.9415	43.0904	3242.3326
	Model Set3	0.8858	55.5065	41.8948	3080.9717

Table 6.2: Comparison Evaluation Metrices of 3 Model set

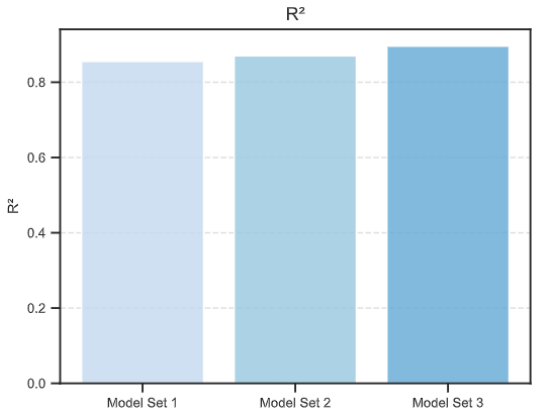


Figure 6.2: Comparison Plot between 3 Main Model

Taken together, these results in Figure 6.2 provide several insights; Firstly, Seismic features alone are highly predictive of total traffic counts, supporting their use as a low-cost, privacy-preserving proxy. Second Air quality covariates strengthen the model by capturing pollution–traffic interactions that are not always visible in seismic data, improving the reliability of stratified vehicle predictions. However, Weather variables enhance robustness, accounting for temporal variability driven by meteorology, and resulting in the most generalisable framework.

6.2.2 Strengths of the Approach

A key strength of this project lies in its multimodal integration of seismic, traffic, air quality, and weather datasets at a uniform 15-minute granularity. This resolution is particularly effective for capturing short-term traffic fluctuations, such as peak-hour congestion or seasonal variation, and ensured that the four-month study period (December 2024–March 2025) covered both winter holiday lulls and subsequent peak commuter activity.

The application of Optuna for hyperparameter optimisation further strengthened the methodology by enabling efficient and adaptive parameter tuning across multiple Random Forest models. This reduced the risks of underfitting or overfitting and ensured that each model configuration was optimised for its respective feature space.

Another notable strength is the inclusion of interpretability techniques. By combining feature importance, SHAP analysis, and PCA, the project moved beyond simple predictive accuracy to provide insights into how different features contribute to traffic variation. For instance, SHAP values identified dominant predictors across traffic states, while PCA highlighted the clustering of seismic and environmental variables under peak versus off-peak conditions.

A further strength is the direct environmental relevance of the multimodal design. By coupling seismic proxies of traffic intensity with observed air-quality measurements and meteorology at a uniform 15-minute resolution, the analysis aligns model behaviour with exposure-relevant time scales and known dispersion dynamics. This makes the results more actionable for air-quality management (e.g., short-term NO_2

exceedance risk) and links naturally to city-level greenhouse-gas objectives, given transport's substantial share of Greater Manchester's CO_2 emissions. The privacy-preserving nature of seismic sensing strengthens this alignment by offering a low-intrusion, low-cost signal suitable for scaling without the privacy trade-offs of conventional camera networks.

6.2.3 Limitations and Challenges

Despite the encouraging performance of the Random Forest models, several limitations constrain the scope and generalisability of this study. The dataset covers a four-month period with winter and early spring conditions. While this period captures both peak commuting activity and seasonal variation, it excludes longer-term dynamics such as summer traffic patterns.

Stratified traffic counts revealed that heavy vehicles constitute a small fraction of total flow, which contributed to systematic under-prediction of heavy-vehicle volumes as models learned predominantly from light-vehicle signatures. The citizen-sensing Telraam dataset underperformed compare to Drakewell ground-truth counts, in part due to its placement on a bus-gate segment of Oxford Road that restricted its field of view and reduced representativeness of overall vehicular flows. While Telraam was excluded from final evaluation to avoid introducing instability, improved siting and calibration against Drakewell could restore its value as a low-cost complement to official monitoring. A final challenge is the potential for traffic and environmental relationships to evolve over time. Without continuous retraining, model performance may degrade in such cases. Taken together, these challenges do not undermine the validity of the results within the defined scope, but they do highlight the boundaries of interpretation.

6.2.4 Broader Implications

Seismic sensing offers a privacy-preserving, low-cost complement to camera and loop systems for short-term nowcasting of flow, providing redundancy where cameras fail or coverage is partial. The unified structure can support congestion alerts, incident validation, and targeted enforcement, making networks resilient.

By integrating seismic, air-quality, and weather data on the same 15-minute grid, the approach helps separate emission-driven variability from meteorological influences. This enhances near-real-time situational awareness such as increased NO_2 risk during peaks and unfavorable dispersion and allows evidence-based intervention by dynamic signage to tactical traffic control that minimizes exposure in sensitive areas such as schools, hospitals.

Given transport's high share of city-level CO_2, a scalable and privacy-respecting traffic proxy is operationally valuable. When paired with standard per-vehicle-class emission factors, seismic-informed traffic estimates could underpin screen-level CO_2e indicators, helping authorities track the co-benefits of measures such as bus-priority, low-emission zones, and signal optimisation. While this study did not compute emissions, it demonstrates the data pathway needed to do so credibly. Unlike CCTV, seismic signals do not capture personally identifying information, improving social acceptability. This characteristic is material for sustained deployment in dense urban settings where public trust is a precondition for smart-city infrastructure.

6.3 Future Work

While this study demonstrated the feasibility of multimodal seismic traffic nowcasting, several extensions could enhance robustness, generalisability, and policy relevance.

First, the methodological aspect of the study should be improved by increasing the dataset to varying seasons to make the dataset more representative in light of traffic and environmental variability. The present analysis was limited to December–March, which reflects winter-specific dynamics; including summer would allow the models to adapt to seasonal fluctuations in traffic demand and pollutant dispersion. Although other ensemble approaches like Gradient Boosting or LightGBM may be used as other comparators, the Random Forest framework has already shown to be effective in short term nowcasting.

Second, research needs to be enhanced in terms of information coverage and infrastructure in the future. Portability of the framework to other cities, beyond the dynamics of the Oxford Road, would be tested through validation in several sites in

Manchester, or other cities in the UK. The reliability of citizen sensing is also important to be improved; the Telraam devices could be placed better, and the count of vehicles in the stratified sample could be calibrated to the official Drakewell counters to decrease the instabilities. Simultaneously, a real-time pipeline to execute inference and dashboards would also be built and would change the framework into an accessible system potentially capable of supporting real-time decision-making.

Lastly, the future directions must be focused on policy and research relevancy. Combining the framework with the currently available air quality monitoring networks across cities would allow taking actionable interventions between traffic management and pollution control. The use of causal analysis as an example before-and-after study of road closures or bus-gate enforcement may be useful to not only decouple the impacts of policy actions of natural variability. Adding quantification of uncertainty would also promote more trust in decision-making, by clearly conveying the model output reliability to the stakeholders. Combined, these extensions would not only enhance the scientific value of seismic-based nowcasting, but hasten its conversion into practical urban monitoring devices, which would assist in traffic management as well as environmental governance.

Chapter 7 - Conclusion

This dissertation examined the possibility of multimodal nowcasting of urban traffic flows with seismic signals processed together with the traffic, air-quality and weather data on the Oxford Road in Manchester. Inspired by the dual imperative of public-health protection and climate response with transport potentially representing a significant portion of Greater Manchester CO_2 this work aimed at a scalable, privacy-enabling alternative to more traditional monitoring.

The research objectives set out in the Terms of Reference were met. Raspberry Shake, Drakewell, DEFRA and Open-Meteo data were harmonised at 15-minute resolution over the period of December-March to give a single dataset that could be used operationally. A baseline Linear Regression was used to set a benchmark. Random Forest regression, optimized by Optuna, was repeatedly more accurate than the baseline, making it a strong predictor in the short term.

Experimentally formalized models revealed that seismic-only models give a good proxy of light-vehicle flow, whereas the incorporation of air-quality and weather features enhanced resiliency due to the consideration of emission dispersion interactions. Interpretability techniques helped to clarify the contribution of low-frequency seismic bands and pollution covariates during peaks, which inspired confidence in the fused model.

These technical results have clear environmental and climate relevance. It is the combination of the two structures that correlates the intensity of traffic, which can be measured seismically and confirmed by the cameras that the air-quality fluctuates at the exposure scale, which can facilitate specific short-term intervention that can limit the risk of exceeding the limit of NO_2. While the study did not estimate emissions, the same pipeline could, with standard per-class emission factors, yield screen-level CO_2 indicators, enabling cities to track the co-benefits of traffic management, bus-priority, or low-emission measures.

The temporal scope and corridor specificity constrain generalisability; heavy-vehicle sparsity and citizen-sensing placement affected stratified performance; and winter

meteorology may amplify pollutant concentrations independent of emissions. These do not undermine internal validity but delimit interpretation limits and encourage seasonal extension, multi-site validation, and better sensor location.

Overall, this paper demonstrates that seismic environmental convergence can provide interpretable, accurate, and privacy-protective short-term traffic nowcasts that can be acted upon by both traffic operations and policy consistent with air-quality/CO 2 policy. The priorities of the future relate to seasonal and geographical scaling, calibration of citizen-sensing, real-time practitioner-friendly dashboard, and combination of emission/exposure estimation and uncertainty. The dissertation can help create more intelligent, cleaner, resilient urban mobility systems by promoting a viable, morally sound, and solid monitoring channel.

References

- Adam, M.G., Tran, P.T. and Balasubramanian, R. (2021) 'Air quality changes in cities during the COVID-19 lockdown: A critical review', *Atmospheric Research*, 264, p. 105823. DOI: 10.1016/j.atmosres.2021.105823.Availableat:https://www.sciencedirect.com/science/article/pii/S0169809521003793 [Accessed 17 August 2025].
- Ahmad, R., T. Tsuji, and T. Saito (2021). "Traffic Monitoring System Based on Deep Learning and Seismometer Data". *Applied Sciences* 11(10), p. 4590. DOI: 10.3390/app11104590. Available at: https://www.mdpi.com/2076-3417/11/10/4590 [Accessed: 16 August 2025].
- Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) 'Optuna: A next-generation hyperparameter optimization framework', in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, 4–8 August 2019, Anchorage, Alaska, USA. New York: ACM, pp. 2623–2631. doi: 10.1145/3292500.3330701. Available at: https://dl.acm.org/doi/10.1145/3292500.3330701 [Accessed 28August 2025].
- Ait Ouallane, A., Bahnasse, A., Bakali, A. and Talea, M. (2022) 'Overview of road traffic management solutions based on IoT and AI', *Procedia Computer Science*, 198, pp. 518–523. DOI: 10.1016/j.procs.2021.12.278. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921025067 [Accessed 17 August 2025].
- Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M. and Ayyash, M. (2015) 'Internet of things: A survey on enabling technologies, protocols, and applications', *IEEE Communications Surveys & Tutorials*, 17(4), pp. 2347–2376. doi: 10.1109/COMST.2015.2444095. Available at: https://ieeexplore.ieee.org/document/7123563 [Accessed 17 August 2025].
- Barua, S. and Nath, S.D. (2021) 'The impact of COVID-19 on air pollution: Evidence from global data', *Journal of Cleaner Production*, 298, p. 126755. doi:10.1016/j.jclepro.2021.126755.Availableat:https://www.sciencedirect.com/science/article/pii/S0959652621008855 [Accessed 17 August 2025].
- BEIS (2019) - https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019
- Birnie, C., Ravasi, M., Liu, S. and Alkhalifah, T. (2021) 'The potential of self-supervised networks for random noise suppression in seismic data', *Artificial Intelligence in Geosciences*, 2, pp. 47–59. doi: 10.1016/j.aiig.2021.05.001. Availableat:https://www.sciencedirect.com/science/article/pii/S2666544121000149 [Accessed 17 August 2025].
- Boese, C.M., Wotherspoon, L., Alvarez, M. and Malin, P. (2015) 'Analysis of anthropogenic and natural noise from multilevel borehole seismometers in an urban environment, Auckland, New Zealand', *Bulletin of the Seismological*

Society of America, 105(1), pp. 285–299. doi: 10.1785/0120130288. Available at: https://pubs.geoscienceworld.org/ssa/bssa/article/105/1/285/323691 [Accessed 17 August 2025].

- Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.
- Buch, N., Velastin, S.A. and Orwell, J. (2011) 'A review of computer vision techniques for the analysis of urban traffic', *IEEE Transactions on Intelligent Transportation Systems*, 12(3), pp. 920–939. doi: 10.1109/TITS.2011.2119372. Available at: https://ieeexplore.ieee.org/document/5732967 [Accessed 17 August 2025].
- Chen, M., Yu, X. and Liu, Y. (2018) 'PCNN: Deep convolutional networks for short-term traffic congestion prediction', *IEEE Transactions on Intelligent Transportation Systems*, 19(11), pp. 3550–3559. doi: 10.1109/TITS.2018.2795381. Available at: https://ieeexplore.ieee.org/document/8259241 [Accessed 17 August 2025].
- Chicco, D., Warrens, M.J. and Jurman, G. (2021) 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ Computer Science*, 7, p. e623. doi: 10.7717/peerj-cs.623. Available at: https://peerj.com/articles/cs-623/ [Accessed 27 August 2025].
- Díaz, J., Ruiz, M., Sánchez-Pastor, P.S. and Romero, P. (2017) 'Urban seismology: On the origin of earth vibrations within a city', *Scientific Reports*, 7, 15296. doi: 10.1038/s41598-017-15499-y. Available at: https://www.nature.com/articles/s41598-017-15499-y [Accessed 17 August 2025].
- Dou, S., Lindsey, N., Wagner, A.M., Daley, T.M., Freifeld, B., Robertson, M., Peterson, J., Ulrich, C., Martin, E.R. and Ajo-Franklin, J.B. (2017) 'Distributed acoustic sensing for seismic monitoring of the near surface: A traffic-noise interferometry case study', *Scientific Reports*, 7(1), p. 11620. doi: 10.1038/s41598-017-11986-4. Available at: https://www.nature.com/articles/s41598-017-11986-4 [Accessed 17 August 2025].
- Egan, S., Fedorko, W., Lister, A., Pearkes, J. and Gay, C. (2017) 'Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC', *arXiv preprint*, arXiv:1711.09059. Available at: https://arxiv.org/abs/1711.09059 [Accessed 17 August 2025].
- Essien, A., Petrounias, I., Sampaio, P. and Sampaio, S. (2019) 'Improving urban traffic speed prediction using data source fusion and deep learning', in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 27 February–2 March 2019, Kyoto, Japan. Piscataway, NJ: IEEE, pp. 1–8. doi: 10.1109/BIGCOMP.2019.8679274. Available at: https://ieeexplore.ieee.org/document/8679274 [Accessed 17 August 2025].

- Fuchs, F., Lenhardt, W., Bokelmann, G. and AlpArray Working Group (2018) 'Seismic detection of rockslides at regional scale: examples from the Eastern Alps and feasibility of kurtosis-based event location', *Earth Surface Dynamics*, 6, pp. 955–970. doi: 10.5194/esurf-6-955-2018. Available at: https://esurf.copernicus.org/articles/6/955/2018/ [Accessed 17 August 2025].
- García-Sigüenza, J., Llorens-Largo, F., Tortosa, L. and Vicent, J.F. (2023) 'Explainability techniques applied to road traffic forecasting using graph neural network models', Information Sciences, 645, p. 119320. doi: 10.1016/j.ins.2023.119320. Available at: https://www.sciencedirect.com/science/article/pii/S0020025523005634 [Accessed 17 August 2025].
- Global quieting of high-frequency seismic noise due to COVID-19 pandemic lockdown measure: https://www.science.org/doi/full/10.1126/science.abd2438
- Greenhouse gas reporting: conversion factors 2019: https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019
- James, J.Q., Markos, C. and Zhang, S. (2021) 'Long-term urban traffic speed prediction with deep learning on graphs', *IEEE Transactions on Intelligent Transportation Systems*, 23(7), pp. 7359–7370. doi: 10.1109/TITS.2021.3055978.Availableat:https://ieeexplore.ieee.org/document/9346043 [Accessed 17 August 2025].
- Healy, D. (2023) 'Listening to Manchester: Using citizen science Raspberry Shake seismometers to quantify road traffic', *EarthArXiv*, preprint, published 31 May 2023. Doi: 10.31223/X57D47. Available at: https://eartharxiv.org/repository/view/5440/ [Accessed 17 August 2025].
- Kumar, S.V. and Vanajakshi, L. (2015) 'Short-term traffic flow prediction using seasonal ARIMA model with limited input data', *European Transport Research Review*, 7(3), p. 21. doi: 10.1007/s12544-015-0162-8. Available at: https://link.springer.com/article/10.1007/s12544-015-0162-8 [Accessed 17 August 2025].
- Lecocq, T., Hicks, S.P., Van Noten, K., Van Wijk, K., Koelemeijer, P., De Plaen, R.S., Massin, F., Hillers, G., Anthony, R.E., Apoloner, M.T. and Arroyo-Solórzano, M. (2020) 'Global quieting of high-frequency seismic noise due to COVID-19 pandemic lockdown measures', *Science*, 369(6509), pp. 1338–1343. DOI: 10.1126/science. abd2438. Available at https://www.science.org/doi/10.1126/science.abd2438 [Accessed 17 August 2025].
- Little, R.J.A. and Rubin, D.B. (2019) *Statistical analysis with missing data*. 3rd edn. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781119482260.
- Lu, J. (2021) 'A rigorous introduction to linear models', *arXiv preprint*, arXiv:2105.04240. Available at: https://arxiv.org/abs/2105.04240 [Accessed 29 August 2025].

- Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 4–9 December 2017, Long Beach, CA, USA. Curran Associates, 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Accessed 17 August 2025].
- Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F.Y. (2014) 'Traffic flow prediction with big data: A deep learning approach', *IEEE Transactions on Intelligent Transportation Systems*, 16(2), pp. 865–873. doi: 10.1109/TITS.2014.2345663. Available at: https://ieeexplore.ieee.org/document/6876013 [Accessed 17 August 2025].
- Masek, P., Masek, J., Frantik, P., Fujdiak, R., Ometov, A., Hosek, J., Andreev, S., Mlynek, P. and Misurec, J. (2016) 'A harmonized perspective on transportation management in smart cities: The novel IoT-driven environment for road traffic modeling', *Sensors*, 16(11), p. 1872. doi: 10.3390/s16111872. Available at: https://www.mdpi.com/1424-8220/16/11/1872 [Accessed 17 August 2025].
- Miller, C., Portlock, T., Nyaga, D.M. and O'Sullivan, J.M. (2024) 'A review of model evaluation metrics for machine learning in genetics and genomics', *Frontiers in Bioinformatics*, 4, p. 1457619. doi: 10.3389/fbinf.2024.1457619. Availableat:https://www.frontiersin.org/articles/10.3389/fbinf.2024.1457619/full [Accessed 17 August 2025].
- Open-Meteo.com: https://open-meteo.com/en/docs
- Raspberry Shake Data Visualization Tool: https://dataview.raspberryshake.org/#/AM/R6C8A/00/EHZ
- Telraam - Smart traffic counters for all transport modes - indoor and outdoor models: https://telraam.net/en/location/9000005312/2024-12-27/2024-12-27
- TFGM Drakewell traffic camera data: https://manchester-i.com/projects
- Tang, J., Chen, X., Hu, Z., Zong, F., Han, C. and Li, L. (2019) 'Traffic flow prediction based on combination of support vector machine and data denoising schemes', *Physica A: Statistical Mechanics and its Applications*, 534, p. 120642. doi: 10.1016/j.physa.2019.122358. Available at: https://www.sciencedirect.com/science/article/pii/S0378437119310471 [Accessed 19 August 2025].
- Williams, B.M. and Hoel, L.A. (2003) 'Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results', *Journal of Transportation Engineering*, 129(6), pp. 664–672. doi: 10.1061/(ASCE)0733-947X(2003)129:6(664). Available at: https://ascelibrary.org/doi/10.1061/(ASCE)0733-947X(2003)129:6(664) [Accessed 07 August 2025].

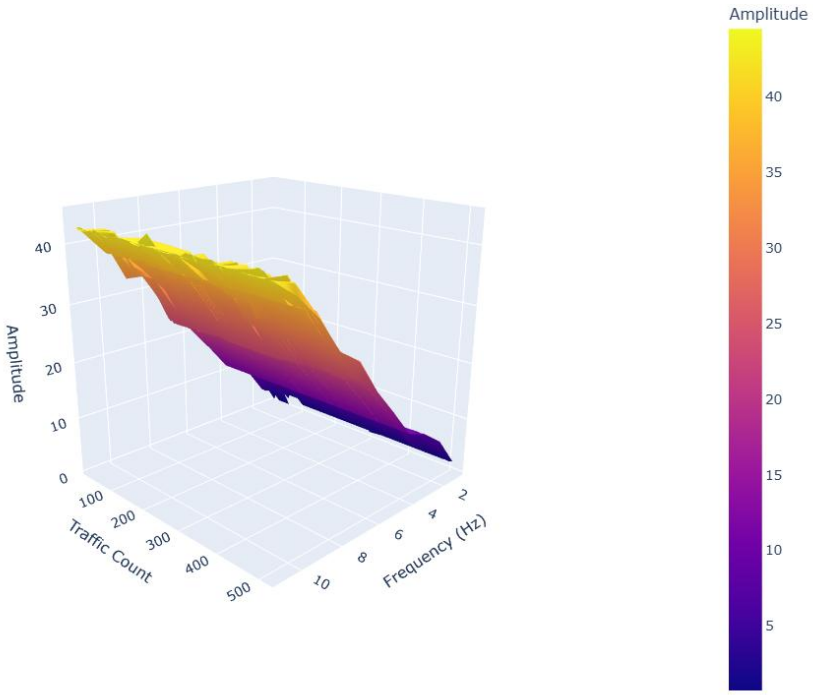
Appendix A

Below are the outcomes of methodology.

1. Spectral Extracted

date_time	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0	3.25	3.5
2024-12-24 00:00	62.26	80.65	65.20	53.98	50.15	59.13	93.55	90.52	63.62	65.20	75.12	77.59	68.63	75.20	74.93
2024-12-24 00:15	61.13	82.53	64.63	49.63	39.59	40.55	59.45	67.55	54.55	56.63	54.74	60.03	62.73	71.00	74.14
2024-12-24 00:30	62.80	81.95	64.60	49.17	40.16	39.94	57.35	68.35	55.59	56.61	58.84	60.67	62.80	67.07	72.15
2024-12-24 00:45	61.27	82.45	64.48	46.91	36.95	36.80	55.18	65.13	53.00	50.62	51.84	55.13	61.08	66.94	70.32
2024-12-24 01:00	62.58	81.82	65.42	49.34	39.07	38.68	59.57	66.07	52.28	50.49	52.77	54.88	59.46	69.55	69.54
2024-12-24 01:15	61.16	80.94	62.25	49.04	39.02	37.52	55.12	69.32	54.82	63.30	66.02	67.22	79.64	87.35	92.96
2024-12-24 01:30	60.51	82.05	64.71	46.93	39.46	39.50	55.52	66.17	55.07	52.53	52.71	55.43	58.38	63.58	66.70
2024-12-24 01:45	59.99	80.35	63.78	47.14	38.74	45.43	54.70	65.92	52.60	52.40	52.61	54.67	56.27	63.82	65.75
2024-12-24 02:00	59.61	81.69	64.22	47.18	40.01	40.00	56.86	69.14	55.97	53.90	57.54	60.91	67.92	71.76	75.29
2024-12-24 02:15	61.23	80.54	64.30	53.79	48.25	46.31	60.23	63.95	54.07	55.80	55.84	57.03	64.49	71.16	71.52
2024-12-24 02:30	59.24	81.91	63.04	48.30	39.03	37.43	55.68	63.74	55.32	56.00	57.75	62.32	67.67	74.33	80.90
2024-12-24 02:45	60.42	79.50	63.42	48.51	39.21	37.08	52.79	61.23	50.93	50.05	50.94	57.76	62.20	71.25	68.47
2024-12-24 03:00	61.73	81.09	61.84	49.05	40.45	39.63	54.90	62.97	52.61	54.35	53.20	55.24	59.05	66.55	72.13
2024-12-24 03:15	59.71	79.24	62.35	47.03	37.15	38.51	55.25	66.10	53.09	56.17	59.88	68.46	70.75	77.31	92.41
2024-12-24 03:30	61.27	79.95	63.30	48.21	41.43	41.73	56.33	63.29	50.53	53.98	54.30	58.37	61.51	76.23	77.26
2024-12-24 03:45	60.65	79.79	61.96	48.30	39.85	38.97	55.25	62.35	48.27	49.63	48.02	51.55	56.63	63.44	64.89
2024-12-24 04:00	59.62	78.81	63.27	50.07	42.81	42.16	58.07	67.24	54.76	54.25	63.65	68.69	74.95	81.88	79.83
2024-12-24 04:15	62.11	78.95	62.21	51.61	41.93	41.38	58.38	63.26	54.83	52.92	52.24	56.20	60.65	74.02	76.46
2024-12-24 04:30	60.55	80.74	62.64	50.75	45.37	46.81	62.99	70.14	57.96	59.12	63.67	70.31	79.21	88.11	86.76
2024-12-24 04:45	59.82	78.77	62.21	47.50	41.25	43.74	62.23	67.94	53.31	53.59	56.70	60.31	66.50	71.07	72.15
2024-12-24 05:00	60.88	78.40	61.91	48.91	42.39	43.95	61.40	69.38	57.71	58.08	63.46	71.34	76.34	80.54	81.51
2024-12-24 05:15	60.43	80.05	61.81	47.45	43.40	46.72	60.08	69.93	60.29	56.75	60.51	66.53	70.15	74.45	75.49
2024-12-24 05:30	58.38	76.44	61.35	48.78	44.13	48.33	65.04	71.63	60.64	57.42	60.25	63.64	70.95	81.43	84.92
2024-12-24 05:45	59.01	78.50	62.37	48.38	42.34	47.38	63.04	69.78	59.05	58.22	61.95	68.14	74.13	79.55	79.62

2. Interactive 3D Plot: Frequency vs Traffic Count vs Amplitude



3. Grid search CV selection

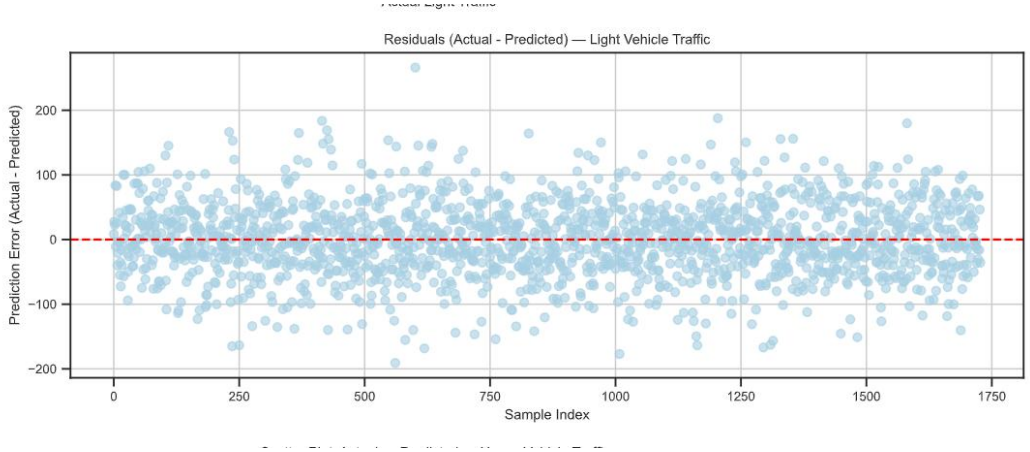
```
Fitting 3 folds for each of 36 candidates, totalling 108 fits
----- Training Performance -----
MAE: 18.81
MSE: 679.64
RMSE: 26.07
R²: 0.9713

----- Testing Performance -----
MAE: 44.99
MSE: 3369.03
RMSE: 58.04
R²: 0.8561

Best Hyperparameters: {'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300}
```

4. Optuna Tunning Process

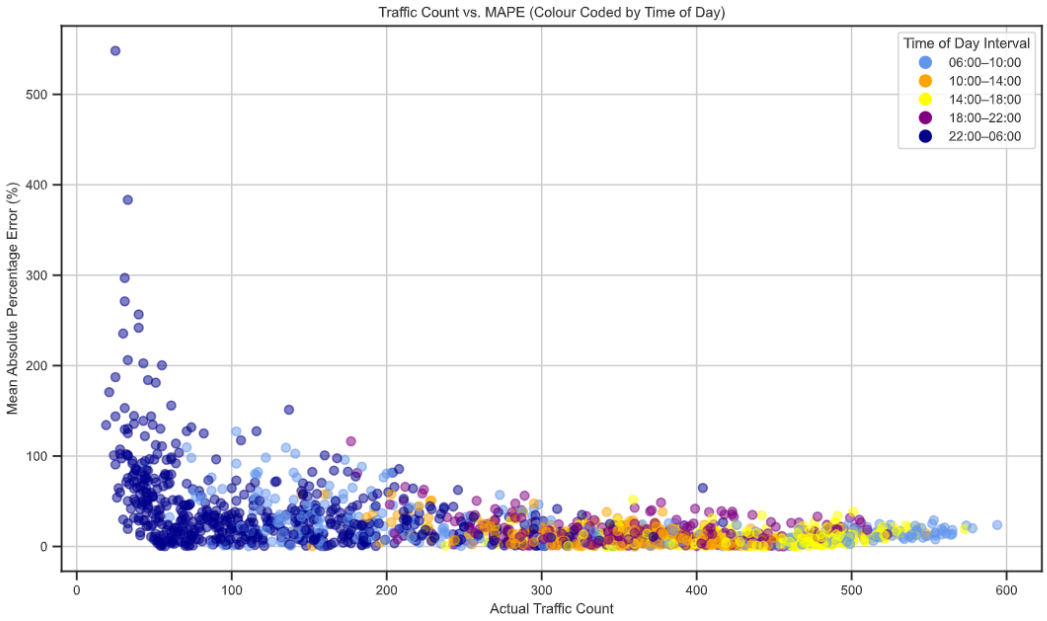
```
[I 2025-09-13 19:58:26,485] Trial 0 finished with value: -3789.216930525197 and parameters: {'n_estimators': 250, 'max_depth': None, 'min_samples_spl
it': 2, 'min_samples_leaf': 5, 'max_features': 0.5, 'bootstrap': True}. Best is trial 0 with value: -3789.216930525197.
[I 2025-09-13 19:58:34,062] Trial 1 finished with value: -4093.9319773547504 and parameters: {'n_estimators': 150, 'max_depth': 40, 'min_samples_spli
t': 3, 'min_samples_leaf': 2, 'max_features': 'log2', 'bootstrap': True}. Best is trial 0 with value: -3789.216930525197.
[I 2025-09-13 19:59:45,551] Trial 2 finished with value: -3818.224962643062 and parameters: {'n_estimators': 350, 'max_depth': 30, 'min_samples_spli
t': 4, 'min_samples_leaf': 1, 'max_features': None, 'bootstrap': True}. Best is trial 0 with value: -3789.216930525197.
[I 2025-09-13 19:59:59,058] Trial 3 finished with value: -3979.3778625667896 and parameters: {'n_estimators': 350, 'max_depth': 40, 'min_samples_spli
t': 8, 'min_samples_leaf': 5, 'max_features': 'log2', 'bootstrap': False}. Best is trial 0 with value: -3789.216930525197.
[I 2025-09-13 20:00:07,224] Trial 4 finished with value: -4107.53118833188915 and parameters: {'n_estimators': 250, 'max_depth': 10, 'min_samples_spli
t': 3, 'min_samples_leaf': 5, 'max_features': 'sqrt', 'bootstrap': True}. Best is trial 0 with value: -3789.216930525197.
[I 2025-09-13 20:00:45,264] Trial 5 finished with value: -3762.9259767501717 and parameters: {'n_estimators': 400, 'max_depth': 40, 'min_samples_spli
t': 7, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': True}. Best is trial 5 with value: -3762.9259767501717.
[I 2025-09-13 20:01:02,872] Trial 6 finished with value: -3678.110271897321 and parameters: {'n_estimators': 150, 'max_depth': 30, 'min_samples_spli
t': 6, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 6 with value: -3678.110271897321.
[I 2025-09-13 20:01:05,111] Trial 7 finished with value: -4251.794089982943 and parameters: {'n_estimators': 200, 'max_depth': 10, 'min_samples_spli
t': 3, 'min_samples_leaf': 5, 'max_features': 'log2', 'bootstrap': True}. Best is trial 6 with value: -3678.110271897321.
[I 2025-09-13 20:01:17,421] Trial 8 finished with value: -3849.75827176468 and parameters: {'n_estimators': 450, 'max_depth': None, 'min_samples_spli
t': 9, 'min_samples_leaf': 5, 'max_features': 'sqrt', 'bootstrap': False}. Best is trial 6 with value: -3678.110271897321.
[I 2025-09-13 20:01:42,931] Trial 9 finished with value: -3674.5764211879073 and parameters: {'n_estimators': 200, 'max_depth': 30, 'min_samples_spli
t': 4, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 9 with value: -3674.5764211879073.
[I 2025-09-13 20:02:03,759] Trial 10 finished with value: -4049.6679227623295 and parameters: {'n_estimators': 100, 'max_depth': 20, 'min_samples_spl
it': 5, 'min_samples_leaf': 3, 'max_features': 0.8, 'bootstrap': False}. Best is trial 9 with value: -3674.5764211879073.
[I 2025-09-13 20:02:15,971] Trial 11 finished with value: -3693.464761214491 and parameters: {'n_estimators': 100, 'max_depth': 30, 'min_samples_spli
t': 6, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 9 with value: -3674.5764211879073.
[I 2025-09-13 20:02:38,405] Trial 12 finished with value: -3676.632602987331 and parameters: {'n_estimators': 200, 'max_depth': 30, 'min_samples_spli
t': 6, 'min_samples_leaf': 4, 'max_features': 0.5, 'bootstrap': False}. Best is trial 9 with value: -3674.5764211879073.
```

```
it': 7, 'min_samples_leaf': 1, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:14:28,467] Trial 29 finished with value: -3762.9259767501717 and parameters: {'n_estimators': 400, 'max_depth': None, 'min_samples_s
plit': 7, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': True}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:16:32,284] Trial 30 finished with value: -7350.348060861722 and parameters: {'n_estimators': 500, 'max_depth': 20, 'min_samples_spli
t': 7, 'min_samples_leaf': 2, 'max_features': None, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:17:30,352] Trial 31 finished with value: -3670.6658185053616 and parameters: {'n_estimators': 450, 'max_depth': 30, 'min_samples_spl
it': 8, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:18:22,271] Trial 32 finished with value: -3685.9123134226406 and parameters: {'n_estimators': 400, 'max_depth': 30, 'min_samples_spl
it': 9, 'min_samples_leaf': 1, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:18:29,178] Trial 33 finished with value: -3925.7669476530755 and parameters: {'n_estimators': 350, 'max_depth': 40, 'min_samples_spl
it': 7, 'min_samples_leaf': 2, 'max_features': 'log2', 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:18:57,276] Trial 34 finished with value: -3757.291075093486 and parameters: {'n_estimators': 350, 'max_depth': 30, 'min_samples_spli
t': 6, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': True}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:19:11,090] Trial 35 finished with value: -3817.0228837999275 and parameters: {'n_estimators': 450, 'max_depth': 30, 'min_samples_spl
it': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:19:21,377] Trial 36 finished with value: -4088.1415139527903 and parameters: {'n_estimators': 400, 'max_depth': 40, 'min_samples_spl
it': 8, 'min_samples_leaf': 1, 'max_features': 'log2', 'bootstrap': True}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:19:49,299] Trial 37 finished with value: -3778.627707019801 and parameters: {'n_estimators': 300, 'max_depth': 10, 'min_samples_spli
t': 9, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:20:09,926] Trial 38 finished with value: -3758.148487544068 and parameters: {'n_estimators': 250, 'max_depth': 30, 'min_samples_spli
t': 2, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': True}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:21:22,267] Trial 39 finished with value: -4052.833359424265 and parameters: {'n_estimators': 350, 'max_depth': None, 'min_samples_sp
lit': 7, 'min_samples_leaf': 1, 'max_features': 0.8, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:22:17,577] Trial 40 finished with value: -3668.485041196804 and parameters: {'n_estimators': 400, 'max_depth': 40, 'min_samples_spli
t': 6, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:22:52,559] Trial 41 finished with value: -3668.109920241223 and parameters: {'n_estimators': 300, 'max_depth': 30, 'min_samples_spli
t': 10, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:23:34,518] Trial 42 finished with value: -3669.62157157257 and parameters: {'n_estimators': 350, 'max_depth': 30, 'min_samples_spli
t': 10, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:24:06,305] Trial 43 finished with value: -3670.4681885764185 and parameters: {'n_estimators': 250, 'max_depth': 30, 'min_samples_spl
it': 9, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:24:12,391] Trial 44 finished with value: -3957.3013386141533 and parameters: {'n_estimators': 200, 'max_depth': 30, 'min_samples_spl
it': 10, 'min_samples_leaf': 3, 'max_features': 'log2', 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:24:39,841] Trial 45 finished with value: -3771.0240333367124 and parameters: {'n_estimators': 250, 'max_depth': 10, 'min_samples_spl
it': 8, 'min_samples_leaf': 3, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:24:46,000] Trial 46 finished with value: -3988.911908573818 and parameters: {'n_estimators': 300, 'max_depth': 30, 'min_samples_spli
t': 9, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'bootstrap': True}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:25:25,884] Trial 47 finished with value: -3671.649602803058 and parameters: {'n_estimators': 350, 'max_depth': 30, 'min_samples_spli
t': 10, 'min_samples_leaf': 4, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:27:11,825] Trial 48 finished with value: -7333.831531274828 and parameters: {'n_estimators': 450, 'max_depth': 30, 'min_samples_spli
t': 8, 'min_samples_leaf': 3, 'max_features': None, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
[I 2025-09-07 16:27:36,364] Trial 49 finished with value: -3675.798545866361 and parameters: {'n_estimators': 150, 'max_depth': 30, 'min_samples_spli
t': 4, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': False}. Best is trial 26 with value: -3665.3974731025132.
Best CV score (neg MSE): -3665.3974731025132
Best hyperparameters: {'n_estimators': 400, 'max_depth': 30, 'min_samples_split': 7, 'min_samples_leaf': 2, 'max_features': 0.5, 'bootstrap': False}
```

5. Residual Plot



6. MAPE Plot Model 1



Appendix B

Jupyter Notebook code File :