Data-Driven Solutions for Transport Policy Interventions Using Machine Learning and Optimisation Methods

Farzaneh Farhadi

Supervisors: Prof. Roberto Palacin
Prof. Phil Blythe CBE

Advisor: Mr Tim Gammons, Arup Group

School of Engineering
Newcastle University

This thesis is submitted for the degree of
Doctor of Philosophy

November 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in Newcastle University, or any other higher education institute. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text.

Farzaneh Farhadi
November 2024

Abstract

The use of data-driven approaches, machine learning techniques, and optimisation methods in transport policy-making and the subsequent implementation of policy commitments has seen a substantial growth in recent years. The potential benefits of big data in transportation are significant, but the challenge lies in *extracting knowledge from data* to inform policy design, implementation, and validation. There is a significant gap in the literature on the application of machine learning and optimisation methods applied for policy validation and its implementation in transportation.

The aim of this PhD project it to study the potential of data-driven techniques for analysing and validating the objectives of policy interventions, and implementing policy commitments in the transport arenas. To achieve the aim of this PhD research, the following research questions are specifically addressed: (a) Given the large volume of data gathered from the transportation network, how to find data types that are relevant to a policy objective? (b) What machine learning techniques are suitable for combining large datasets, processing the data, and validating a policy objective? (c) Can these large dataset techniques be integrated in the implementation of policy commitments?

The study's methodology involves identifying relevant data types for the proposed policy objectives, selecting appropriate machine learning techniques for processing data and validating the policy objectives, and determining the potential use of these techniques for policy commitment implementation. Two frameworks have been designed to tackle the specific challenge of finding datasets related to the policy objective and validating policy interventions using machine learning techniques. A third framework have been designed for finding the best implementation of policy commitments using multi-objective optimisations. The term 'framework' is used since the proposed approaches are high level and flexible, and can be applied to different policy objectives. The details and the choice of machine learning models can be decided depending on the specifics of the policy objective.

The study focuses on two case studies aimed at improving air quality and reducing greenhouse gas emissions, which are essential components for meeting the UK's target of achieving net-zero emissions by 2050. Datasets from the Newcastle Urban Observatory and open-source datasets gathered from the industrial Case-funding partner of this PhD, Arup, is

used for this research with policies focusing on clean air zones and the transition towards electric vehicles (EVs).

The objective of the clean air zone policy is to reduce exposure to harmful levels of NO2, while the most important policy commitment of transitioning towards electric vehicles is to expand the electric vehicle charging infrastructure, ensuring that the EV charging infrastructure meets the demand of users. Two data-driven approaches are employed, including machine learning models for policy objectives and the use of simulation in combination with optimisation for implementing policy commitments. By leveraging these advanced techniques, this research aims to provide valuable insights for policymakers, helping them make more informed decisions when planning and implementing transportation policies.

In the first case study, common machine learning classifiers are used, which include Decision Tree, K-Nearest Neighbours, Gradient-Boosted Decision Trees, and Light Gradient-Boosting Machine (LGBM). It is shown that the constructed models share common conclusions about the importance of features in predicting NO2 concentrations with LGBM performing best in capturing the relations in the dataset with accuracy 88%. Subsequently, historical data is used to model air quality in Newcastle upon Tyne both assuming with and without the implementation of the clean air zone. The long short-term memory model is used to predict the NO2 concentration with root mean square error of 0.95. The approach shows the use of machine learning methods in analysing and validating the objectives of interventions in transportation systems. The role of machine learning can be summarised as predicting what is going to happen in the future if the policy is not implemented (using available historical data), and predicting the air quality and other related variables using transport behaviour changes in response to the implemented policy.

The second case study is the expansion of the EV charging infrastructure of Newcastle upon Tyne, UK. An optimisation model is developed to estimate and optimise the charging points types, charging points quantity, charging points locations, total expenditures, and utilisation of charging points for four different future energy scenarios. Quantitatively, the optimal solutions recommend installing higher number of faster charging points to reduce the percentage of slower charging points from the current 60% to around 25% in the four scenarios. Still, the optimal solutions put priority on the slower charging points (around 25%), with faster charging points having smaller portions each around 10%-13%. The optimisation shows that while 7kW charging dominates the market currently, it is more beneficial to improve charging efficiency and reduce investment costs by having a higher percentage of installations from other types of charging points in the future installations. The results also illustrate the spatial distribution of charging points, with higher concentrations in urban areas and near major roads.

This PhD research contributes to the body of knowledge on using *quantitative methods* to validate the objectives of policy interventions and implement policy commitments in transportation. The findings will provide policymakers with valuable insights to make more informed choices to improve transportation systems. This research provides an opportunity to explore the benefits and challenges of using data-driven approaches, machine learning techniques, and optimisation methods to improve transportation planning and policy-making. The study's methodology and results will be significant for policymakers, stakeholders, and researchers interested in using quantitative methods to improve transportation systems.

I dedicate this thesis to my beloved, a steadfast source of support and inspiration, and to my family. To my father, who passed away during my PhD journey but whose love and encouragement have always been a guiding force, and to my mother, who continues to support and inspire me.

Acknowledgements

I would like to begin by expressing my deepest gratitude to my academic supervisors, Professor Roberto Palacin and Professor Phil Blythe, for their exceptional guidance, unwavering support, and invaluable mentorship throughout the course of this research. Their expertise, dedication, and enthusiasm have been a constant source of inspiration, and I have learned a great deal under their supervision.

I would like to thank the members of my thesis approval and progression panel, Professor Nilanjan Chakraborty, Dr Dilum Dissanayake and Professor Margaret Bell, for their valuable feedback and insightful comments that have helped me refine my work and achieve the highest possible standards.

Furthermore, I would like to acknowledge the generous financial support provided by the EPSRC grant EP/V519571/1 and Arup Group Limited. Special thanks goes to Mr Tim Gammons, the Director of Intelligent Mobility at Arup, for his generous support and invaluable suggestions throughout this research. He was particularly instrumental in facilitating my research goal by connecting me with the right experts within the Arup organisation and ensuring continuous support. Their contributions have been critical in providing the necessary resources to complete this research. I would also like to express my gratitude to Arup Group Limited for providing valuable data, resources, and expertise that have significantly contributed to the success of this research.

I extend my sincere appreciation to my colleagues and friends at the School of Engineering, whose intellectual discussions and assistance have fostered a stimulating and enjoyable research environment. I am deeply grateful for their support throughout my studies.

I would like to acknowledge the staff at Newcastle University for their helpfulness and accommodating attitude, which have made my research experience a pleasant and productive one. To my friends and fellow students, I express my heartfelt thanks for their unwavering support, encouragement, and camaraderie throughout this journey. Their belief in me has been a source of strength and motivation.

Finally, I am eternally grateful to my family for their unconditional love, support, and understanding during my studies. Their unwavering support and encouragement have been the bedrock of my success.

List of Publications Associated with the Thesis

Journals

- A journal article based on the results of Chapter 6: Farhadi, Farzaneh, Shixiao Wang, Roberto Palacin, and Phil Blythe. "Data-driven multi-objective optimization for electric vehicle charging infrastructure." iScience (2023). Published. DOI: https://doi.org/10.1016/j.isci.2023.107737

- A journal article based on the results of Chapters 4–5: Farhadi, Farzaneh, Roberto Palacin, and Phil Blythe. "Machine Learning for Transport Policy Interventions on Air Quality." IEEE Access (2023). Published. DOI: https://doi.org/10.1109/ACCESS.2023.3272662

Conference Papers

- A conference paper based on the results of Chapter 6: Farhadi, F., Wang, S., Palacin, R. and Blythe, P., 2023. Efficient electric vehicle charging infrastructure planning using data-driven optimization. Institution of Engineering and Technology (IET). Published. DOI: https://doi.org/10.1049/icp.2023.3116

- A conference paper based on the results of Chapter 5: Farhadi, Farzaneh, Roberto Palacin, and Phil Blythe. "Data-driven framework for validating policies: Air quality case study." 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022. Published. DOI: https://doi.org/10.1109/ITSC55140.2022.9922587

- A conference paper based on the methodology of Chapters 4–5: Farhadi, F., Palacin, R. and Blythe, P., 2023. Estimating the Impact of Electric Vehicle Adoption on CO2 Concentrations Using Classification and Time Series Methods. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) (pp. 368-374). IEEE. Published. DOI: https://doi.org/10.1109/ITSC57777.2023.10421805

- A conference based on the results of Chapter 6: Farhadi, F., Palacin, R. and Blythe, P., 2023. Data-Driven Framework for Implementing Policy Commitment, Universities Transport Study Group (UTSG), Cardiff (July 2023), Nominated for Smeed prize.
- A conference based on the results of Chapter 6: Farhadi, F., Palacin, R. and Blythe, P., 2023. Multi-Objective Optimisation Method for Electric Vehicle Charging Infrastructure, 8th Annual Electric Vehicle Conference, Edinburgh Napier (June 2023).
- A conference based on the results of Chapter 5: Farhadi, F., Palacin, R. and Blythe, P., 2023. Machine Learning Methods for Policy Interventions. ECTRI Young Researchers, Lisbon, Portugal (May 2023).
- A conference based on the results of Chapter 4: Farhadi, F., Palacin, R. and Blythe, P., 2022. Machine Learning Methods for Identifying Relevant Data in Transport Policy Interventions. Universities Transport Study Group (UTSG), Edinburgh Napier (July 2022).

Presentations and Posters

- May 2023: Presentation: Data-Driven Approach using Simulation and Optimisation for Transport Policy Commitments: Department for Transport, London.
- April 2023: Presentation: Accelerating Electromobility in the UK and Singapore: a shared challenge: Singapore.
- January 2023: Abstract, Poster Presentation: Data-Driven Solutions in Transport Systems Using Machine learning and Optimisation Methods: PGR Conference, Newcastle University.
- September 2022: Poster Presentation: Data-Driven Framework for Validating Policies Objective: Air Quality Case Study, 50 years of Transport Research and Teaching, Newcastle University.
- July 2022, Presentation: Complex Transport Systems Bringing Together Technology, People and Process, Intelligent Mobility Arup Global Call, Arup Group.
- June 2022, Presentation: Using Data-Driven Approaches in Decarbonisation Interventions in Transport Systems, Smart Energy and Decarbonisation Summer School, Durham University.

- May 2022, Poster Presentation: Multi-Modal Data-Driven Solutions for Designing Intervention Objectives in Transport Systems, Centre for Mobility and Transport launch event, Newcastle University.
- June 2021, Poster Presentation: Data-Driven Solutions for Validating Policies Intervention in Transport System, PGR conference in engineering, Newcastle University.

Session Chair, Conference Committee, and Reviewed Papers

- June 2023, Sessions Chair: 8th Annual Electric Vehicle Conference, Edinburgh Napier University.
- June 2023, Reviewed Papers: IEEE Intelligent Transport Systems Conference.
- January 2023, Sessions Chair: PGR Conference in Engineering, Newcastle University.
- October 2022, Sessions Chair and Reviewed Papers: IEEE Intelligent Transport Systems Conference.
- June 2021, Committee Member and Sessions Chair: PGR Conference in Engineering, Newcastle University.

Contents

List of figures

List of tables

Nomenclature

Roman Symbols

AC	Alternating current
ACO	Ant colony optimisation
ANN	Artificial neural networks
AQI	Air quality index
BEIS	Department for Business, Energy, and Industrial Strategy
BEV	Battery electric vehicles
BRT	Boosted regression trees
CDF	Cumulative distribution function
CNN	Convolutional neural networks
CO_2	carbon dioxide
DC	Direct current
Defra	Department for Environment, Food & Rural Affairs
DfT	Department for Transport
DNN	Deep neural networks
DT	Decision tree
EFT	Emissions factors toolkit
EV	Electric vehicle

GA	Genetic algorithm
GBDT	Gradient boosted decision trees
GIS	Geographic information systems
GMCAZ	Greater Manchester Clean Air Zone
GRU	Gated recurrent unit
HGVs	Heavy goods vehicles
ICT	Information and communication technologies
IoT	Internet of Things
KNN	K-nearest neighbours
LAs	Local authorities
LGBM	Light gradient boosting machine
LSOAs	Lower layer super output areas
LSTM	Long short-term memory
MOGA	Multi-objective genetic algorithm
MSOAs	Middle layer super output areas
NO_x	Nitrogen oxides
NTEM	National trip end model
NTS	National travel survey
OAs	Output areas
PM_x	particulate matters
ppb	Parts per billion
SDGs	Sustainable development goals
SVM	Support vector machines
SVR	Support vector regression

UKCRIC	UK collaboratorium for research in infrastructure and cities
UO	Urban observatory

Chapter 1

Introduction

This introductory chapter sets the scene for the research area of the thesis and provides a background on the use of data-driven methods in transport systems. The research gap in the literature is briefly explained with focus on the use of data-driven solutions in transport policy interventions and two case studies one focused on clean air zones and the other considering the expansion of electric vehicles charging infrastructure. The aim, research questions, and objectives of the thesis are then provided. The chapter also describes the potential benefits, limitations, and finally the outline of the thesis with the interrelation of the chapters.

For the purpose of reproducibility of the results of this thesis, the links to original datasets used for generating the results are provided in a separate section on Data and Code Availability. All original code has been deposited at Github and is publicly available with links reported in the section.

1.1 Background

The current increasing use of Information and Communication Technologies (ICT) in the transport sector allows for collection of unprecedented volumes of data across all modes and transport systems. This 'big data' has generated a strong interest in the transport research community as well as in the relevant industries, transport stakeholders and delivery bodies, and among policy makers (Urbanek, 2018). In order to utilise the potential benefits of ICT, additional research is needed in terms of digital infrastructure, development of digital skills, and understanding of the underlying data resources, as described in a report from Government Office for Science (Cottrill, 2018).

The considerable amount of data available from all modes of transport can be used to improve performance, efficiency, service provision, safety and security of the transport network for its users (Teoh et al., 2019). Since the large amount of data comes from different

sources, it is a challenging task to extract knowledge from the data as it is brought together. Transport sector is the second-largest contributor of carbon emissions in Europe, and data from the transport network can be used to influence the behaviour of the users and reduce the carbon emissions (Hoen et al., 2014). The newly available data has been successfully applied to solve many important problems in transport (Veres and Moussa, 2019) including the prediction of traffic status, destination, and demand.

Big data points to the datasets that are large, complex with different types of data and can be very challenging to analyse. For instance, in a single project on Transport for London (TfL), over 10 billion data points were collected. One of the most important challenges of big data in transport is the fact that data is gathered, stored, and managed by multiple organisations, with various levels of data accessibility (Jain et al., 2016).

The Open Data Institute (Northhighland worldwide consulting, 2018) has divided the data accessibility in three different groups of *closed*, *shared* and *open* data. The closed data is available only for internal people of an organisation, shared data is available for the Group-based access by authentication, and open data is available for all the businesses, organisations and people. Open data infrastructure is essential to enhance transport and elevate the public transport system's reliability, comfort, and safety for individuals of diverse conditions. According to a paper that was published by Transport Systems Catapult (Knuth, 2016), it was estimated that not making the transport datasets open and accessible for everyone can result in Âč15bn in lost direct and indirect benefits to the UK by 2025. Transport organisations are consisting of many local authorities and private sectors. These can create many barriers in data sharing as each of these organisations have their own policy and concerns for sharing data between different organisations.

According to a report published by the Northhighland worldwide consulting (2018) barriers and concerns about data sharing can be divided into the three categories of external barriers, internal barriers, and cultural barriers (Catapult Transport Systems, 2017). Note however that in case datasets are used for developing sensitive policy options, such barriers may be created intentionally to preserve confidentiality of the developed data-driven approach and the results. Massive data are collected from various sources by different public agencies and private sectors which rarely communicate with each other. This means the data is only used and analysed for a particular piece of a transport system, such as an intersection, a stretch of freeway, or specific bus routes. In order to achieve optimal operation of the transport network, various *multi-modal datasets* need to be collected and integrated from roadway traffic, public transit, parking, incidents, bicycles, buildings, social media, and so on (Litman, 2017). These multi-modal datasets can then be analysed using novel statistical

learning algorithms for optimal decision making and designing interventions for one mode of transport while taking into account the impact on other modes.

The recent advances in sensing and the storage of large volumes of data have introduced a revolution in the way transport systems are designed and managed (Torre-Bastida et al., 2018; Zhang et al., 2011). Real-time and historical information gathered from the transport network enable us to learn, and develop interventions to improve the efficiency of the network and make it sustainable. This has opened new research directions in the transport research community and has generated a strong interest in the relevant industries and among policy makers to move towards intelligent transport systems in which data collection and analysis plays a fundamental role in decision making, design, and increased efficiency (Hoen et al., 2014; Teoh et al., 2019; Urbanek, 2018). The survey papers by Zhang et al. (2011) and by Torre-Bastida et al. (2018) analyse the latest research efforts on developing data-driven intelligent transport systems while discussing the functionality of the key components and future research directions to tackle the related challenges.

Using large volume of data effectively in the transport sector bring its own challenges and opportunities for further research. Specific data-driven solutions that have been studied by researchers recently to address challenges in the transport sector are as follows. The work by Sarabia-Jácome et al. (2020) studies automation of operations in seaport logistics and proposes a Big Data architecture for secure data sharing and promoting an intelligent transport multimodal terminal for improving decision making. The work by Dai et al. (2019) proposes a data-driven approach to construct an accurate model for predicting short-term traffic flow by combining the spatio-temporal analysis with a Gated Recurrent Unit. The article by Ma et al. (2020) proposes a convolutional neural network architecture for predicting multi-lane short-term traffic flow. Other applications of data-driven methods in the transport sector include building preventive maintenance decision models of urban transport systems (Li et al., 2019), optimising fuel consumption and sulfur oxide (SOx) emissions using big data analytics techniques to make environmentally sustainable operations in maritime shipping (Zhao et al., 2019), and predicting transport carbon emission using urban big data to mitigate climate change (Lu et al., 2017; Sharma and De, 2022).

The previous research on validating the objectives of transport policy interventions using data-driven methods is very limited. Current model-driven approaches are not adequate to cope with dynamic urban environments and the increasing complexity of transport networks. The exponential growth in data availability from diverse sources in transport infrastructure have created the potential of harnessing this data for more insightful, dynamic, and predictive transport policy-making. Therefore, substantial amount of research is needed to effectively integrate data-driven methods in various stages of designing transport policies, validating the

objectives of policy interventions, and implementing policy commitments. Technological advancements in machine learning, artificial intelligence, and big data analytics is currently at a stage where their application can significantly transform transport policy design and implementation, making systems more adaptive and responsive to real-world conditions.

Addressing the complexities of modern transport requires interdisciplinary research that combines data science, transport policy-making, urban planning, and environmental science. This approach can ensure that data-driven methods are not only technologically sound but also socially equitable and environmentally sustainable. A successful example is *Transport for London*, where big data is used efficiently to design new interventions in case of disruptions for improving the public transport for millions of passengers. This showcases that the availability of data creates possibilities for having more efficient implementation of policy commitments.

1.1.1 Clean Air Zones

Clean air zones are being designed and implemented by local authorities to improve the air quality, reduce pollution, improve public health, and create a more sustainable urban environment. The clean air zone can apply specific requirements to both commercial operators and private motorists. This include buses and coaches, taxis, vans, light goods vehicles, minibuses, heavy good vehicles, and private hire vehicles that are not compliant with the intended emission standards. The UK governmentâĂŹs Clean Air Strategy (DEFRA, 2019) includes the implementation of clean air zones in large UK cities. By charging road vehicles of specific classes, the main purpose of introducing and implementing clean air zones is to reduce the air pollution levels and satisfy the legal requirements on keeping specific pollutants below the allowable limits. The pollutants include nitrogen oxides (NO_x), particulate matters (PM_x), carbon dioxide (CO_2), and other greenhouse gases. In particular, breathing air that has high concentrations of NO_2 can create irritation in the human respiratory system, can cause diseases such as asthma, may compromise long function, and increases the risk of respiratory infections (Chen et al., 2007; Keast et al., 2022). Once properly designed and implemented, a clean air zone reduces traffic emissions that are harmful, improves air quality and protects people's health (see the work by Holman et al. (2015) on the performance of such zones in European cities and the work by Liu et al. (2023) for Birmingham). Designing effective clean air zone interventions is essential for achieving an improved air quality, and data will play a central role in such a design.

The methodologies for designing effective clean air zones have the following steps: identifying the geographical area, stakeholder engagement, modelling and impact assessment, choosing measures and incentives, implementation, and finally, monitoring and evaluation.

The core technical aspect of these methodologies is to develop computational models, which is done using physics laws and models of physical processes (e.g., chemical transport models, computational fluid dynamics, and whether forecast models (Jacob, 1999)). Building such computational models involves making appropriate assumptions to simplify complex atmospheric processes and to manage the computational constraints. Given the recent advances in machine learning and artificial intelligence, an alternative approach to building these complex physics-based computational models is to develop models directly from data.

The report by Defra Joint Air Quality Unit (February 2020) provides the general guidelines for the operation of clean air zones in England. It recommends the approach to be taken by local authorities when implementing and operating a clean air zone. An example of this approach is the clean air zone in Greater Manchester (CAGM), which has the goal of improving air quality by encouraging some vehicle owners to upgrade to cleaner vehicles or pay a daily charge. The technical reports published on the website of CAGM [1] clearly show the role of data is designing the related policies and evaluating such policies when they are implemented. In general, datasets play two main roles: (1) datasets are used to select and tune parameters of the physical models developed for air quality, and (2) datasets are used for monitoring and checking if the target of the policies are achieved (e.g., reducing the pollutant level to some value).

Newcastle City Council has been in the process of designing and implementing a clean air zone In the past few years. The zone was launched on January 2023. The availability of large volumes of data from Newcastle collected and stored by Newcastle Urban Observatory, *makes the Newcastle clean air zone a perfect candidate for applying the methodologies developed in this PhD project for data-driven validation of policy objectives*.

1.1.2 Expansion of EV Charging Infrastructure

With the advances in science and technology to better understand and evidence the effects of climate change, individuals and governmental organisations are paying more attention to alternative forms of energy obtained from solar, wind, hydroelectric, and geothermal power (Nehrir et al., 2011). The UK committed in 2019 to a legally binding net zero target by 2050 and introduced new interim targets to reduce emissions by 78% by 2035 (Logan et al., 2022). As a key tenant of the new technologies to reduce carbon emissions, electric vehicles (EVs) are making a rapid sales progress with a yearly sales increase of 20% in 2022 (ZapMap, 2022b). As of the end of February 2024, there are over one million EVs on UK roads. To address the inevitable increasing demand for charging EVs, Department

[1] https://cleanairgm.com/technical-documents

for Business, Energy & Industrial Strategy (BEIS) has committed to a minimum of 2,500 charging points across the strategic road network in the UK (HM Government, 2021). Many studies look at the market, the economy and the environment as entry points for designing the location of charging stations (Cai et al., 2014; LaMonaca and Ryan, 2022; Wang et al., 2016; Yang et al., 2017). However, there are still some deficiency in the study of the actual quantity and types of charging points.

In order to help City Councils and other governmental organisations plan for the EV charging infrastructure, optimising the following factors are needed to design and expand the EV charging infrastructure: charging point type, charging point location, charging point quantity, total capital and operational expenditures, and operating hours of charging points. *This observation makes the optimal expansion of EV charging infrastructure a good candidate for applying the methodologies developed in this PhD thesis on data-driven optimisation for implementation of policy commitments.*

1.2 Aim, Research Questions, and Objectives

Based on the emergence of new technologies and the collection of unprecedented amounts of data from various sources in the transport sector, this PhD project aims to investigate the potential of quantitative methods that can effectively deal with large, diverse, and complex datasets for validating the objectives of proposed policy interventions and implementing policy commitments. Based on the literature review presented in Chapter 2 and the identified research gaps and challenges, this PhD project will focus on the following aim:

"To develop data-driven techniques that can integrate and deal with large diverse complex datasets in transport for validating the objectives of policy interventions and for efficient implementation of policy commitments with Air Quality and EV charging infrastructure as case studies."

In order to achieve the aim of this PhD project, the following research questions are identified:

RQ1. Given the large volume of data gathered from the transport network, what data types are relevant to the objectives of a policy intervention?

RQ2. What machine learning techniques are suitable for combining large datasets, processing the data, and validating the objectives of policy intervention?

RQ3. Could these datasets and machine learning techniques be used for efficient optimal implementation of policy commitments?

These research questions are linked together with respect to the extremely large volume of data and the need for efficient analysis and learning methods. This PhD project will address the challenging task of developing a methodology for using machine learning techniques to validate the objectives of a policy intervention and find optimal implementation of policy commitments.

In order to achieve the aim of this PhD research and answer the above research questions, the following objectives are considered:

O1. Identify, gather, preprocess and analyse data types relevant to a policy from different sources.

O2. Develop suitable machine learning models based on the input processed data and the considered policy objectives and commitments.

O3. Analyse and simulate future scenarios under the implementation of the policy commitments to gain insights on their impact in the transport network.

O4. Study methods for validating the outcome of machine learning methods. Select and use metrics that can best describe the accuracy of the outcome and validate the outcome against domain knowledge.

O5. Determine the potential use of optimisation methods for transport policy commitment implementation.

O6. Apply the designed frameworks to case studies on validating the policy objectives of clean air zone and the expansion of the electric vehicle charging infrastructure, which are critical for achieving the UK's target of net-zero emissions by 2050.

1.3 Contributions and Potential Benefits

The main contributions of this thesis are as follows.

- This thesis demonstrates the use of machine learning methods for validating policy intervention objectives in transport systems. The validation could be part of the initial policy creation, policy refinement, or a post implementation review.
- It proposes a framework for finding data types that are relevant to the intervention objective, and for validating the intervention and checking how well the objectives of the intervention are achieved.

- It contributes to the AI-based support and assistance of implementing policy commitments by capturing the focus of the implementation problem, and building a framework that integrates data with simulation and optimisation.
- It provides new ideas and solutions for intelligent decision-making against multiple conflicting performance criteria. The approach of this thesis creates computational models that can understand statistical data and generate acceptable, intelligent decisions with good robustness.

The obtained results of this PhD research have been presented in the form of **two journal manuscripts** and **seven conference papers** including a journal paper published at **iScience**, a journal paper published at **IEEE Access**, two papers at the **IEEE International Conference on Intelligent Transportation Systems**, two conference papers at the **Universities Transport Study Group (UTSG)**, a conference paper at the **Institution of Engineering and Technology, Powering Net Zero (IET)**, a paper at the **ECTRI Young Researchers conference**, and a paper at the **Annual Electric Vehicle Conference** in Edinburgh Napier.

The study's findings will contribute to the body of knowledge on using data-driven approaches, machine learning techniques, and optimisation methods to improve policy planning and decision-making in the transport sector, while also exploring the possibilities of quantitative modelling impact on validating policy objectives. The insights gained from this study will be valuable to policymakers and stakeholders in the transport sector and will assist in making informed choices to improve transport systems.

The research of this thesis contributes to a more sustainable urban environment by providing valuable insights into effective clean air zone interventions, which can improve air quality and promote sustainable transport solutions. The optimisation approach proposed in this thesis is general and can be applied to any baseline model that can simulate future transport scenarios.

1.4 Scope and Limitations

Based on the research questions, and the performed literature review, a research gap that emerges is the limited research on the application of machine learning and optimisation methods to validate and implement transport policy commitments. While there has been growing interest in using these technologies in transport, there is still limited research on their practical application for validation of policy objectives and its implementation. Therefore, this study aims to fill this gap by exploring the potential use of machine learning and optimisation techniques for validation of policy objectives and its implementation in the transport sector.

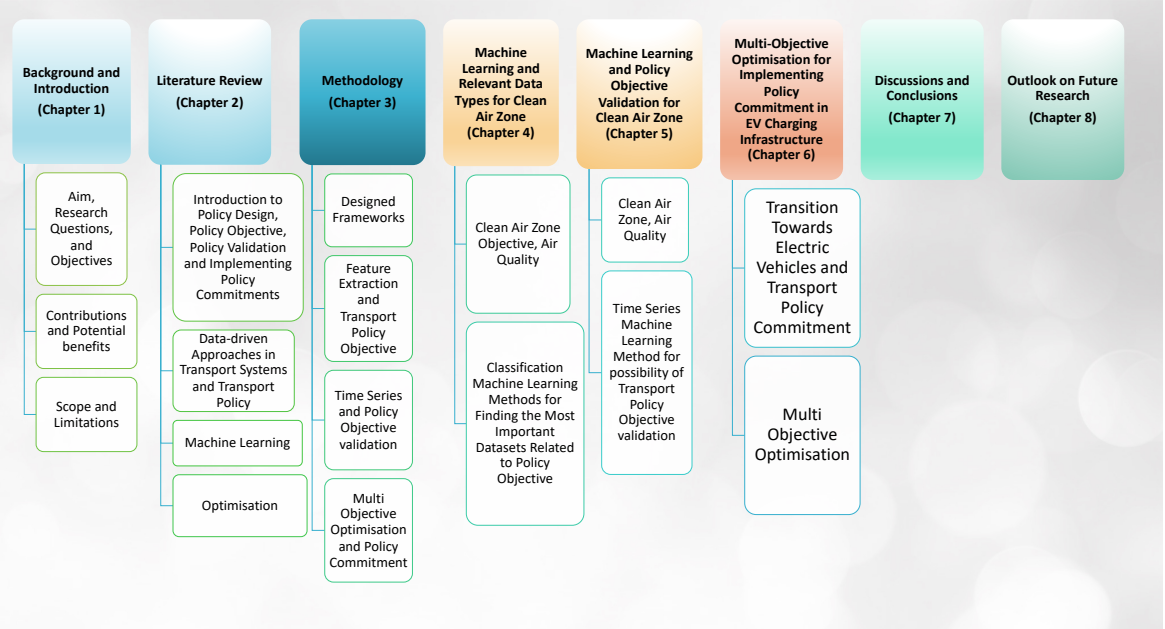

Fig. 1.1 High-level overview of the content of the thesis chapters

Additionally, there is a need to investigate the challenges and opportunities of applying data-driven approaches and quantitative methods to transport policy-making, particularly in the context of large, diverse, and complex datasets. This research aims to provide valuable insights into these research gaps and contribute to the effective implementation of transport policy commitments and decision-making processes.

The scope of this study is limited to the application of machine learning and optimisation methods for validating policy objectives and its implementation in transport, with a focus on air quality improvement and electric vehicle charging infrastructure expansion. While the methodology of this thesis is general and the designed frameworks are applicable to other transport policy interventions, the quantitative results of this study needs to be interpreted within the context of the specific case studies being investigated.

Several limitations may impact the study's results, including the availability and quality of data, and the selection of machine learning and optimisation techniques. To address these limitations, the research will adopt a rigorous methodology for data collection and analysis, carefully selecting appropriate machine learning techniques and optimisation methods after an extensive literature review.

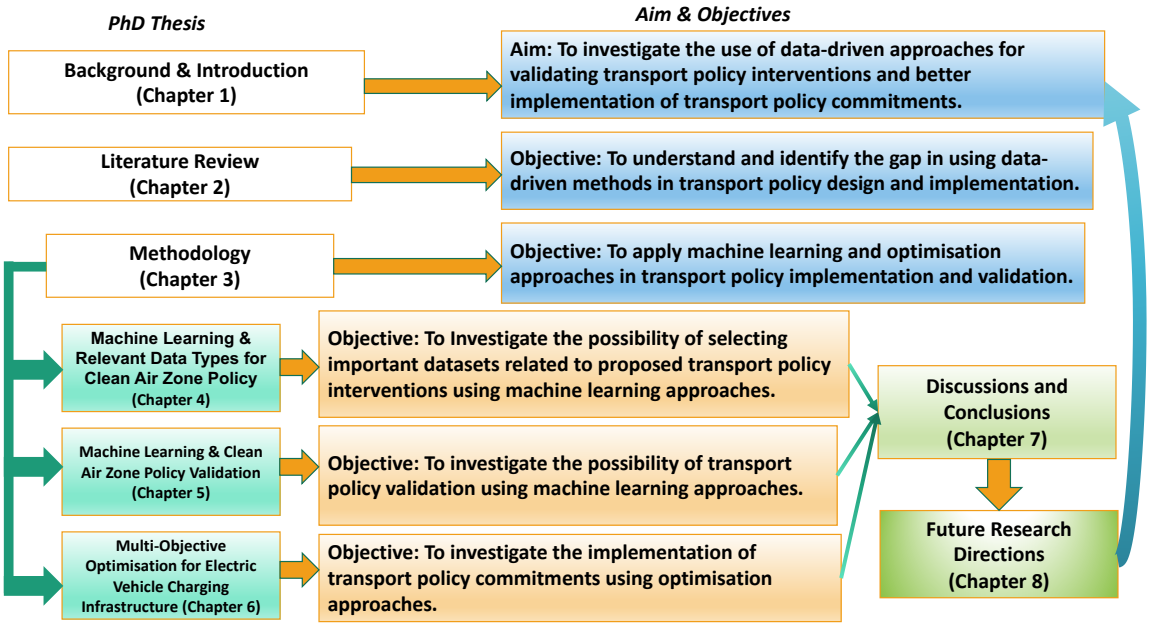

Fig. 1.2 Structure of the thesis

1.5 Thesis Outline

This thesis consists of eight chapters which detail the steps and processes undertaken in order to achieve the aim and objectives of the study described in Section 1.2. The high-level overview of the content of the thesis chapters are provided in Figure 1.1. The connection between different chapters is presented in Figure 1.2. These chapters are:

- **Chapter 1: Introduction** provides an overview of the research background, research questions, scope, limitations, and the structure of the thesis.
- **Chapter 2: Literature Review** presents a review of the existing literature on data types, machine learning techniques, and optimisation methods with focus on their application in transport.
- **Chapter 3: Methodology** describes the research design, the current state of practice in evidence-based policy-making, the impact of data-driven methods on policy-making processes, the proposed data-driven frameworks, and their integration within the cycle of policy design and implementation.
- **Chapter 4: Machine Learning & Relevant Data Types for Clean Air Zone** presents data collection and data analysis methods, the results and analysis of the clean air zone

case study, focusing on finding relevant data types related to the policy objective using machine learning methods.

- **Chapter 5: Machine Learning & Validating the Objective of Clean Air Zone** presents the results and analysis of the clean air zone case study, focusing on validating the policy objectives using machine learning methods.
- **Chapter 6: Optimisation & Electric Vehicle Charging Infrastructure** presents data collection and analysis, and the results for the electric vehicle charging infrastructure expansion, focusing on policy commitment implementation using optimisation methods.
- **Chapter 7: Discussions and Conclusions** provides a synthesis of the research findings, discussing the implications for policy and practice, as well as the study's limitations with respect to technical and data constraints as well as their limitations in being integrated in the policy design cycle.
- **Chapter 8: Future Research Directions** summarises the key findings and contributions of the thesis, and offers recommendations for policymakers, stakeholders, and researchers in the field of transport policy-making.

The above chapters will contribute to the main aim of this thesis, which is to provide a comprehensive understanding of the potential of data-driven approaches, machine learning techniques, and optimisation methods for validating the objectives of policy interventions and implementing policy commitments. The employed approaches include integrating simulation models with optimisation, putting appropriate assumptions for building prediction models, training machine learning models using data, and performing multi-objective optimisations.

1.6 Data and Code Availability

Original datasets used for generating the results of this thesis are publicly available with links listed in Table 1.1. All original code has been deposited at Github and is publicly available with links reported in Table 1.1. The Python code for Chapter 4 is included in Appendix B. The Python code for Chapter 5 is included in Appendix C. The Python code for Chapter 6 is included in Appendix D. While the author has reported in this thesis any information required to reanalyse the data, any further information and requests for resources should be directed to the author at farhadi.farzane@gmail.com.

The first year of the PhD research considered also other case studies such as road pricing schemes. Multiple organisations were contacted to gain access to appropriate datasets. Due

to the barriers and concerns around data sharing encountered extensively in the initial stage of this PhD research, it was decided to continue with clean air zone as the first case study since appropriate datasets was available through Newcastle Urban Observatory. The second case study on EV charging infrastructure was chosen in consultation with the industrial partner of the PhD research, Arup Group Limited, that facilitated accessing a baseline model with publicly available datasets.

Table 1.1 Links to data sources and code developed for generating the results of this thesis

DATA & CODE	SOURCE	IDENTIFIER
Deposited data		
Air Quality Data	Newcastle Urban Observatory	https://newcastle.urbanobservatory.ac.uk/
Car Ownership	UK Department For Transport	https://www.data.gov.uk/dataset/11bc7aaf-ddf6-4133-a91d-84e6f20a663e/national-trip-end-model-ntem
Vehicle Availability	UK Office for National Statistics (Nomis)	https://www.nomisweb.co.uk/home/Search?context=&term=Car+or+van+availability
Power Profile	UK Department For Transport	https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-domestics
Trip Statistics	UK Office for National Statistics (Census)	https://www.ons.gov.uk/census/2011census/2011censusdata/originanddestinationdata
EV Efficiency and EV Battery Size	Bloomberg	https://www.bloomberg.com/professional/datasets/?bbgsum-page=DG-WS-PROF-SOLU-DATACONT&mpam-page=21140&tactic-page=429341
Future Energy Scenarios	National Grid	https://www.nationalgrideso.com/future-energy/future-energy-scenarios
Software and algorithms		
Python 3.10	Python Software Foundation	https://www.python.org
Code for Air Quality results in Chapters 4–5	Github repository	https://github.com/farzanehfar/Air_Quality_MachineLearning
Code for EV charging infrastructure and evaluation in Chapter 6	Github repository	https://github.com/farzanehfar/MultiObjective-Optimization

Chapter 2

Literature Review

This chapter undertakes a thorough examination of the existing literature derived from prior studies, aiming to establish the research direction for this thesis. Employing a top-to-bottom approach, the literature review takes a broader perspective of the subject, gradually refining the discussion to identify the fundamental issues. The exploration starts by sequentially addressing key topics, commencing with a comprehensive overview and progressively delving into more specific areas of the analysis. The sequence of the topics reviewed begins by defining policy and its different stages as described in Section 2.1. The next Section 2.2 describes policy design, implementation, and validations in the transport sector. Section 2.3 reviews the essential role of quantitative methods in transport followed by a review of machine learning methods in Section 2.4 and optimisation methods in Section 2.5. This section provides information on the choice of the case studies for applying the methodologies developed in this thesis together with the limitations of the current research are presented in Section 2.6. Finally, the research gaps are presented in Section 2.7 along with a conclusion of this chapter.

2.1 Policy

The word *policy* is defined in Cambridge dictionary as *"a set of ideas or a plan of what to do in particular situations that has been agreed to officially by a group of people, a business organisation, a government, or a political party"*. The specific forms and types of policies can vary significantly depending on the context and the organisation or government body developing them. Policies could be in the form of laws, regulations, targets, ways of doing things, banning certain behaviours, and incentivising specific actions. Policies could be set from national governments, regional governments, local government, or stakeholders (e.g., in the transport sector being National Highways or Network Rail). A policy could

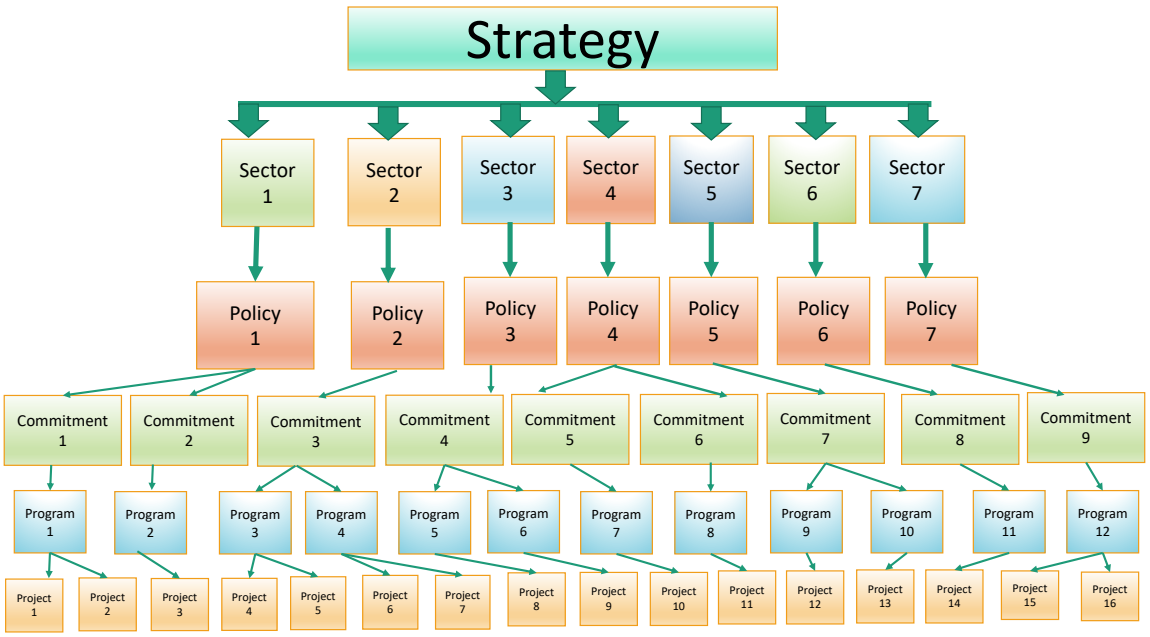

Fig. 2.1 Different stages for achieving an overarching vision or goal starting from setting a strategy down to defining multiple projects

also be delivered by authorities that are different from the one setting the policy. Figure 2.1 sketched based on the works by Palfrey et al. (2012) and Nellthorp and Mackie (2017) shows different stages for achieving an overarching vision or goal by the UK Government, which starts by setting a strategy, having policies for the related sectors, defining policy commitments, and finally having multiple programs and projects to implement those policy commitments by the related authorities and public or private organisations. This thesis will focus on policy in the transport sector. An example strategy set by the UK government is the net zero emission strategy with policies specific to how the transport sector can help deliver on this strategy, which will also be used in this thesis for demonstrating the developed methodologies. Figure 2.2 shows the key transport policy commitment and objective of the Net Zero Emission Strategy based on the UK policy paper on Net Zero Strategy,[1] which will be used for applying the data-driven quantitative research methodology of this thesis.

2.2 Policy in Transport

Transport policy is an important aspect of modern society, as it impacts the movement of people and goods and influences economic development and environmental sustainability

[1] https://www.gov.uk/government/publications/net-zero-strategy

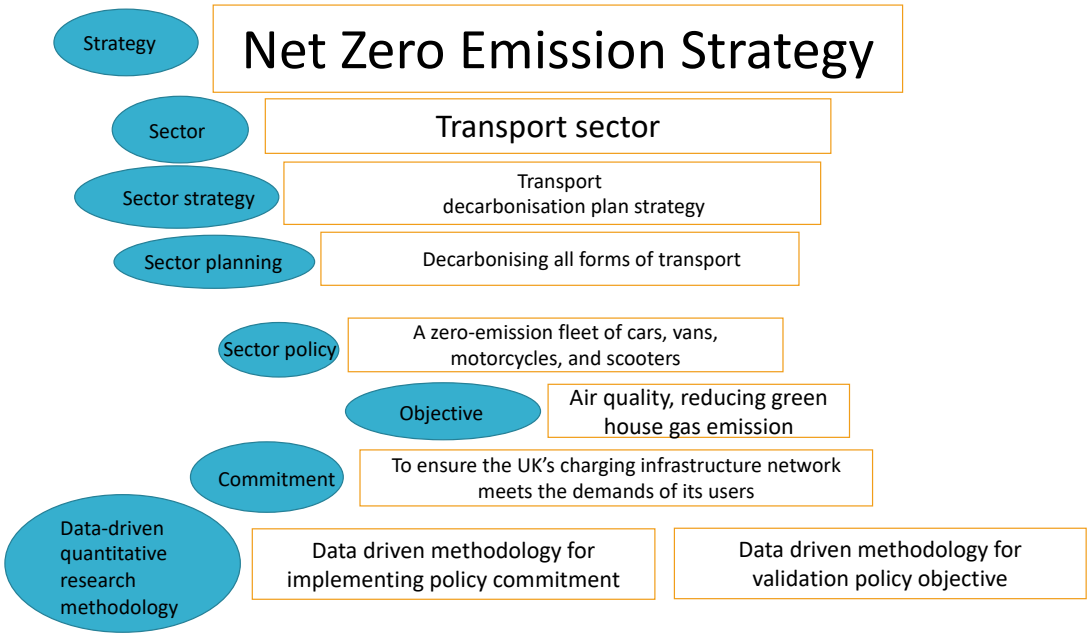

Fig. 2.2 Key transport policy commitment and objective of the Net Zero Emission Strategy for applying the data-driven quantitative research methodology of this thesis

(Marsden and Reardon, 2017). This section will discuss the importance of policy in transport and the challenges facing transport policy design, implementation, and validation.

One of the key reasons why policy in transport is important is that it can help to shape the development of transport systems in ways that align with broader social, economic, and environmental objectives (Rodrigue et al., 2023). For example, policy can promote the use of public transport, walking, and cycling, which can help to reduce carbon emissions, congestion, and air pollution (Poudenx, 2008). In addition, transport policy can support economic development by providing access to employment, education, and healthcare, and by facilitating the movement of goods and services.

The work by Kii et al. (2016) provides a review of the interaction between transportation and spatial development with respect to urban policy and infrastructure planning. The authors review policy objectives, such as safety, efficiency, and sustainability, and the policies that have been implemented to achieve these objectives. The authors conclude that the co-evolution of two approaches of (a) intensive modelling at the local scale and (b) simplified modelling at regional, national, and global scales will contribute to sustainability science to satisfy environmental and energy constraints. The work by Kaufmann et al. (2008) examines the political, social, and economic factors that have influenced the development and implementation of transport policies. The researchers analysed existing literature and policy documents, as well as conducted interviews with key stakeholders involved in the

transport sector. The study found that transport policies have been shaped by a complex interplay of political, social, and economic factors, including urbanisation, environmental concerns, technological developments, and institutional arrangements. The researchers identified key challenges facing transport policy, such as the need to balance competing policy goals and the tensions between national and local governance structures. The work by Rastogi (2011) focuses on transport policy and planning in developing countries. The authors review the challenges facing transport systems in these countries, including rapid urbanisation and increasing motorisation. They conclude that policy solutions such as promoting public transport, improving non-motorised transport, and implementing low-carbon transport options will make non-motorised transportation sustainable both within and across travel modes.

The work by Chapman (2007) provides an overview of transport policies that aim to reduce carbon emissions and mitigate climate change. The authors review policy instruments such as fuel taxes, emissions standards, and public transport subsidies, and their effectiveness in reducing emissions. They conclude that a combination of policy instruments is needed to achieve significant emissions reductions. This observation becomes important when setting an objective for a policy intervention and extracting trends in datasets gathered before and after implementing a clean air zone. The work by Leite de Almeida et al. (2021) discusses the role of transport policy in achieving the Sustainable Development Goals (SDGs). The authors review the links between transport and the SDGs, and the policy measures needed to achieve them. Their results show that these SDGs can be achieved by setting appropriate policy interventions including implementation of zero emission zones, subsidy schemes for the uptake of clean vehicle technology, and the digitalisation of the transport system.

The development and implementation of transport policy can also face significant challenges. One of the main challenges is the complex and fragmented nature of transport systems, which can involve multiple stakeholders and levels of the government (Givoni, 2014). This can make it difficult to coordinate policy and ensure that it is effective and sustainable over the long term. Another challenge is the limited resources available for transport policy, which can constrain the ability of policymakers to invest in new infrastructure, improve services, or address emerging issues such as climate change (Givoni, 2014). Moreover, setting wrong transport policies could create obstacles and unintended consequences including increased traffic congestion, negative environmental impacts, economic inefficiency, social inequalities, Infrastructure Strain, energy Dependence, negative health impacts, safety concerns, and public dissatisfaction (Santos et al., 2010).

Despite the challenges mentioned above, there are several key elements that are critical to effective transport policy. First, policy should be based on a clear understanding of the

objectives and outcomes that policymakers want to achieve (Rodrigue et al., 2023). This should involve consultation with stakeholders during the policy design and the development of robust evidence-based evaluation of the policy implementation using data. Second, transport policy should be integrated with other policy areas, such as land use, environment, and social policy, to ensure that it is aligned with broader social and economic goals. Third, policy should provide have an optimal mix of stability and adaptability, so that it can respond to changing circumstances and emerging challenges such as new technologies or climate change (Browne and Ryan, 2011).

Several factors contribute to the often observed inflexibility in transport policies. This includes limited financial resources, long planning horizons in transport infrastructure projects, institutional inertia that can slow down the adaptation of transport policies, short political cycles that may discourage long-term planning and investment in flexible infrastructure, and lack of comprehensive data and information about emerging technologies (Gifford, 2005).

All the works reviewed above highlight the need for a comprehensive evaluation of policy implementations against their objectives. This will require gathering data through digitalisation of the transport sector, processing the data to extract useful knowledge, building computational models for such evaluations, and use them to improve the policy implementations. These elements form the high-level aim of this PhD research.

2.2.1 Policy Design in Transport

Transport policy design is a critical element in planning at all levels, from local urban frameworks to broader regional and national planning. One of the aims of transport policy design is to provide safe, efficient, and sustainable transport options to the public (Pojani and Stead, 2018). This section will discuss the key principles of transport policy design, provide successful examples, and examine the role of *evidence-based research* in such a design.

One of the key principles of transport policy design is the identification of clear objectives (gov.uk, 2021). Policy designers need to have a clear understanding of what they want to achieve through the policy (Pojani and Stead, 2018). For example, the objective of London's Congestion Charge was to reduce traffic congestion and use the generated revenues to support London's transport system. This objective was clearly stated in the legislation and together with the technology available at the time guided the design of the policy (gov.uk, 2021). The Results of the London Congestion Charging scheme was re-assessed in the work by Givoni (2012) using data-driven methods to build computational models to better understand its effectiveness.

Another important principle of transport policy design is the use of evidence-based research (Davies and Nutley, 2000). Evidence-based research provides policymakers with

objective information about the problem they are trying to address and the most effective solutions. For example, the use of evidence-based research played a critical role in the development of the Vision Zero Safety, which is a multi-national road traffic safety project for achieving zero fatalities in the road traffic.[2] It started in Sweden in 1997 as a vision that galvanised stakeholders to work towards the target and was based on the scientific evidence linking traffic crashes to human behaviour and the built environment (nyc.gov, 2020). Since then, many cities in different countries have implemented measures to move towards this target including Canada, Netherlands, Sweden, United Kingdom, and the United States. Successful design and implementation of the related projects have been based on data-driven solutions linked with the aim of this PhD research.

Policy design also requires a consideration of available resources (Hatzopoulou and Miller, 2009). Policymakers need to understand the costs and benefits of different policy options and the resources required to implement them. The introduction of electric buses in Shenzhen, China, is an example of a policy design that considered available resources (Mahmoud et al., 2016). The policy was designed to address air pollution in the city and was based on the availability of renewable energy sources and the affordability of electric buses.

Successful transport policy designs often involve collaboration between different stakeholders (Fischer, 2004). For example, the development of the Copenhagen Bicycle Strategy involved input from a range of stakeholders, including cyclists, business groups, and city officials (Gössling, 2013). This collaboration ensured that the policy design addressed the concerns and needs of all stakeholders, which results in a policy is a form of compromise between the different interested parties.

The works reviewed above show that transport policy design is a complex process that requires a clear understanding of the problem, the identification of clear objectives, the use of evidence-based research, consideration of available resources, and collaboration between stakeholders. Evidence-based research plays a critical role in transport policy design by providing policymakers with objective information about the problem and potential solutions. Moreover, the limitations of the works reviewed above together with the technological advances in machine learning, artificial intelligence, and big data analytics show that more work is needed to integrate data-driven approaches in assessing the policy design and refining its implementation using the additional knowledge gained through data.

[2] https://en.wikipedia.org/wiki/Vision_Zero

2.2.2 Validating Policy Objectives in Transport

Validating transport policy objectives is an essential step in the policy-making process to ensure that policies will meet their intended objectives and deliver the desired outcomes. The validation could be part of the initial policy creation, policy refinement, or a post implementation review. It also needs to understand the art of the possible of the science, engineering and technology that may be required to introduce the policy effectively. Transport policies can have significant impacts on social, environmental, and economic factors, and therefore require careful consideration and assessment to avoid unintended consequences (Jonsson, 2008). Validating the policy objectives involves a systematic evaluation of the proposed policy against the intended goals and objectives both at the design stage of the policy and when the policy is introduced. It is a critical step in the policy-making process as it allows policymakers to identify any potential gaps, weaknesses, or unintended consequences of the policy before its implementation (Jonsson, 2008). Validation policy objectives can help to improve the effectiveness and efficiency of policies by identifying areas for improvement and suggesting modifications to ensure that policies meet their intended outcomes.

There are several reasons why validating policy objectives is crucial in transport. First, transport policies can have far-reaching effects on social, economic, and environmental factors. For example, a policy to promote cycling may have positive impacts on public health and air quality but may also lead to increased traffic congestion and reduced road capacity (Commission, 2014). Therefore, it is important to assess the potential impacts of the policy holistically and identify any unintended consequences.

Second, transport policies often involve significant financial and resource investments. Validating policy objectives can help policymakers to ensure that resources are allocated effectively and efficiently to achieve the intended outcomes (Halim et al., 2018). By identifying any potential gaps or weaknesses in the policy, policymakers can modify the policy to ensure that resources are used optimally.

Third, transport policies are often complex and require multiple stakeholders' involvement. Validating policy objectives can help to ensure that all stakeholders' interests are taken into account and addressed adequately in the policy (Le Pira et al., 2017). This can help to build stakeholder support and promote the policy's implementation and success.

An example of validating policy objectives in transport is the European Union's policy on alternative fuel infrastructure (Policy, 2017). In 2014, the European Union adopted a policy to promote the deployment of alternative fuel infrastructure, including electric vehicle charging points, hydrogen refuelling stations, and natural gas filling stations, across Europe. The policy aimed to reduce the dependency on fossil fuels and promote the use of low-emission vehicles (Policy, 2017). To validate the policy objectives, the European

Commission conducted a comprehensive impact assessment, including economic, social, and environmental impacts. The impact assessment identified the potential benefits of the policy, such as reducing greenhouse gas emissions and improving air quality, but also highlighted potential challenges, such as high infrastructure costs and regulatory barriers. The findings of the published report (Policy, 2017) highlights the need for secure and flexible data handling, producing real-time data to encourage rational use of the services, and collecting coordinated data through monitoring activities.

Therefore, Section 2.3 will discuss the details of quantitative methods in transport.

2.3 Quantitative Methods in Transport

With the increasing complexity and demands of transport systems, *modelling* plays a critical role in enhancing their efficiency and sustainability (Bates, 2007; Iacono et al., 2008). Transport modelling can be classified into two main categories: quantitative and qualitative modelling (Queirós et al., 2017). While both approaches have their advantages and disadvantages, their effective use depends on the specific transport problem at hand. Quantitative modelling involves the use of mathematical or statistical techniques to analyse transport data and make predictions about future transport patterns (Philips, 2023). This approach uses numerical data and statistical models to quantify relationships between variables and predict the outcomes of transport decisions. Commonly used quantitative modelling techniques in transport include regression analysis, optimisation models, and simulation models (Iacono et al., 2008). The main advantage of quantitative modelling is its ability to provide precise and quantitative estimates of transport outcomes (Queirós et al., 2017). This approach is particularly useful in analysing large and complex transport systems that involve multiple variables and factors. For instance, optimisation models can be used to determine the optimal route for a fleet of trucks, minimising the total travel time and fuel consumption (Pečenỳ et al., 2020). Similarly, simulation models can be used to predict the impact of a new transport policy on traffic flow and congestion (Chao et al., 2020).

However, quantitative modelling also has its limitations. The major challenge is the availability and accuracy of data (Milne and Watling, 2019). transport data is often complex and difficult to collect, requiring specialised tools and techniques (Milne and Watling, 2019). Furthermore, quantitative models are based on assumptions and simplifications, which may not always reflect the real-world complexities of transport systems (Jiang et al., 2022).

There are several types of quantitative modelling techniques that are commonly used in transport research, which are reviewed next. Features, advantages, disadvantages, and limitation of each of the techniques are summarised in Table 2.1.

Table 2.1 Overview of quantitative modelling techniques

Quantitative Modelling Technique	**Core Features**	**Advantages**	**Disadvantages**	**Limitations**
Regression Analysis (Yang, 2015) (Bates, 2007)	Uses data to find out how a change in one thing is linked to a change in another (Estimates relationships between variables).	Good for making predictions and finding trends.	May not capture non-linear relationships	Needs large, good-quality data; can be misled by unusual data points.
Optimisation Models (Mamoun et al., 2021), (Hensher and Button, 2007), (Chen et al., 2016)	Searches for the best way to do something, like minimising costs or travel time.	Finds the best option among the available many options to achieve something.	Can be hard to solve computationally; requiring large computational resources; may not find the absolute best solution.	Sometimes oversimplifies real problems; needs good data to work well.
Simulation Models (Jacyna et al., 2014), (Jacyna et al., 2014), (Möller, 2014)	Creates a model that acts like a real system to see what might happen.	Can try out different scenarios safely.	Can be complex and need large computer power.	Relies on the model being a good copy of the real world.
Discrete Choice Models (Bierlaire, 1998), (Brownstone, 2001), (Möller, 2014), (Tiwari et al., 2003)	Looks at how people make choices from a set of options, like which transport to use.	Helps understand why people prefer one option over another.	Assumes people always make logical choices.	Relies on detailed information about why people make the choices they do.
Machine Learning Methods (Tizghadam et al., 2019), (Tizghadam et al., 2019), (Polson and Sokolov, 2017), (Omrani, 2015), (Evans et al., 2019)	Algorithms that improve from experience and can find patterns in data.	Can deal with complex and big data.	Needs a lot of data and can be hard to understand why it made a certain prediction.	Success depends on the quality of data; can make mistakes if the data is not representative.

- Regression Analysis: Regression analysis is a statistical technique that is used to estimate the relationship between two or more variables (Yang, 2015). In transport research, regression analysis is often used to predict travel behaviour. The book by Bates (2007) provides a historical account of demand modelling in transport, where regression analysis plays a role in predicting different modes of travel behaviour.

- Optimisation Models: Optimisation models are used to find the optimal solution to a transport problem (Mamoun et al., 2021). In transport research, optimisation models are commonly used to optimise route planning and scheduling, vehicle routing, and supply chain management. The book by Hensher and Button (2007) provides different paradigms in fundamentals of transport modelling and describes how optimisation models can be used to improve the use of transport network. Optimisation models are also used recently in a study by Chen et al. (2016) to optimise the route of a fleet of electric vehicles with respect to the availability of charging stations and travel time.

- Simulation Models: Simulation models are used to simulate the behaviour of transport systems and predict transport outcomes (Jacyna et al., 2014). In transport research, simulation models are commonly used to predict traffic flow and congestion, evaluate the impact of transport policies and strategies and simulate emergency response scenarios (Jacyna et al., 2014). The book by Möller (2014) provides an overview of simulation tools in transport using research-oriented use cases in transport sector.

- Discrete Choice Models: Discrete choice models are used to predict travel behaviour and demand by modelling the choices that travellers make (Bierlaire, 1998). In transport research, discrete choice models are commonly used to predict mode choice, route choice, and departure time choice (Brownstone, 2001). The book by Möller (2014) shows how discrete choice models can be integrated as part of simulation models in various transport use cases. A study by Tiwari et al. (2003) have used a discrete choice model to predict the mode choice of commuters.

- Machine Learning Methods: Machine learning methods are increasingly being used in transport research to analyse and predict travel behaviour. They can be used to develop predictive models for simulating the behaviour of transport systems and predicting transport outcomes (Tizghadam et al., 2019). These methods are used to identify patterns and relationships in large datasets. Machine learning techniques are used to develop traffic flow simulation models that can predict traffic flow and congestion (Polson and Sokolov, 2017; Tizghadam et al., 2019). The paper by Omrani (2015) presents a study that uses machine learning methods to predict the travel mode choice

of individuals. The study aims to identify the factors that influence travel mode choice and to develop a model that accurately predicts the mode of travel for individuals based on these factors. The work by Evans et al. (2019) uses machine learning methods to forecast road traffic conditions while incorporating more accurately the contexts such as public holidays and sporting events. The details of machine learning techniques will be reviewed in the next section 2.4.

Depending on the specific research question and transport problem at hand, different modelling techniques may be more appropriate than others. Here are the advantages and disadvantages of quantitative models in transport research, along with application examples. The key advantages are:

- Cost-effectiveness: quantitative models can be used to evaluate the performance of transport systems cost-effectively, without the need for expensive field experiments (Rinaldi et al., 2022). For example, a study by Mishra et al. (2010) used a simulation model to evaluate the impact of different bus rapid transit designs on passenger travel time and waiting time.

- Flexibility: quantitative models can be easily modified to incorporate different scenarios and parameters, allowing researchers to evaluate the performance of transport systems under a variety of conditions (Zaied, 2008). For example, a study by Cao et al. (2019) used a simulation model to evaluate the impact of different parking policies on traffic congestion in a downtown area.

- Safety: quantitative models can be used to evaluate the safety of transport systems and identify potential hazards and risks (Shyur, 2008). For example, a study by Cunto and Saccomanno (2008) used a simulation model to evaluate the safety of a signalised intersection under different traffic conditions.

The key disadvantages are:

- Assumptions: quantitative models are based on a number of assumptions about the behaviour of transport systems (Keith et al., 2020). The accuracy of these assumptions can affect the accuracy of the simulation results. For example, a study by Bonsall et al. (2005)) noted that the accuracy of a simulation model for traffic flow depends on the assumptions made about driver's behaviour.

- Data Requirements: quantitative models require large amounts of data, and the accuracy and representativeness of this data can affect the accuracy of the results (Sargent, 2010). For example, a study by Moosavi et al. (2020) noted that the accuracy of a simulation model for public transit depends on the quality of the transit network data.

- Complexity: quantitative models can be complex and difficult to understand. The accuracy of the quantitative results can depend on the skill and experience of the modeller (Sargent, 2010). For example, a study by Alghamdi et al. (2022) noted that the accuracy of a simulation model for signalised intersections depends on the skill of the modeller in selecting appropriate model parameters.

Types of Data in Transport. Data available from the transport system can be used to improve the accuracy of the quantitative models described above, thus plays a central role in transport policy design, implementation, and evaluation. Relevant types of data in the transport sector include:

- Quantitative data: Numerical information, such as travel demand, vehicle counts, emissions levels, and performance data can be collected through surveys, sensors, or administrative records (Cheng, 2022).
- Qualitative data: Descriptive information, such as stakeholder opinions, user experiences, and institutional arrangements, can be gathered through interviews, focus groups, or document analysis (Fossey et al., 2002).
- Geospatial data: Geographic information, such as spatial distribution of transport infrastructure, land use patterns, and accessibility levels, can be analysed using Geographic Information Systems (Breunig et al., 2020).
- Big data and IoT: Large and complex datasets generated by various sources, such as social media, mobile devices, and Internet of Things (IoT) sensors, can provide valuable insights for transport policy-making and performance monitoring (Hajjaji et al., 2021).

As mentioned in the background Section 1.1, there are barriers in the effective use of data in transport including incomplete and inaccurate data, data privacy and security concerns, lack of data format standardisations, fragmented data sources, regulatory barriers, and technological and financial limitations (Catapult Transport Systems, 2017). The next section provides the details of machine learning methods and how they can address a subset of these barriers to develop accurate computational and quantitative models.

From the quantitative modelling techniques discussed above, discrete choice models will not be considered for answering the research questions of this thesis. This is mainly due to the limitation of discrete choice models focusing on the individualistic behaviours of people than their aggregate effects on the whole transport system. Therefore, machine learning and

optimisation models will be used for addressing the research questions while considering the simulation models as a building block in the designed frameworks. These are reviewed next.

2.4 Machine Learning Methods

Machine learning and optimisation methods are chosen among the available quantitative modelling techniques described in the previous section to address the research questions of this thesis on validation of the objectives of policy interventions and efficient implementation of policy commitments (cf. Section 1.2). In this section, the aspects of machine learning methods required in the rest of the thesis are reviewed.

Machine learning methods have increasingly been applied in the transport domain to analyse large, diverse, and complex datasets for various purposes, including a few applications in transport policy (Löfgren and Webster, 2020). These methods can help in identifying patterns and trends, predicting future outcomes, and evaluating the potential impacts of policy interventions on transport systems.

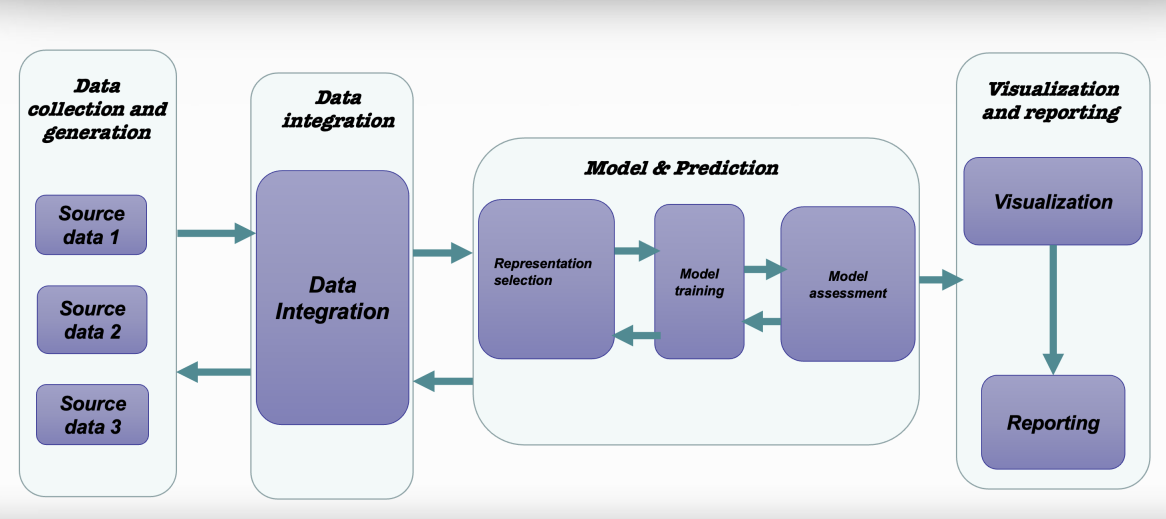

Fig. 2.3 The overview of steps in using machine learning for performing a task (Goodfellow et al., 2016)

Figure 2.3 shows an overview of the steps in using machine learning methods for performing a task (Goodfellow et al., 2016). First data from multiple different sources are collected. These datasets could be in different formats and convey different information. The next step is to integrate these different datasets and preprocess them to make them suitable for the next stage. Then an appropriate model is selected and is trained over the data (Nguyen et al., 2018). The quality of the model is assessed and the model is revised if needed. The model is

also used for making predictions. Finally, the outcome of the model is visualised and reports are created for making decisions (Xyntarakis and Antoniou, 2019).

As it can be seen in Figure 2.3, there are lots of feedback arrows which shows how different steps provide inputs and feedback to each other. For example, if the quality of the data is not good enough, more data needs to be generated. If the quality of the model is not good, a different model needs to be selected or the datasets should be improved (Nguyen et al., 2018). If the quality of the visualised outcome is not good, then the visualisation technique may need to be changed depending on the selected model (Xyntarakis and Antoniou, 2019).

The next section reviews the commonly used machine learning techniques in transport and validating the objectives of transport policy interventions, along with the relevant literature. These techniques can be classified into supervised learning, unsupervised learning, and reinforcement learning (Goodfellow et al., 2016). A summary of these techniques are provided in Table 2.2 and explained in the next subsections.

2.4.1 Supervised Learning

Supervised learning is a widely used machine learning approach in transport, where the model is trained using labelled datasets, and the learning process is guided by a known target variable (Cunningham et al., 2008). Two main types of supervised learning techniques are regression and classification.

Regression: Regression techniques are used to model the relationship between a quantity that is considered as a dependent variable and other quantities in the dataset considered as independent variables. This makes them suitable for predicting continuous outcomes (Nasteski, 2017). Examples of regression techniques in transport include predicting traffic flow (Dell'Acqua et al., 2015), estimating vehicle emission values (Kim and Lee, 2010), and forecasting travel demand (Zhang et al., 2019b).

Classification: Classification techniques are employed to categorise data into distinct classes based on their features, making them appropriate for predicting categorical outcomes (Kesavaraj and Sukumaran, 2013). In transport, classification methods have been used for tasks such as identifying high-risk crash zones, detecting traffic congestion levels, and predicting public transport ridership (Ramli and Mohamed Rawi, 2020).

Beyond regression and classification, supervised learning methods also include:

- **Time Series Forecasting**: Time series forecasting involves predicting future values based on historical data. This class of methods is particularly useful for temporal data where the goal is to make predictions over time.

Table 2.2 Machine learning methods and their applications in transport

Machine Learning Method	Features	Advantages	Disadvantages	Applications in Transport
Supervised Learning (regression)	Trained with labelled data; models guided by a known target variable.	Being accurate on known tasks.	Requires large labelled datasets; may not generalise well.	Traffic flow prediction, vehicle emissions estimation, travel demand forecasting (Dell'Acqua et al., 2015; Kim and Lee, 2010; Zhang et al., 2019b).
Supervised learning (classification)	Utilises labelled data to categorise into predefined classes; output is a discrete label.	High accuracy in prediction for known categories; effective in risk assessment and decision-making.	Requires substantial labelled datasets; less effective with unseen data or overlapping class boundaries.	Identifying high-risk crash zones, detecting traffic congestion levels, predicting public transport ridership (Ramli and Mohamed Rawi, 2020).
Unsupervised learning (clustering)	Finds patterns or structures without labelled data; no predefined target variable.	Discovers hidden structures in data; useful for segmenting data into meaningful groups.	May find irrelevant patterns; less control over outcomes; challenging to validate.	Detecting traffic patterns, clustering similar transport areas, identifying high-risk crash zones, understanding urban mobility patterns (Hong et al., 2021; Kim, 2019; Münz et al., 2007).
Reinforcement learning	Learns through trial and error to achieve a defined goal in a dynamic environment.	Adapts to changing environments; optimises performance through rewards.	Requires a well-defined reward system and an interactive dynamic environment for collecting data.	Optimising traffic signal timings, managing public transport systems, and infrastructure planning investment in a simulation environment (Balaji et al., 2010; Box and Waterson, 2013; Garcia-Flores et al., 2017; Nama et al., 2021).

- **Ranking**: Ranking algorithms are designed to order items in a specific sequence based on their relevance or importance. This is commonly used in search engines, recommendation systems, and information retrieval (Chauhan et al., 2015).
- **Sequence Prediction**: Sequence prediction focuses on predicting the next item in a sequence of events, often used in text generation, speech recognition, and bioinformatics (Welleck et al., 2019).
- **Metric Learning**: Metric learning aims to learn a distance metric that reflects the similarity between data points. It is commonly used in tasks like image retrieval, face verification, and clustering (Hoi et al., 2010).
- **Structured Prediction**: Structured prediction involves predicting complex outputs that are structured, such as sequences, trees, or graphs. This class of methods is used in tasks like natural language processing, parsing, and object detection (Dev et al., 2021).

From the above list, only Time Series Forecasting is applicable to the dataset available for this research (which is the time series from Newcastle Urban Observatory) and will be discussed in details in Section 2.4.6 together with classification and regression methods. The other methods require datasets that include images, textual information, or online webpage data.

2.4.2 Unsupervised Learning

Unsupervised learning techniques involve identifying patterns and structures in datasets without the use of labelled data or known target variables (Hastie et al., 2009).

The main form of unsupervised learning in transport is *clustering algorithms*, which group similar data points together without requiring labelled data. This has been useful for identifying patterns and trends in transport data (Madhulatha, 2012), and has been applied to tasks such as detecting traffic patterns (Münz et al., 2007), analysing travel behaviour (Hong et al., 2021), and identifying areas with similar transport characteristics (Kim, 2019).

2.4.3 Reinforcement Learning

Reinforcement learning techniques involve learning optimal actions based on trial-and-error interactions with a dynamic environment (Kaelbling et al., 1996; Kazemi and Soudjani, 2020; Kazemi et al., 2022, 2024; Sutton and Barto, 2018). In transport, reinforcement learning has been employed for tasks such as optimising railway networks (Subramanian et al., 2023) and traffic signal timings (Balaji et al., 2010), building an automated signalised junction controller (Box and Waterson, 2013), managing public transport systems (Nama et al., 2021),

and planning infrastructure investments (Garcia-Flores et al., 2017). Reinforcement learning algorithm has been applied by Arel et al. (2010) to optimise traffic signal timings in a simulated environment, demonstrating the potential of these techniques for improving traffic flow and reducing congestion.

2.4.4 Machine Learning Methods in Validation of Transport Policy Objectives

The works on the use of machine learning methods for validating the objectives of policy interventions are very limited. There is only a few recent work that study the potential of machine learning methods for answering such a research question:

- The research by Elfar et al. (2018) has employed a supervised learning algorithm to predict real-time traffic congestion, which can be useful for evaluating the effectiveness of congestion management policies.
- The work by Damsere-Derry et al. (2019) has applied regression analysis to evaluate the effectiveness of traffic calming measures on road safety.
- Javanmard and Ghaderi (2022) have used classification techniques to estimate the impact of transport policies on greenhouse gas emissions.

Based on reading these recent preliminary works, it is identified that from the machine learning methods described in the previous subsections, reinforcement learning and unsupervised learning are not suitable for validating the objectives of policy interventions. This is mainly due to (a) the inherent limitation of reinforcement learning that requires a well-defined reward system and an accurate interactive dynamic environment which is not available with the current technological developments in the transport sector; and (b) the limitations of clustering techniques that cannot handle large datasets with its outcome being subjective to interpretation and is difficult to judge the quality of the outcome of the clustering. The research questions of this thesis on validating the objectives of policy interventions are best being studied by regression and classification methods from supervised learning depending on the nature of measured quantities. In the next subsection 2.4.5, techniques for dealing with large datasets are described. The choice of models for regression and classification will be studied in Subsection 2.4.6.

2.4.5 Methods for Dealing with Large Datasets

There are two methods to deal with large volume of data, which are *feature selection* and *dimentionality reduction* (Goodfellow et al., 2016; Hooker et al., 2018). The purpose, considerations, and the preferred use cases of these methods are summarised in Table 2.3. Feature selection is performed via feature importance, which is a technique used in machine learning to identify which features (or quantitative variables) have the most influence on the outcome of a model (Zien et al., 2009). The goal of feature importance analysis is to determine which features contribute the most to the accuracy or performance of a model, and which features can be safely ignored or removed without significantly affecting the model's accuracy. There are two ways to measure feature importance, depending on the type of the model and the specific goals of the analysis (Zien et al., 2009):

- Permutation feature importance: The permutation feature importance was introduced by Breiman (2001) for random forest models and was extended by Fisher et al. (2019) to other machine learning models. In this method, the importance of a feature is computed by first training the model on the train dataset, then permuting the feature and computing the increase in the prediction error of the model. If a feature is important in making predictions, the prediction error should increase after permutation. If a feature is unimportant, the change in the prediction error will be negligible by doing the permutation. Features with the largest decrease in performance are considered the most important.

- Correlation-based methods: Features that are highly correlated with the outcome variable are often considered important, as they are likely to have a strong relationship with the target variable (Blessie and Karthikeyan, 2012).

Dimensionality reduction techniques aim to reduce the number of features in a dataset while preserving its essential structure, which can be valuable for visualising high-dimensional data and improving the efficiency of machine learning algorithms (Sorzano et al., 2014). In transport, dimensionality reduction methods have been used for tasks such as visualising traffic patterns (Li et al., 2017) and simplifying travel demand models (Vovsha et al., 2002).

For the objective of this thesis on validating the objectives of policy interventions, feature selection will be used instead of dimentionality reduction due to the fact that dimentionality reduction can potentially creates new variables as nonlinear functions of quantitative variables in transport. This makes it difficult to assess and interpret the outcomes.

Table 2.3 Methods for dealing with large datasets in machine learning

Method	Purpose	Considerations	Preferred Use Case
Feature Selection (Goodfellow et al., 2016), (Breiman, 2001), (Fisher et al., 2019), (Zien et al., 2009), (Blessie and Karthikeyan, 2012)	To identify which features significantly influence the outcome of a model.	Relies on measuring feature importance, such as permutation feature importance and correlation-based methods.	Used when model interpretability and retaining original features are important.
Dimensionality Reduction (Sorzano et al., 2014), (Li et al., 2017), (Vovsha et al., 2002)	To reduce the number of features while preserving the dataset's structure.	Can create new variables that are nonlinear functions of the original features, which may complicate interpretation.	Used for visualising high-dimensional data and improving algorithm efficiency, but less suitable when interpretability is crucial.

2.4.6 Machine Learning Models for Supervised Learning

As discussed in Subsection 2.4.4, validating the objectives of policy interventions using data is suitable to be studied by regression and classification methods from supervised learning than unsupervised or reinforcement learning. Different supervised learning algorithms may exhibit varying levels of robustness to noise, outliers, and changes in data distribution. Moreover, the characteristics of the dataset, such as size, dimensionality, and noise, can impact the performance of the learning algorithms (Goodfellow et al., 2016). Due to the reasons mentioned above, it is common practice that researchers in machine learning apply multiple learning algorithms to (a) determine which ones are more suitable for the specific data properties at hand; (b) understand the robustness and generalisability of models across diverse datasets; and (c) compare their accuracy to assess their performance in terms of predictive accuracy and efficiency (Goodfellow et al., 2016). These factors become more critical, especially for transport datasets used in the case studies of this thesis, which are noisy with missing or wrong data points.

For validating the objective of the policy intervention studied in Chapter 5, the following classification methods are reviewed: Decision Tree (DT), k-Nearest Neighbors (KNN), Gradient Boosted Decision Trees (GBDT), and Light Gradient Boosting Machine (LGBM) (Goodfellow et al., 2016). A summary of these supervised learning models is presented in Table 2.4 and is explained next.

- Decision Tree (DT) classifiers were initially developed by Messenger and Mandell (1972). Since then, it has been used extensively as one of the most powerful classifiers. Recent use of DT classifiers in transport systems includes the work by Sekhar et al. (2016) that predicts the mode choice behaviour of commuters and by Chen and Wang (2009) for automatic freeway incident detection. DT has a tree structure. At each node of the tree, the data is compared with a constant and depending on the direction of the comparison, one child node is selected. The leaf nodes of the tree holds the class labels. DT classifier assumes the labels to be a function of features. It tries to sequentially divide the space of features into two parts using comparison with a constant until the right label is identified. The depth of the tree is a hyper parameter that shows the required number of comparisons needed to assign a label to a data point.

- Light Gradient Boosting Machine (LGBM) classifier (Ke et al., 2017) is designed to be efficient and more effective for handling big data (large number of features and data instances). The general idea of LGBM is to speed up learning and reduce the computational complexity by focusing on the training examples that result in a larger gradient. LGBM is used in transport applications, e.g., by Niyogisubizo et al. (2023) for predicting traffic crash severity.

- K-Nearest Neighbours (KNN) classifier was developed first by Fix and Hodges (1989). It assumes that data points that are near to each other in the feature space have the same label. The KNN algorithm predicts the label of a given data point as follows: it computes the distance of all the points from the current data point; it then sorts the computed distances from smallest to largest; it picks the first K smallest distances; it finally gets the labels of the selected K data points associated with those distances and returns the label with the highest repetition. The KNN classifier has been used extensively in transport applications including the work by Sun et al. (2018) for short-term traffic forecasting and by Wang et al. (2020) for imputation of missing traffic data.

- Gradient Boosted Decision Tree (GBDT) combines multiple machine learning models (as weak learners) into a single machine learning model (as a strong learner) in an iterative fashion (Elith et al., 2008). GBDT uses regression decision trees as weak models (a numerical value is assigned to each region of the feature space, which is the average of training data in that region). The loss function is also the log-loss function, which is then passed to a sigmoid function to find the predicted label. GBDT is used in transport applications for example by Gallo et al. (2022) to predict the occupancy of public transport vehicles.

From the above four classifiers, GBDT and LGBM are developed recently and can have complex structures with larger number of hyper-parameters. The training of GBDT and LGBM could require more computational resources. DT and KNN are relatively simple and can be easily implemented. The KNN algorithm becomes slower for increasing data sizes (number of data points and features).

Deep learning models have also been applied to variety of tasks in transport, including traffic flow prediction (Lv et al., 2014; Zhang et al., 2019a), travel mode detected (Nam et al., 2017), and traffic incident classification (Li et al., 2022) due to their ability to learn hierarchical representations from data (LeCun et al., 2015). For the case studies of this thesis the above classification models are selected instead of deep learning methods due to the following reasons (Cai et al., 2018; Goodfellow et al., 2016):

- **Data Characteristics**: The datasets used in this research are relatively structured and tabular, with a moderate number of features and records. DT, GBDT, and LGBM are well-suited for such data types, often outperforming deep learning models in scenarios where the data is structured and does not have a vast number of features or records that typically benefit from deep learning architectures. KNN also excels in cases where the decision boundaries are non-linear, and the feature space is well-defined.

- **Performance on Medium-Sized Datasets**: Deep learning models tend to perform best when there is a large volume of data for training. However, for medium-sized datasets, which were the focus of this research, DT, GBDT, and LGBM showed to achieve high performance without the need for extensive tuning and large datasets that deep learning models require.

- **Training Efficiency and Resource Constraints**: DT, GBDT, LGBM, and KNN are generally more computationally efficient and less resource-intensive compared to deep learning models, which often require extensive computational resources, large datasets, and longer training times. In the context of this research, where efficiency and the ability to quickly iterate on models were important, these methods provided a good balance between performance and computational cost.

- **Model Interpretability**: A key objective of this research was to maintain a high degree of model interpretability. Dt, GBDT, and LGBM provide clear insights into feature importance and decision-making processes, making them easier to understand than the often complex and hard-to-interpret deep learning models.

Although examples of using the above supervised learning models in data-driven analysis of transport systems were mentioned, there is no prior work on using these models for

validating the objectives of transport policy interventions. For instance, searching the term "validating policy objective" and "decision tree" and "transport" in Google Scholar reveals no relevant result apart from the work published during this PhD research.

Making Predictions on Time Series Data. The datasets gathered and stored from transport systems often come in the form of time series, which are quantities that change over time. Examples of these datasets are traffic flow, travel time, time-stamped weather data, occurrence of accidents, and air quality monitoring data. The datasets available for the case studies of this research are also in the form of time-stamped data (cf. Section 4.2). Specific supervised learning models have been developed to make predictions on time series data (Bontempi et al., 2013).

Recurrent neural networks (RNNs) are a special class of neural networks with a particular structure designed for processing sequential data (Goodfellow et al., 2016). The basic idea of RNN is to introduce hidden states h_t that can encode some form of memory for capturing the essential information from previous data points in the sequence. The relation between hidden states h_t, features x_t, and labels y_t can be summarised with the equations

$$h_t = f(x_t, h_{t-1}; b_h)$$
$$y_t = g(h_t; b_y),$$

where at each time point, the hidden state h_t is a function of current features x_t and previous hidden states h_{t-1}. The labels y_t are also functions of the hidden states h_t. The parameters b_h, b_y are learned from the data. The functions f, g are usually in the form of neurons that take the weighted sum of their inputs together with some appropriate activation functions.

Long Short-Term Memory (LSTM) networks are a type of RNNs for learning order dependence in sequence prediction problems (Gers et al., 2002). The basic idea behind an LSTM is to introduce a "memory cell" that can store information over long periods of time (Van Houdt et al., 2020). The memory cell is controlled by three "gates" that regulate the flow of information in and out of the cell: the input gate, the forget gate, and the output gate. The LSTM cell can be broken down into four main components:

1. Input gate: This gate controls the input to the memory cell. It takes the input vector and the previously hidden state vector as inputs and produces a vector of values between 0 and 1 that represent the amount of each input to let into the memory cell (Staudemeyer and Morris, 2019).

2. Forget gate: This gate controls the retention of information in the memory cell. It takes the input vector and the previously hidden state vector as inputs and produces a vector

Table 2.4 Machine learning models: features, parameters, advantages, and applications in transport systems

Method	Features and Parameters	Advantages	Application in Transport
Decision Tree (DT) (Messenger and Mandell, 1972),	Tree-like model; parameters include tree depth, min samples split, min samples leaf.	Easy to interpret; handles both types of data well.	Predicting the mode choice behaviour of commuters (Sekhar et al., 2016), automatic freeway incident detection (Chen and Wang, 2009).
Light Gradient Boosting Machine (LGBM) (Ke et al., 2017)	Gradient boosting; parameters include learning rate, number of leaves, max depth.	Fast training; efficient for large datasets.	Predicting traffic crash severity (Niyogisubizo et al., 2023).
k-Nearest Neighbours (KNN) Fix and Hodges (1989)	Instance-based; main parameter is the number of neighbours.	Simple implementation; effective for small datasets.	Short-term traffic forecasting (Sun et al., 2018), imputation of missing traffic data (Wang et al., 2020).
Gradient Boosted Decision Trees (GBDT) (Elith et al., 2008)	Ensemble of trees; parameters include number of trees, learning rate, depth.	High accuracy; reduces bias and variance; strong model from weak learners.	Predict the occupancy of public transport vehicles (Gallo et al., 2022).
Time Series Forecasting with LSTM (Gers et al., 2002), (Van Houdt et al., 2020)	Type of recurrent neural network; parameters include LSTM units, learning rate, batch size, layers.	Captures long-term dependencies in series; optimal for sequential data, especially with long-range dependencies.	Short-term traffic forecast (Zhao et al., 2017), air quality modelling (Krishan et al., 2019).

of values between 0 and 1 that represent the amount of each piece of information to forget from the memory cell (Van Houdt et al., 2020).

3. Memory cell: This is the main storage unit of the LSTM. It stores information over time and is controlled by the input and forget gates (Graves, 2012).

4. Output gate: This gate controls the output of the memory cell. It takes the input vector and the previously hidden state vector as inputs and produces a vector of values between 0 and 1 that represent the amount of each piece of information to output from the memory cell (Van Houdt et al., 2020).

Although LSTM as a time series machine learning model have been used for making predictions on time-stamped transport data (Krishan et al., 2019; Zhao et al., 2017), there is no prior work on using these models for validating the objectives of transport policy interventions. Searching the term "validating policy objective" and "LSTM" and "transport" in Google Scholar reveals no relevant result apart from the works published during this PhD research.

2.5 Optimisation Methods

Machine learning and optimisation methods are chosen among the available quantitative modelling techniques described in the Section 2.3 to address the research questions of this thesis on validation of the objectives of policy interventions and efficient implementation of policy commitments (cf. Section 1.2). This section reviews the principles of optimisation methods and prior research results to identify the related research gaps.

Optimisation algorithms play a key role in solving complex problems in various fields including transport engineering, economics, computer science, and biology (Vanderbei, 2014). The structure of optimisation algorithms can be divided into the following main components (Bertsekas, 2016):

- Decision variables: These are the variables that can be adjusted to find the optimal solution. The values of decision variables determine the value of the objective function.

- Constraints: These are the restrictions that the solution must satisfy. Constraints may be expressed as equality or inequality conditions.

- Objective function: This is the function that needs to be optimised. It defines the quantity that is to be maximised or minimised.

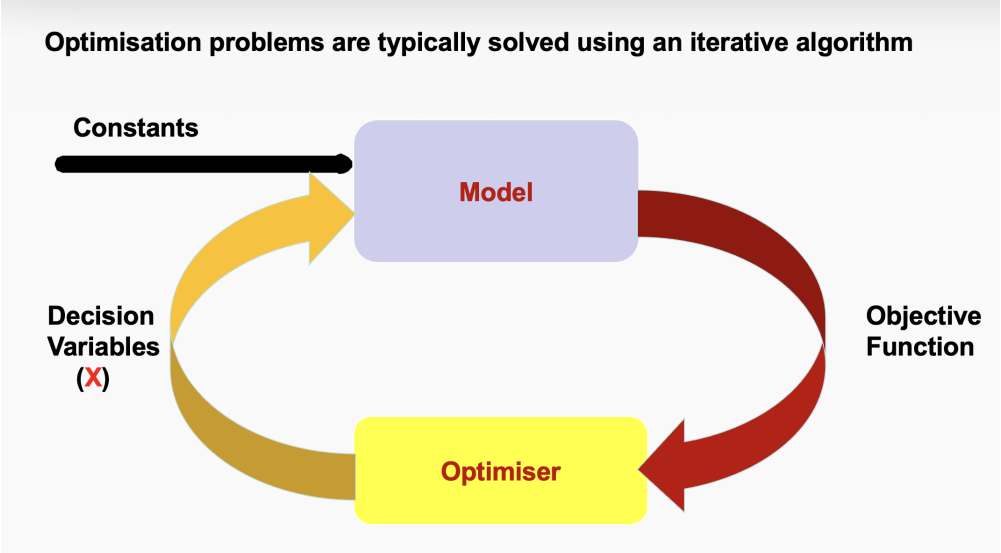

Fig. 2.4 The overview of optimisation problems (Bertsekas, 2016)

The overview of the optimisation components has been shown in Figure 2.4.

In the next subsections, classic deterministic optimisation algorithms, evolutionary randomised methods, and multi-objective algorithms will be discussed.

2.5.1 Classic Deterministic Optimisation Algorithms

Deterministic optimisation refers to a class of optimisation algorithms that follow a systematic and rule-based approach to find the optimal solution (Cavazzuti, 2013). These algorithms are based on mathematical programming and rely on precise mathematical formulations of the objective function and constraints. Deterministic optimisation is often used interchangeably with classical optimisation (Rader, 2010). Both terms refer to optimisation methods that rely on mathematical and computational models to find the optimal solution to a given problem (Rader, 2010). Deterministic optimisation assumes that the inputs to the optimisation model are known with certainty and that the objective function is deterministic, meaning that there is no uncertainty or randomness involved (Lin et al., 2012). Deterministic optimisation algorithms are a class of optimisation methods that aim to find the optimal solution for a single objective function (Persson et al., 2005).

Classic deterministic optimisation algorithms include gradient descent (Ruder, 2016), NewtonâĂŹs method (Polyak, 2007), and conjugate gradient (Nazareth, 2009). These algorithms have the following limitations: (a) sensitivity to the initial conditions, (b) the risk of getting trapped in local optima, and (c) the difficulty in handling non-convex problems. Moreover, these algorithms mainly handle single optimisation objectives, thus not suitable

for implementing transport policy commitments in which multiple conflicting objectives need to be addressed at the same time.

2.5.2 Evolutionary Randomised Algorithms

Evolutionary algorithms are inspired by the principles of natural evolution and genetics (Eiben and Smit, 2012). They are a family of optimisation algorithms that use mechanisms of selection, recombination, and mutation to search for the optimal solution to a problem (Egea et al., 2010). Evolutionary algorithms are implemented using the following steps:

- Initialisation: A population of candidate solutions is randomly generated (Tanskanen, 2002).
- Fitness evaluation: The fitness of each candidate solution is evaluated using an objective function that measures the quality of the solution (Jin et al., 2000).
- Selection: The fittest candidate solutions are selected for reproduction, based on their fitness.
- Reproduction: New candidate solutions are generated by applying recombination and mutation operators to the selected candidate solutions (Annunziato and Pizzuti, 2000).
- Replacement: The new candidate solutions replace the old ones in the population (Tanskanen, 2002).
- Termination: The algorithm terminates when a termination criterion is met, such as a maximum number of generations or when the optimal solution is found (Annunziato and Pizzuti, 2000).

The main strengths of evolutionary algorithms are:

- Robustness: Evolutionary algorithms are highly robust, as they can handle complex and highly nonlinear optimisation problems that are difficult to solve using traditional optimisation techniques (Vikhar, 2016).
- Global optimisation: Evolutionary algorithms can search the entire solution space and find the optimal solution, even in the presence of multiple local optima (Andrzej and Stanislaw, 2006).
- Adaptability: Evolutionary algorithms can adapt to changes in the problem environment, making them suitable for dynamic optimisation problems (Díaz-Manríquez et al., 2011).

- Black-box optimisation: Evolutionary algorithms do not require any prior knowledge of the problem, making them suitable for black-box optimisation problems (Bennet et al., 2021).

On the other hand, evolutionary algorithms can be computationally expensive, especially for high-dimensional optimisation problems (Bennet et al., 2021), and require the tuning of several parameters, which can be a challenging task for complex optimisation problems (Bennet et al., 2021).

In contrast to classic deterministic optimisation algorithms described in the previous section, evolutionary algorithms can handle multiple conflicting objectives (Deb and Deb, 2013) to generate a set of solutions that are optimal with respect to multiple objectives instead of a single objective (Fieldsend and Everson, 2005). Implementing transport policy commitments in general includes multiple conflicting objectives that need to be addressed at the same time. Therefore, for the research questions of this thesis, a class of evolutionary algorithms called *multi-objective genetic algorithm* (Gao et al., 2000) is selected for optimisation as a popular evolutionary algorithm that can handle multiple objectives and do not require any prior knowledge of the problem thus suitable for being integrated with machine learning models.

In a genetic algorithm (GA), solutions are represented as genes, and a population of these genes constitutes a generation (Forrest, 1996). The algorithm evolves through successive generations, applying genetic operators such as selection, crossover (recombination), and mutation to create offspring solutions. The fittest individuals, determined by their fitness values in relation to the problem's objective(s), are more likely to be selected for reproduction. This process is iteratively performed until a termination condition is met, such as a predefined number of generations, a specific fitness threshold, or a convergence criterion (Kumar et al., 2010). By mimicking natural evolution processes, GAs effectively explore and exploit the search space to find optimal or near-optimal solutions for complex optimisation problems.

Genetic algorithms have been used as an optimisation tool for tuning the hyperparameters of machine learning algorithms (Mehta, 2022). A prominent example of this is the work by Nikbakht et al. (2021), who proposed using GA to optimise the hyperparameters of deep neural networks. Similarly, other studies have used GA to optimise the parameters of support vector machines (Huang and Wang, 2006) and decision trees (Stein et al., 2005). Conversely, machine learning techniques have been used to enhance the performance of GA. One example is the use of neural networks as surrogate models to speed up the evaluation of fitness functions in GA (Rasheed et al., 2005). Another example is the use of machine learning to guide the search process in GA, such as the work by Zhou (2002), who proposed

a hybrid algorithm combining GA and reinforcement learning for solving optimisation problems.

A survey of multi-objective evolutionary algorithms can be founds in the work by (Zhou et al., 2011). A survey on the application of evolutionary algorithms in transport is provided by (Chen et al., 2022). Among these methods, Ant Colony Optimisation (ACO) is a heuristic algorithm inspired by the foraging behaviour of ant colonies (DâĂŹAcierno et al., 2010). It is shown by (Alexander and Sriwindono, 2020) that while ACO is able to find better solutions than the Genetic Algorithms when the fitness function is known, the Genetic Algorithm shows a better speed of completion. This feature is in particular important for the problem setting of this thesis where there are multiple fitness functions and the optimisation domain is very large.

Although the works reviewed above and summarised in Table 2.5 show promising results in the integration of GA and machine learning methods, there is no previous research on integrating them in a framework for finding optimal implementation of policy commitments.

2.6 Case Studies

This section provides information on the choice of the case studies for applying the methodologies developed in this thesis. Due to the global challenge of climate change and the growing environmental concerns, the focus will be on the Net Zero Emission strategy set by the UK government (Government, 2021a). The first case study is clean air zone with an objective for the policy intervention, which will be used to show how machine learning models can validate this policy objective. The second case study is the expansion of electric vehicle charging infrastructure using optimisation methods (cf. Figure 2.2).

With the advances in science and technology to better understand and evidence the effects of climate change, individuals and governmental organisations are paying more attention to alternative forms of energy obtained from solar, wind, hydroelectric, and geothermal power (Nehrir et al., 2011). Many countries are progressing in the path of introducing appropriate policies to reduce carbon emissions and mitigate the effects of climate change. Most notably, the 2021 United Nations Climate Change Conference (COP26) which was held in Glasgow, Scotland, which hosted delegates from 200 countries. The outcome of COP26 was a new deal, known as the Glasgow Climate Pact. The UK already committed in 2019 to a legally binding net zero target by 2050 and introduced new interim targets to reduce emissions by 78% by 2035 (Logan et al., 2022). Figure 2.5 shows the percentage of global total greenhouse gas emissions in 2018 with China, the USA, and the EU being responsible for almost half of global emissions (data from CAIT Climate Data Explorer via ClimateWatch).

Table 2.5 Integration of genetic algorithms and machine learning techniques

Reference	Purpose	Machine Learning Integration	Contributions
Mehta (2022)	Using GA for hyperparameter optimisation of machine learning models.	Generic, not specific to a model.	Demonstrated GA's utility in machine learning hyperparameter optimisation.
Nikbakht et al. (2021)	Using GA for hyperparameter optimisation of deep neural networks.	Deep Neural Networks	Proposed GA for deep learning model optimisation, enhancing performance.
Huang and Wang (2006)	Using GA for hyperparameter optimisation of SVMs.	Support Vector Machines	Used GA to fine-tune SVM parameters, resulting in improved classification accuracy.
Stein et al. (2005)	Using GA for hyperparameter optimisation for decision trees.	Decision Trees	Applied GA to determine optimal decision tree parameters, improving decision-making accuracy.
Rasheed et al. (2005)	Enhancing GA performance.	Neural networks as surrogate models	Utilised neural networks to accelerate GA fitness evaluations, improving optimisation efficiency.
Zhou (2002)	Guiding the GA search process.	Hybrid with Reinforcement Learning	Introduced a hybrid algorithm combining GA and reinforcement learning for optimisation problems.

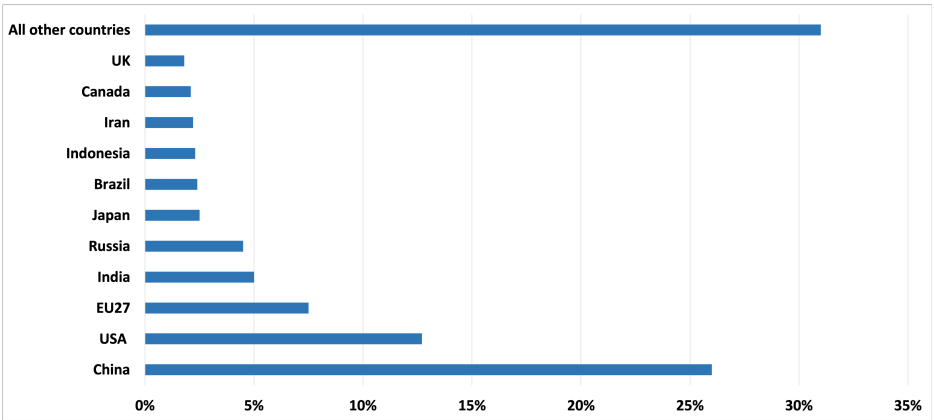

Fig. 2.5 Percentage of global total greenhouse gas emissions in 2018 (data from Climate-Watch)

In recent years, there has been growing concern about the impact of air pollution and greenhouse gas emissions on public health and the environment. In response, policymakers and experts have focused on improving air quality and reducing greenhouse gas emissions through various strategies and policies (National Institute of Environmental, 2019). Improving air quality is a complex challenge that requires a range of strategies, policies, commitments, programs, and projects across different sectors (Nations, 2015). Many of these strategies are interconnected and can have co-benefits for both air quality and greenhouse gas emissions (Trust, 2021).

Department for Business, Energy, and Industrial Strategy (BEIS) of the UK government has proposed a "Net Zero Strategy" to support the successful implementation of the strategy by investing over £100 billion in efforts to eliminate carbon emissions (BEIS, 2021). As a key tenant of the new technologies to reduce carbon emissions, electric vehicles (EVs) are making a rapid sales progress with a yearly sales increase of 20% in 2022 (ZapMap, 2022b). The UK plans to achieve 100% zero emissions on all new cars and trains by 2035 by actively pursuing the development and sale of Ultra Low Emission Vehicles and Zero-Emission Vehicles. To address the inevitable increasing demand for charging EVs, BEIS has committed to a minimum of 2,500 charging points across the strategic road network in the UK (HM Government, 2021).

The UK's **Net Zero Emissions Strategy**, which aims to achieve net-zero greenhouse gas emissions by 2050, includes a range of policies and initiatives across various sectors, including energy, transport, buildings, industry, and agriculture (Government, 2021a). Some of these key policies include (Government, 2021a):

- Increasing renewable energy sources: The UK has set a target to generate at least 50% of electricity from renewable sources by 2030.

- Encouraging energy efficiency: The UK has set a target to improve energy efficiency in homes, businesses, and industry.
- Promoting low-emission vehicles: The UK has set a target for the ban on the sale of new petrol and diesel cars from 2030 and for all new cars and vans to be zero-emission by 2035 by transitioning toward electric vehicles.
- Reducing emissions from buildings: The UK has set a target for all new homes to be built to a zero-carbon standard by 2025.
- Reducing emissions from industry: The UK has set a target to reduce emissions from industry by at least two-thirds by 2050.

The next two subsections give reviews the principles of clean air zones and transition towards electric vehicles in the context of the UK's Net Zero Emission Strategy. Table 2.6 gives an overview of these two case studies used in this thesis.

Table 2.6 Overview of the case studies used in this thesis to demonstrate the designed frameworks

Aspect	**Clean Air Zone**	**Transition Towards Electric Vehicles**
Policy Focus	Encourage the use of low-emission vehicles, introduce daily charges	Increase the number of charging points, improve accessibility
Objective	Improve air quality and reduce pollution in cities	Reduce carbon emission and improve air quality
UK Government's Commitment	Target net-zero carbon emissions by 2050, introduce clean air zones in cities	Expanding the charging infrastructure to support the growing number of EVs on the road.

2.6.1 Clean Air Zone

The UK government has set a clean air zone policy to improve air quality and reduce air pollution in cities across the country (UK Government, 2021). The policy aims to reduce the harmful emissions from vehicles by encouraging the use of low-emission vehicles and reducing the number of high-polluting vehicles on the roads.

The clean air zone policy was first introduced in 2017, following a court ruling that the UK government was failing to meet EU air quality standards. The policy was designed to

help reduce the levels of nitrogen dioxide and particulate matter in the air, which are harmful to human health and the environment (UK Government, 2021). Under the policy, local authorities in England are required to identify areas with poor air quality and implement measures to reduce pollution, including the introduction of clean air zones. These zones are areas where certain types of vehicles are required to pay a daily charge to enter, based on their level of emissions.

The policy is being rolled out gradually across the country, with several cities already introducing clean air zones, including Birmingham, Bath, and Leeds. London has also introduced its own Ultra-Low Emission Zone, which operates in the same way as a clean air zone (Transport for London, 2021).

The UK government has set a target of reaching net-zero carbon emissions by 2050, and the clean air zone policy is an important part of achieving this goal (Transport for London, 2021). The policy is also in line with the UK's obligations under the Paris Agreement, which aims to limit global warming to below 2ÂřC above pre-industrial levels.

The UK Department for Transport (DfT) and the Department for Environment, Food & Rural Affairs (Defra) have jointly produced a report that lays out a framework for the design and operation of clean air zones in England (Defra Joint Air Quality Unit, February 2020). It recommends the approach to be taken by local authorities when implementing and operating a clean air zone. These recommendations apply also to the clean air zone being considered for implementation in Newcastle. An example of such an implementation in England is the Greater Manchester Clean Air Zone (GMCAZ) (Clean Air Greater Manchester, 2020). The interventions designed in the GMCAZ was launched on 30 May 2022 and requires the vehicles that do not meet the emissions standards, to pay a fee when entering the clean air zone. The non-compliant vehicles are: Heavy goods vehicles, buses, coaches, vans, minibuses, private hire cars, and motorcaravans that have a EUROV or earlier diesel engine that have a EUROIII or earlier petrol engine. Private cars are not currently affected by the intervention. This new scheme also provides financial support with more than £120m of government funding to help businesses in the region, organisations and people switch to compliant vehicles by either replacing or retrofitting non-compliant vehicles.

2.6.2 Transition Towards Electric Vehicles

The UK government has introduced several policies aimed at promoting the transition towards electric vehicles (EVs) in recent years. These policies are part of the UK's efforts to reduce carbon emissions and improve air quality (Wills, 2021). This section will discuss some of the key policies that have been introduced to promote the uptake of EVs in the UK.

- Plug-in Car Grant: The UK government offers a grant to drivers who purchase a new electric vehicle. This grant was first introduced in 2011 and has been extended several times since then (UK Government., 2021). Currently, drivers can receive a grant of up to £2,500 towards the cost of a new electric car, and up to £350 towards the cost of a new electric motorcycle.

- Workplace Charging Scheme: The UK government offers a grant to businesses that install EV charging points at their premises (Government, 2021b). The Workplace Charging Scheme provides a grant of up to £350 per charging socket, up to a maximum of 40 sockets per business.

- Electric Vehicle Homecharge Scheme: This scheme provides a grant of up to 75% of the cost of installing a home charging point, up to a maximum of £350 (Gov.UK, 2021). This scheme is designed to make it easier and more convenient for people to charge their EVs at home.

- Road to Zero Strategy: In 2018, the UK government published its Road to Zero strategy, which outlines its plans for transitioning towards zero-emission vehicles (HM Government, 2018). The strategy includes a target for at least 50% of new cars and 40% of new vans to be ultra-low emission by 2030, as well as plans to invest in EV charging infrastructure and support for research and development.

- Zero Emission Vehicle Mandate: In 2020, the UK government announced plans to introduce a Zero Emission Vehicle Mandate, which would require car manufacturers to sell a certain percentage of zero-emission vehicles each year (HM Government, 2018). The mandate is expected to be introduced in 2030, following a public consultation.

- Ban on new petrol and diesel vehicles: The UK government has announced a ban on the sale of new petrol and diesel cars and vans from 2030 (HM Government, 2018). This policy is part of the government's plan to achieve net-zero greenhouse gas emissions by 2050 and will accelerate the transition towards EVs (HM Government, 2018).

These policies have already had a significant impact on the uptake of EVs in the UK (OFGEM, 2018). In 2020, more than 108,000 new electric cars were registered in the UK, accounting for 6.6% of all new car registrations. This represents a significant increase from 2019 when electric cars accounted for just 1.6% of new car registrations (OFGEM, 2018). The commitments that the UK government has made to facilitate the transition towards electric vehicles (HM Government, 2021) are as follows:

- Grants for EV purchases: The government offers a grant of up to £2,500 towards the purchase of eligible EVs to encourage the uptake of cleaner vehicles. This grant is available to all UK residents and can be applied for through the Office for Zero Emission Vehicles.
- Funding for EV infrastructure: The UK government has committed £1.3 billion to support the rollout of EV charging infrastructure across the country. This funding will help to expand the network of charging points and make EVs more accessible to drivers.
- Company car tax incentives: The government has introduced tax incentives to encourage businesses to choose EVs for their company car fleets. The incentives include lower tax rates for EVs and a zero rate of company car tax for pure electric vehicles.

As the UK moves towards electric vehicles, one of the most important commitments is expanding the charging infrastructure to support the growing number of EVs on the road. This expansion is necessary to address range anxiety and increased demand, stimulate the economy, and reduce the environmental impact of the transport sector.

Chapters 4 and 5 of this thesis are focused on applying the methodologies developed in this thesis to clean air zone, which aimed to improve air quality using machine learning methods. Chapter 6 examines the the application of the methodologies to expansion of electric vehicle charging infrastructure using optimisation methods. In the next two sections, the literature on the use of machine learning methods and optimisations on these two case studies are reviewed to identify the gaps.

2.6.3 Machine Learning Techniques Applied to Air Quality

The technical reports published in the website of Greater Manchester Clean Air Zone[3] demonstrate the role of data in designing the related interventions and evaluating them when they are implemented. In general, datasets play two main roles: (1) datasets are used to select and tune parameters of the physical models developed for air quality, and (2) datasets are used for monitoring and checking if the policy objectives are achieved (e.g., reducing the pollutant level to some value).

In contrast to a clean air zone in which the local authority is actively trying to improve the quality of the air, *low emission zones* put restrictions specifically on vehicles that do not meet a minimum standard for vehicle emissions, e.g., the European Union's emissions standards (European Environment Agency, 2020) on harmful air pollutants and greenhouse

[3] https://cleanairgm.com/technical-documents

gases. Ellison et al. (2013) studied how the low emission zone implemented in London impacted the vehicle usage and air pollution by using the data from the registration and enforcement information. The authors focus on concentrations of particulate matter PM_{10} and NO_x. This choice was made due to the fact that in London, approximately 25% of PM_{10} and 57% of NO_x emissions from road transport come from heavy vehicles (Transport for London, 2008). Using ambient air quality measurements, they showed that concentrations of particulate matter have dropped by 2.5–3.1% within the low emission zone, but they did not find any noticeable differences for all measured NO_x concentrations.

Another line of research studies various methods for specifying and deciding on boundaries of a clean air zone. Karaca (2012) uses data from the air quality monitoring stations and combines statistical analyses with interpolation techniques to identify the areas with the highest concentrations of particulate matter. Sturman and Zawar-Reza (2002) integrate an atmospheric model with a kinematic model to identify the boundaries of the catchment of air affecting concentrations of air pollution. The available data is used to initialise and calibrate the models. Yearley et al. (2003) discuss an empirical approach called 'participatory modelling' to find spatial representations of local knowledge about air pollution and use them in local governance of air quality. They present empirical data from a three-city case study and generate maps using local knowledge, which can then be used as a form of consultation for the local governance of the politics on air pollution.

Although the above works show the potential of data-driven techniques for modelling, analysing, and addressing challenges related to air quality, developing data-driven methods for validating policy objectives on air quality has not received attention. The rest of this section reviews machine learning models used to solve problems on air quality.

A systematic review of data-driven methods for air pollution prediction is provided by Iskandaryan et al. (2020). Their work provides an overview of machine learning techniques used in the smart city domain to predict air quality, and classifies the temporal resolutions analysed with these techniques. Prediction of NO_2 concentration in an air quality monitoring site of the Greater Manchester Area, United Kingdom, is performed by Catalano and Galatioto (2017) using a moving average model and a neural network model. Their results show that the accuracy of moving average in the prediction of extreme air pollution events is 27% better than the standard statistical methods and 113% better than using neural network models. This shows that methods for time series modelling and analysis are better in making predictions of time-stamped data with temporal dependencies. This observation has been considered in the choice of the machine learning models of this thesis.

Navares and Aznarte (2020) have focused on air quality prediction in Madrid, Spain. Their work provides a method to predict CO, NO_2, O_3, PM_{10}, SO_2 and pollen concentrations

using long short-term memory (LSTM). They try to find the best configuration in the LSTM (e.g., how neurons are connected) for reducing the prediction error and having robust one day-ahead air quality predictions. Their work does not include any results on using these predictions for validating policy objectives or their design.

Rybarczyk and Zalakeviciute (2021) consider the travel restrictions and the lockdowns imposed due to the COVID-19 pandemic, and studies their impact on Air Quality using machine learning methods. Their work uses Gradient Boosting Machine algorithm to assess the impact of full or partial lockdown on air quality. Their approach is to use pre-lockdown data to predict the pollution levels without lockdown and then compare the predictions with pollution levels measured under lockdown measures. Another study is performed by Turner et al. (2020) to understand the effect of COVID-19 restrictions on CO_2 emissions. They use CO_2 observations and an atmospheric transport model to compute changes in CO_2 emissions caused by the imposed lockdown. They predict a 30% decrease in CO_2 emissions and conclude that this reduction is mainly due to changes in road vehicles as opposed to other non-traffic emissions which showed small changes.

Castelli et al. (2020) apply support vector regression (SVR) as a machine learning approach to predict the air quality index (AQI). AQI is an index for quantifying the level of pollution of air. Its values range from 0 to 500 parts per billion (ppb), where higher values indicate larger pollution. The SVR model used by Castelli et al. (2020) is a nonlinear mapping that maps the dataset into a feature space and fits a linear regression model to the dataset in the new feature space. A kernel function is defined as a mapping from the input space to the new feature space. They show empirically that radial basis function as kernel functions gives the best result for prediction on the chosen dataset. Their approach is implemented on a dataset containing hourly data measured from the state of California, USA, between January 1, 2016, and May 1, 2018. The result shows that the pollutant concentrations can be successfully predicted to using SVR method as a regression problem, and the six categories of AQI can be predicted as a classification problem with an accuracy of 94.1%. They do not study the effect of interventions for improving air quality.

Suleiman et al. (2019) use three machine learning methods for predicting concentrations of PM_{10} and $PM_{2.5}$ using road traffic, meteorological data and pollutant data measured and stored at Air Quality Monitoring sites of London. The machine learning models used in their study include Artificial Neural Networks (ANN), Boosted Regression Trees (BRT) and Support Vector Machines (SVM). The reported implementations show that ANN and BRT are better than SVM in predicting PM_{10} and $PM_{2.5}$ concentrations and these two models can be applied in managing the traffic-related particulate matter concentrations. The authors also conceptualised a hypothetical scenario to demonstrate the use of machine learning models in

air quality management. The scenario assumed that the study area permits only EUROIV petrol and EUROVI diesel vehicles to be driven in that area. The dataset is the revised using the Emissions Factors Toolkit (EFT) (Defra, 2022) and new machine learning models are constructed on the modified data. The paper demonstrate that machine learning methods can be used to forecast concentrations of pollutants PM_{10} and $PM_{2.5}$ whenever rich datasets are available.

Orun et al. (2018) have developed a Bayesian network method with an optimised configuration to provide a probabilistic traffic data analysis and to predict traffic-related air pollution. Machine learning predictive models are developed by Harishkumar et al. (2020) for predicting particulate matter concentration using Taiwan Air Quality Monitoring datasets from 2012 to 2017. The developed predictive models were compared with the traditional models and cross-validation is used to select the best model with the highest performance. Vosough et al. (2020) have studied the reduction of ambient air pollution and congestion using weather forecasts and predictive cordon tolls. The authors used a model of emission dispersion to forecast air quality using recorded weather data for Tehran in 2016. It is shown that the constructed pricing scheme decreases the daily average CO concentrations.

Although the works reviewed above and summarised in Table 2.7 make promising observations on the use of machine learning methods for making predictions on air quality, these works do not give a framework for efficient analysis of policy objectives related to air quality. Chapters 4 and 5 of this thesis provide a general data-driven framework for analysis of the policy objectives that have machine learning methods as a core modelling and computation component. As a proof of concept, the reduction in the concentration of NO_2 by the implemented clean air zone will be studied.

2.6.4 Optimisation Models for Electric Vehicle Charging Infrastructure

With the rapid increase in the variety and number of EVs, it is becoming challenging to meet the growing demand for charging these EVs. The challenges include the electricity grid overloading, forecasting the required charging load, and charging time and traffic-crowd management at charging stations (Pareek et al., 2020). The research conducted by Illmann and Kluge (2020) studies the relation between an increasing availability of public charging infrastructure and consumers decisions to switch to EVs. They find evidence of a positive long-run relationship, and conclude that consumers also attach more importance to the charging speed. Therefore, governments and experts are interested in satisfying the demand for EV charging and maximising the economy and benefits. Cai et al. (2014), Yang et al. (2017), and Wang et al. (2016) have studied and simulated the optimal location of charging

Table 2.7 Overview of previous research works applying machine learning methods to air quality

Reference	Methodology/ Techniques	Focus and Contribution
Ellison et al. (2013)	Data analysis using registration and enforcement information	Impact of low emission zone in London on vehicle usage and air pollution
Karaca (2012)	Statistical analyses with interpolation techniques	Identifying areas with high particulate matter concentrations
Sturman and Zawar-Reza (2002)	Integration of atmospheric and kinematic models	Identifying boundaries of air pollution catchment areas
Yearley et al. (2003)	Participatory modelling	Empirical approach for spatial representation of local knowledge on air pollution
Iskandaryan et al. (2020)	Systematic review	Overview of machine learning techniques in smart cities for air quality prediction
Catalano and Galatioto (2017)	Moving average model and neural network	Prediction of NO_2 concentration; comparison with traditional methods
Navares and Aznarte (2020)	Long short-term memory (LSTM)	Air quality prediction in Madrid; robust one day-ahead predictions
Rybarczyk and Zalakeviciute (2021)	Gradient Boosting Machine	Assessing the impact of COVID-19 lockdown on air quality
Turner et al. (2020)	Atmospheric transport model	Studying the effect of COVID-19 restrictions on CO_2 emissions
Castelli et al. (2020)	Support vector regression (SVR)	Predicting air quality index using SVR
Suleiman et al. (2019)	ANN, BRT, SVM	Predicting PM_{10} and $PM_{2.5}$ concentrations using traffic data
Orun et al. (2018)	Bayesian network	Probabilistic traffic data analysis for air pollution prediction
Harishkumar et al. (2020)	Machine learning models	Forecasting particulate matter concentration in Taiwan
Vosough et al. (2020)	Emission dispersion model	Forecasting air quality and congestion reduction strategies

stations for taxis and buses and have extended this to the location of charging stations for private EVs.

Although the expectations and constraints for taxi and bus charging are different to that for private vehicles, the studies mentioned above highlight the importance of developing simulation models for EV charging infrastructure. On the other hand, LaMonaca and Ryan (2022) provide a review of the EV charging infrastructure market, and identify the relationship between governments, investors and individuals by presenting and analysing its essential functions and the roles of the players that will fuel the deployment of large-scale EV infrastructure. They find that assigning clear roles to the public and private actors and funders is needed to achieve efficient development of the required infrastructure for large-scale EVs.

A summary of the research reviewed above is provided in Table 2.8 that overall look at the market, the economy and the environment as entry points for the location of charging stations. However, there are still deficiency in the study of the actual quantity and types of charging points. There is a lack of appropriate analysis of the problem at the micro level (i.e., the spatial distribution of the charging points in a geographical region). This may lead to an inappropriate utilisation of resources and an unnecessary burden on the power grid: having an unbalanced charging demand in different areas of the region may violate the constraints of the existing electricity grid distribution. Satisfying these constraints is detrimental to the development and implementation of policy commitments on the new EVs in the long run. Therefore, it is necessary to comprehensively analyse and optimise the type, quantity, location, and total capital and operational expenditures of charging points from a combination of micro and macro levels.

While the existing literature on optimisation and electric vehicle charging infrastructure presents valuable insights and approaches for planning and managing charging networks (Neaimeh et al., 2015; Sadati et al., 2020; Zhang et al., 2020; Zhou et al., 2022), there is a need for a comprehensive framework that integrates advanced techniques to effectively address the challenges associated with the rapid growth of EVs. Chapter 6 shows how the novel data-driven optimisation framework of this thesis can combine machine learning methods, such as long short-term memory (LSTM) and fuzzy logic, with multi-objective genetic algorithms to efficiently plan and deploy electric vehicle charging infrastructure. This framework not only considers multiple objectives, such as cost minimisation and charging network coverage but also accounts for real-world constraints and uncertainties to ensure its practical applicability.

Previous studies generally rely on restrictive assumptions that limit the scope for application of the results. Studies that provide optimisation for a single type of charging point are available (Bayram et al., 2022; Sundström and Binding, 2010). In order to better investigate

Table 2.8 Summary of research on optimisation models for EV charging infrastructure

Reference	Study Focus/Objective	Key Findings and Contributions
Pareek et al. (2020)	Challenges in EV charging infrastructure	Discussed grid overloading, charging load forecasting, and traffic management at charging stations
Illmann and Kluge (2020)	Relation between public charging infrastructure and EV adoption	Found a positive long-run relationship; importance of charging speed emphasised
Cai et al. (2014)	Optimal location of charging stations for taxis	Simulated and studied optimal station locations, extended to private EVs
Yang et al. (2017)	Data-driven approach for EV charging stations	Simulated optimal locations for bus and taxi charging stations
Wang et al. (2016)	Simulation of charging station locations	Focused on charging station location for private EVs
LaMonaca and Ryan (2022)	Review of EV charging infrastructure market	Identified roles of public and private actors in infrastructure development
Zhou et al. (2022)	Building a social cost model for placement of charging stations	The social cost is sensitive to the number of charging stations, demand at intersection points and probability of charging each day
Zhang et al. (2020)	Solving EVs charging scheduling problem using reinforcement learning	minimising the total charging time of EVs while maximising reduction in the origin-destination distance
Sadati et al. (2020)	Study solar-based EV car parks with private owners	Profit maximisation and cost minimisation
Neaimeh et al. (2015)	Probabilistic approach to combine smart meter and electric vehicle charging data	Investigate impacts on the distribution network
Bayram et al. (2022)	Finding a closed-form expression for the plug-in electric vehicles charging station capacity	Calculation of the optimal service capacity for charging locations
Sundström and Binding (2010)	EV charging schedule optimisation while minimising charging costs	Achieving satisfactory state-of-charge levels and optimal power balancing.

multiple aspects of EV infrastructure planning at the same time, this research uses genetic algorithm, which is improved based on the concepts of long short-term memory and fuzzy logic.

2.7 Research gaps

This chapter has reviewed extensively the range of research topics related to the main subject of this study, starting from policy design, implementation, and validations in the transport sector, followed by the role of quantitative methods in transport and a review of machine learning and optimisation methods.

The reviewed research papers revealed two key observations: (a) limited work in which the objectives of a policy intervention is validated using machine learning methods; and (b) a need for a framework to evaluate policy implementations against their objectives using data gathered through digitalisation of the transport sector. More work is needed to integrate data-driven approaches in assessing the objectives of policy interventions and refining their implementations using the additional knowledge gained through data. These observations shaped the high-level aim of this PhD research that has elements from processing the data to extract useful knowledge, building computational models for the objectives of policy interventions, and use them to improve the implementations of policy commitments and their outcomes.

Next, quantitative modelling techniques were reviewed, concluding that machine learning and optimisation models are better suited for validating the objective of policy interventions and improving the implementation of policy commitments in comparison with discrete choice models. This is mainly due to the limitation of discrete choice models focusing on the individualistic behaviours of people than their aggregate effects on the whole transport system.

Then, classes of machine learning methods where reviewed with their properties, advantages, and disadvantages. It is concluded that *supervised learning methods* are appropriate for the quantitative policy objectives in transport than reinforcement learning or unsupervised learning. This is mainly due to (a) the inherent limitation of reinforcement learning that requires a well-defined reward system and an accurate interactive dynamic environment which is not available with the current technological developments in the transport sector; and (b) the limitations of unsupervised techniques that cannot handle large datasets with its outcome being subjective to interpretation and is difficult to judge the quality of the outcome.

Classes of optimisation methods where reviewed with their properties, advantages, and disadvantages. It is concluded that evolutionary randomised algorithms are appropriate for

for handling multiple conflicting objectives in implementing transport policy commitments than classic deterministic optimisation algorithms. Classic optimisations are limited due to (a) sensitivity to the initial conditions; (b) the risk of getting trapped in local optima; and (c) the difficulty in handling non-convex problems.

The reviewed literature were considered to identify the gaps in (a) machine learning methods in validating the objectives of policy interventions, and (b) data-driven optimisation techniques for implementing transport policy commitments. Although the reviewed works show promising results in the integration of evolutionary optimisation and machine learning methods, there is no work integrating them in a framework for finding optimal implementation of policy commitments.

Due to the global challenge of climate change and the growing environmental concerns, two case studies from the UK's Net Zero Emission strategy (Government, 2021a) were selected as candidates for demonstrating the methodologies developed in this thesis. The first case study is clean air zone and the second case study is the expansion of electric vehicle charging infrastructure.

The limitations of the current research on the chosen case studies for applying the methodologies developed in this thesis were also identified, summarised in Table 2.9. In the clean air zone case study, the most relevant papers were the ones reported in the first row and second column of the table. Although these works make promising observations on the use of machine learning methods for making predictions on air quality, they do not give a framework for efficient analysis of policy objectives related to air quality.

In the case study on the EV charging infrastructure case study, the most relevant literature were the ones reported in the second row and second column of the Table 2.9. Although these works provide valuable insights and approaches for planning and managing EV charging networks, there is no comprehensive framework that integrates machine learning techniques with simulation models to effectively address the challenges, real-world constraints, and multiple conflicting objectives associated with the rapid growth of EVs.

2.8 Conclusions

This chapter reviewed over 140 references on the main subject of this study. The identified key gaps in the existing literature are

- Limited work on data-based validation of policy objectives.
- Lack of a comprehensive data-driven framework for validating the objectives of policy interventions.

Table 2.9 Summary of research gaps in case studies

Case Study	Relevant Literature	Identified Gaps in Current Research	Contribution of This Thesis
Clean Air Zone	Ellison et al. (2013), Karaca (2012), Sturman and Zawar-Reza (2002), Yearley et al. (2003), Iskandaryan et al. (2020), Navares and Aznarte (2020), Rybarczyk and Zalakeviciute (2021), Turner et al. (2020), Castelli et al. (2020), Suleiman et al. (2019), Orun et al. (2018), Harishkumar et al. (2020), Vosough et al. (2020).	Works focus on air quality predictions but lack a framework for analysing policy objectives related to air quality.	Developing a framework for efficient analysis and validation of air quality policy objectives using machine learning methods.
Electric Vehicle Charging Infrastructure	Pareek et al. (2020), Illmann and Kluge (2020), Cai et al. (2014), Yang et al. (2017), Wang et al. (2016), LaMonaca and Ryan (2022), Zhou et al. (2022), Zhang et al. (2020), Sadati et al. (2020), Neaimeh et al. (2015), Bayram et al. (2022), Sundström and Binding (2010).	Insights on planning and managing EV charging networks but no integration of machine learning techniques with simulation models for addressing real-world challenges.	Creating a comprehensive framework that combines machine learning with simulation models for EV charging infrastructure challenges.

- Need for integrating machine learning methods with policy intervention analysis.
- Limited work on data-driven optimisation integrated with simulation models for policy implementation.

These identified research gaps serve as crucial areas for the proposed PhD research to address, aiming to contribute to the field by developing a methodology that fill these voids and enhance the understanding and implementation of transport policies.

The extensive review summarised in this chapter has motivated the research proposed in this PhD thesis. It is important to integrate machine learning and optimisation methods for validating the objectives of transport policy interventions and implementing the policy commitments. This research expects to develop novel data-driven frameworks designed specifically for filling the identified gaps. Next chapter describes the methodology and the designed frameworks.

For the selection of case studies, Clean Air Zone and transition towards electric vehicles were considered in the context of the UK's Net Zero Emission Strategy, and the literature related to the topic of this PhD research were reviewed. It is concluded that although promising observations are made in the use of machine learning for predictions in air quality, there is a lack of a cohesive framework for the efficient analysis of policy objectives, particularly related to clean air zones. The reviewed literature notes a lack of work that integrates machine learning techniques with simulation models to effectively address challenges, real-world constraints, and conflicting objectives associated with the rapid growth of electric vehicles. Therefore, clean air zones and expansion of the electric vehicles charging infrastructure will be used in the subsequent chapters to demonstrate the designed data-driven frameworks of the next chapter.

Chapter 3

Methodology

A critical review of the previous research works was covered in Chapter 2. The reviewed papers revealed the limited work in which the success of a policy intervention is assessed using data, and showed the need for a framework to validate policy interventions against their objectives and implement policy commitments using data. The review identified the following research questions in integrating machine learning and optimisation methods for validating transport policy objectives and implementing the policy commitments: **RQ1.** Given the large volume of data, what data types are relevant to the objectives of a policy intervention? **RQ2.** What machine learning techniques are suitable for combining large datasets and validating an intervention? **RQ3.** Could these data-driven techniques be used for efficient optimal implementation of policy commitments?

In order to answer the above research questions, the following objectives are considered: **O1.** Identify, gather, preprocess and analyse data types relevant to a policy from different sources. **O2.** Develop suitable machine learning models based on the input processed data and the considered policy objectives and commitments. **O3.** Analyse and simulate future scenarios under the implementation of the policy commitments to gain insights on their impact in the transport network. **O4.** Study methods for validating the outcome of machine learning methods. Select and use metrics that can best describe the accuracy of the outcome and validate the outcome against domain knowledge. **O5.** Determine the potential use of optimisation methods for transport policy commitment implementation. **O6.** Apply the designed frameworks to case studies on validating the policy objectives of clean air zone and the expansion of the electric vehicle charging infrastructure, which are critical for achieving the UK's target of net-zero emissions by 2050.

Regarding the case studies, Chapter 2 also demonstrated that previous research works do not give a framework for efficient analysis of policy objectives related to air quality. Previous works do not provide a comprehensive framework that integrates machine learning techniques

with simulation models to effectively address the challenges, real-world constraints, and multiple conflicting objectives associated with the rapid growth of electric vehicles. Therefore, this chapter describes the datasets collected and prepared for applying the methodology of the thesis and the designed frameworks to these case studies.

This chapter starts with a description of the current state of practice in evidence-based policy-making in Section 3.1 followed by a discussion on the impact of data-driven methods on policy-making processes in Section 3.2. Three data-driven frameworks that address the research questions of this thesis are described in Section 3.3, with their integration within the policy cycle presented in Section 3.4. The first framework is developed for identifying relevant data types and is presented in Section 3.5. The second framework is developed for validating the objectives of policy interventions and is presented in Section 3.6. The third framework is developed for optimal implementation of policy commitments and is presented in Section 3.7. Finally, the choice of programming language for implementing the methodology and applying the frameworks to case studies in subsequent chapters is presented in Section 3.8 followed by conclusions of this chapter.

3.1 Current State of Practice in Evidence-Based Policy-Making

Policy-making is a complex and multifaceted process that requires the integration of various sources of information, stakeholder perspectives, and evidence-based insights. WEBTAG (Web-based Transport Analysis Guidance)[1] is a framework developed by the UK Department for Transport (DfT) to provide guidance on the appraisal of transport policies. It offers comprehensive guidelines and methodologies for assessing the economic, environmental, and social impacts of transportation initiatives. WEBTAG aims to ensure that transport projects are appraised consistently, transparently, and in line with government objectives. Another guidance for transport policy making is the Green Book[2] issued by HM Treasury in the United Kingdom. The Green Book is a comprehensive guide designed to provide government departments and agencies with best practices for appraisal and evaluation of policies and projects.

The key components discussed in WEBTAG, the Green Book, and other evidence-based policy-making guidelines (Bulmer et al., 2007; De Marchi et al., 2016; Research Oxford University, 2024; Strydom et al., 2010), are as follows:

[1] https://discovery.nationalarchives.gov.uk/details/r/C16957
[2] https://shorturl.at/JGunZ

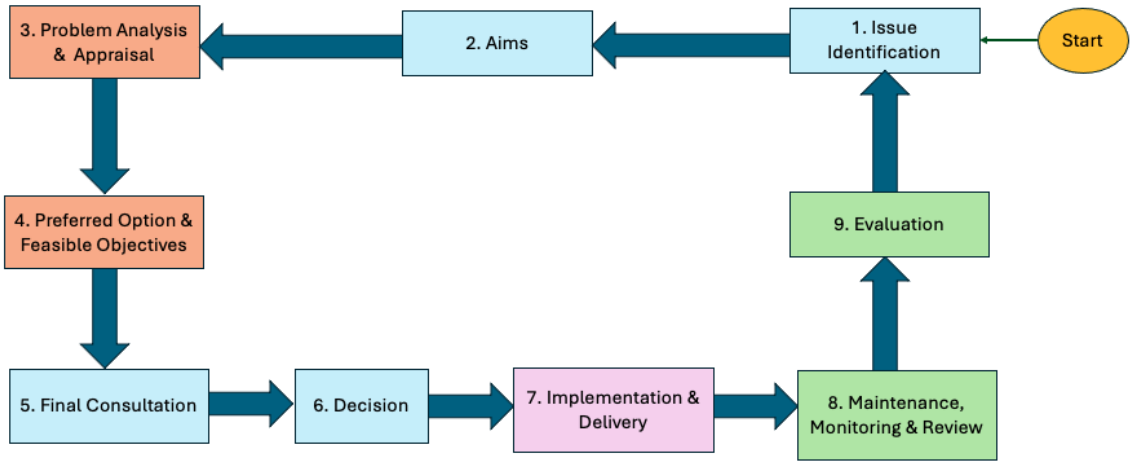

Fig. 3.1 The cycle of policy design and implementation (De Marchi et al., 2016; Research Oxford University, 2024)

1. **Transport Appraisal Process**: developing a strategic case that justifies the project in terms of its alignment with policy objectives and strategic fit; developing an economic case that makes a cost-benefit analysis; developing a financial case that considers the financial affordability, funding, and budgetary implications; and developing a management case that focuses on the deliverability and management of the project, including governance, risk management, and project planning.

2. **Appraisal Framework**: for performing demand forecasting, cost-benefit analysis, environmental impact assessment, and social impact assessment.

3. **Data Collection and Analysis**: providing guidance on conducting surveys, studies, and data collection to support the appraisal process.

4. **Documentation and Reporting**: standardised templates and formats for presenting appraisal results and processes for reviewing and validating appraisal reports.

Based on the available guidelines for evidence-based policy-making including WEBTAG, the Green Book, and Research Oxford University (2024), and the recent works by De Marchi et al. (2016), Strydom et al. (2010), and Bulmer et al. (2007), the cycle of policy design and implementation consists of the following steps, as illustrated in Figure 3.1:

1. **Issue Identification:** This first step in the policy cycle involves identifying the issues that require government intervention. This step includes recognising and defining the problem, understanding its root causes, and determining its scope and impact on

society. It is crucial to clearly articulate the issue to ensure that all stakeholders have a common understanding of the problem that needs to be addressed.

2. **Aims:** Once the issue is identified, the next step is to clearly define the aims of the policy intervention. This involves setting broad goals and objectives that the policy seeks to achieve. The aims should be aligned with the overall strategic priorities of the government or organisation and should address the core aspects of the identified issue.

3. **Problem Analysis & Appraisal:** This step involves a thorough analysis of the problem, generating various options to address it, and appraising these options based on their feasibility, effectiveness, and potential impact. The analysis should consider multiple dimensions of the problem, including economic, social, and environmental factors. Appraisal involves evaluating the pros and cons of each option, considering factors such as cost, implementation complexity, and expected benefits.

4. **Preferred Option & Feasible Objectives:** After appraising different options, the most promising one is selected as the preferred option. This step also involves setting feasible objectives that the preferred option aims to achieve. The objectives should be specific, measurable, achievable, relevant, and time-bound. This ensures that the policy intervention has clear targets and milestones that can be tracked over time.

5. **Final Consultation:** Before finalising the policy, it is essential to consult with stakeholders, including experts, affected parties, and the general public. This consultation helps in refining the policy and ensuring it is comprehensive and acceptable to all relevant parties. Stakeholder engagement is critical for gaining support, addressing concerns, and incorporating diverse perspectives into the policy design.

6. **Decision:** Making the final decision on which policy option to implement.

7. **Implementation & Delivery:** Putting the policy into action through various programs, regulations, and initiatives.

8. **Maintenance, Monitoring & Review:** Continuously monitoring the implementation of the policy to ensure it is achieving its objectives, and making necessary adjustments.

9. **Evaluation:** Assessing the overall impact and effectiveness of the policy.

Each step in the policy cycle is critical for the successful development and implementation of a policy. The data-driven frameworks developed in this research provide support at various stages of this process.

3.2 Impact of Data-Driven Methods on Policy-Making Processes

Traditional policy-making approaches often rely heavily on expert judgement, qualitative analysis, and historical data. While these methods have proven effective in many cases, they can be limited by subjective biases and the static nature of the data used. In an era where data is abundantly available and computational power has significantly increased, there is a growing opportunity to improve policy-making processes through the application of data-driven approaches. The integration of data-driven methods into traditional policy-making processes marks a pivotal shift in how policies are conceptualised, implemented, and refined. Traditionally, policy decisions have heavily relied on qualitative analysis and expert judgement. While these approaches remain valuable, they can be significantly strengthened by incorporating quantitative data analysis, which offers enhanced rigour and precision. According to Sanderson (2002), data-driven methods allow policymakers to ground their decisions in empirical evidence, thereby minimising biases and subjective interpretations that often accompany purely qualitative assessments.

Moreover, the transparency and accountability of the policy-making process are greatly enhanced through data-driven approaches. As Brunswicker et al. (2019) argue, the availability of concrete data enables policymakers to clearly justify their decisions, fostering trust among stakeholders and the general public. This traceability is particularly important in an era where public confidence in governmental decisions is frequently challenged, necessitating a clear and defensible rationale for policy actions.

Another critical advantage of integrating data-driven methods is the ability to make proactive policy adjustments. Continuous data collection and real-time analysis provide a dynamic feedback loop, enabling policymakers to identify emerging issues and optimise policy outcomes. Brunswicker et al. (2019) and van Veenstra and Kotterink (2017) highlight how this flexibility can lead to more responsive and adaptive governance, where policies evolve in tandem with changing circumstances, rather than lagging behind them.

Furthermore, data-driven approaches can facilitate broader stakeholder engagement by making complex data more accessible through advanced visualisation and presentation tools. When data is presented in an understandable format, it empowers stakeholders to provide more informed feedback. This inclusive approach not only democratises the policy-making process but also leads to more comprehensive and accepted policy outcomes. By engaging a wider range of voices, from citizens to experts, data-driven policy design can enhance the legitimacy and acceptance of policies, ultimately contributing to more effective governance.

In light of these benefits, there is a growing corpus of research on data-driven policy design that underscores its transformative potential (Brunswicker et al., 2019; van Veenstra and Kotterink, 2017). Integrating these insights into traditional policy-making frameworks may offer a pathway to more innovative, effective, and equitable policy outcomes. Note that expert knowledge and data-driven methods complement each other at each stage of the policy design cycle, enhancing the robustness, effectiveness, and responsiveness of policies as discussed by Christensen (2021) and Bolger and Wright (2017):

1. **Agenda Setting**: Expert knowledge identifies and prioritises emerging issues based on theoretical and practical insights. Data-driven methods support this by revealing patterns and trends through trend analysis and quantitative evidence. Together, they ensure the policy agenda is both relevant and evidence-based.

2. **Policy Formulation**: Experts design policy options and assess their feasibility, drawing on their specialised knowledge. Data-driven methods can simulate potential outcomes and quantify costs and benefits. Experts interpret these results to refine policy solutions, ensuring practicality and effectiveness.

3. **Decision Making**: Experts provide contextual understanding and balance competing interests during decision-making. Data-driven methods enhance this with decision support systems and scenario analysis, presenting possible outcomes. Together, they guide policymakers to make informed, balanced decisions.

4. **Implementation**: Experts offer technical support and training during policy implementation. Data-driven methods provide monitoring systems and real-time performance metrics, tracking progress and effectiveness. This feedback allows experts to make timely adjustments, ensuring successful implementation.

5. **Evaluation**: Experts interpret evaluation results and recommend policy improvements. Data-driven methods conduct impact and comparative analyses, measuring outcomes against benchmarks. Combining these insights ensures evidence-based recommendations for refining policies.

6. **Policy Maintenance and Termination**: Experts provide ongoing advice on maintaining, adjusting, or terminating policies based on new insights. Data-driven methods continuously monitor performance and analyse trends. This combination ensures policies remain relevant and effective, adjusting as needed.

7. **Policy Learning and Innovation**: Experts disseminate best practices and drive innovation using theoretical insights and practical experience. Data-driven methods

facilitate data sharing and identify opportunities for innovation. Integrating both fosters continuous learning and improvement in policies (Trott et al., 2021).

3.3 Proposed Data-Driven Frameworks

This PhD project will address the challenging task of developing a methodology for using machine learning techniques to validate the objectives of a policy intervention and find optimal implementation of policy commitments. To fill the identified gaps and answer the research questions of this PhD project, three frameworks are developed to address three research questions related to transport policy. The term 'framework' is used since the approaches presented in this chapter are general and flexible, and can be applied to different policy objectives and commitments.

The first research question mentioned above aims to identify datasets relevant to transport policy objectives. To accomplish this, feature importance and classification machine learning techniques are employed. The framework is discussed in detail in Section 3.5 and is employed in Chapter 4 for clean air zone case study. It is based on approaches and concepts from the Machine Learning literature. This framework facilitates the computation of correlations between different datasets and target variables using available data and statistical methods.

The second framework is designed to validate the objectives of policy interventions. Time series machine learning techniques are used to develop this framework, which is described in Section 3.6 and is employed in Chapter 5 for clean air zone case study.

The third framework aims to implement policy commitments in transport systems using optimisation multi-objective methods. This framework is elaborated in detail in Section 3.7 and is employed in Chapter 6 on the case study of the expansion of the electric vehicle charging infrastructure.

Table 3.1 gives an overview of mapping the designed frameworks to the research objectives and the technical chapters of this thesis.

3.4 Integration of Data-Driven Frameworks with the Policy Cycle

Figure 3.2 shows the integration of the data-driven frameworks proposed in this thesis with the policy cycle. Each framework is aligned with specific steps in the process of policy design and implementation as follows:

Table 3.1 Overview of mapping the designed frameworks to the research objectives and the technical chapters

Chapter	Chapter 4	Chapter 5	Chapter 6
Research Questions (RQ)	**RQ1.** Given the large volume of data gathered from the transport network, what data types are relevant to the objectives of a policy intervention?	**RQ2.** What machine learning techniques are suitable for combining datasets, processing the data, and validating the objective of a policy intervention?	**RQ3.** Could data-driven techniques be used for efficient optimal implementation of policy commitments?
Objectives	To Investigate the possibility of selecting important datasets related to proposed policy objective in transport system using data-driven approach.	To investigate the possibility of proposed policy validation before implementing them using data-driven approach.	To investigate the implementation of policy commitment using data-driven approach.
Nature of Study	Quantitative	Quantitative	Quantitative
Methods	Machine Learning Classification	Machine Learning Time Series	Data-Driven Multi-Objective Optimisation
Data Collected	Air Quality, Weather, Time Stamps, Vehicles	Air Quality, Weather, Time Stamps, Vehicles	Car Ownership, Vehicle Availability, Power Profile, Trip Statistics, Electric Vehicle Efficiency, Electric Vehicle Battery Size.
Data Pre-Processing	Collecting, Cleaning, Analysing	Collecting, Cleaning, Analysing	Collecting, Cleaning, Analysing
Programming Language	Python, Libraries (NumPy, Scikit-learn, Pandas, Keras, Matplotlib)	Python, Libraries (NumPy, Scikit-learn, Pandas, Keras, Matplotlib, TorchMetrics)	Python, Libraries (Pandas, NumPy, Math, Matplotlib, GDAL)

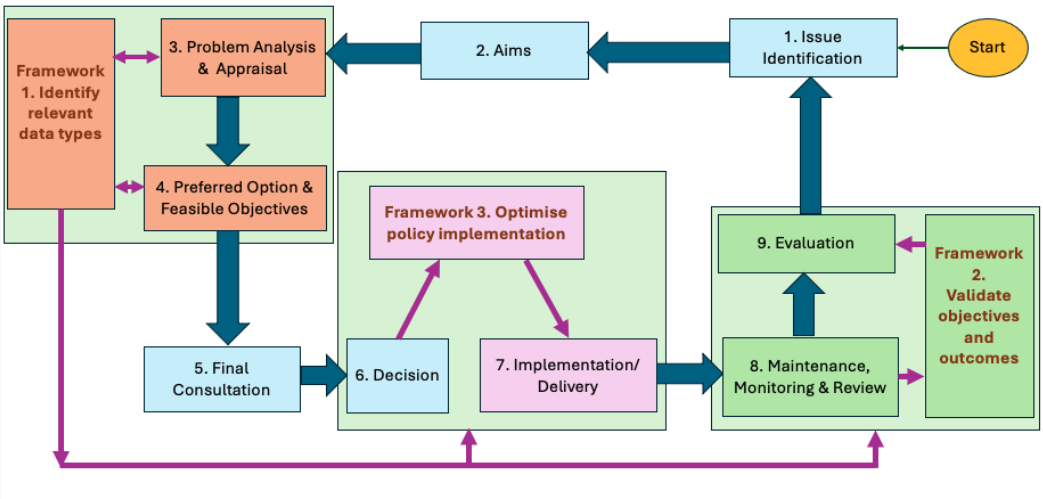

Fig. 3.2 Integration of the data-driven frameworks proposed in this thesis with the policy cycle

- **Framework 1 (Data Analysis):** Receives inputs from steps 3 and 4 of the policy cycle on Problem Analysis, Appraisal, and feasible objectives. Framework 1 focuses on identifying relevant data types and using machine learning techniques to analyse multi-modal datasets. This helps in understanding the problem context and generating feasible options for addressing the issue. By using techniques such as feature selection and importance ranking, policymakers can focus on the most critical data points that influence the problem, thereby enhancing the accuracy and relevance of the analysis. The framework also provides inputs to steps 7,8,9 on implementation, monitoring, and evaluation.

- **Framework 2 (Validation):** Integrated at the Maintenance, Monitoring & Review and Evaluation stages, this framework validates the objectives of policy interventions using historical and new data. It ensures continuous monitoring and assessment, helping to make necessary adjustments and ensuring the policy remains effective.

- **Framework 3 (Optimisation):** Integrated between the Decision stage and the Implementation/Delivery stage, this framework uses simulation and multi-objective optimisation methods to find the most effective approach for policy implementation. It ensures efficient resource allocation and maximises the impact of the policy commitments.

By aligning the data-driven frameworks proposed in this thesis with specific steps in the policy cycle, the proposed frameworks enhance traditional policy-making methods, ensuring that policies are based on robust evidence, continuously monitored, and adaptively improved.

While Framework 2 on validating policy objectives is integrated with steps 8 and 9 on monitoring and evaluation, it could also be part of the initial policy creation (steps 3 and 4) or the policy refinement (next iteration of the policy cycle). The main difference is on the availability of data and which dataset needs to be used in the framework. In the stages of post implementation review and policy refinement, datasets from the transport system in both cases of without policy implementation and with policy implementation are available, and the framework will be applied to these datasets without raising any assumption. In the stages of initial policy creation, only dataset from the transport system without policy implementation is available. In this case, assumptions and plausible scenarios need to be developed to build the most likely datasets for the transport system under the policy implementation. Further details will be provided in Section 3.6.

More generally, data analysis can contributes to the step 1 of the policy cycle on Issue Identification by highlighting trends, identifying emerging issues, and providing evidence to support the need for policy action. By analysing large datasets, policymakers can uncover patterns and correlations that may not be immediately apparent through traditional analysis methods. For instance, machine learning algorithms can sift through vast amounts of data to identify underlying factors contributing to a problem, thereby enabling a more precise definition of the issue. Data analysis can contribute to the step 2 of the policy cycle in setting realistic and measurable aims based on data insights and predictive analytics. By leveraging historical data and predictive models, policymakers can set targets that are both ambitious and achievable.

Data analysis and simulation models can contribute to step 4 of the policy cycle by helping to determine the feasibility of objectives and predicting the outcomes of the preferred option. By running simulations and scenario analyses, policymakers can anticipate potential impacts and identify any risks or challenges associated with the preferred option. This enables a more informed decision-making process and helps in setting realistic and achievable objectives. Data visualisation and presentation tools can be used in step 5 on Final Consultation to effectively communicate the policy options and their implications to stakeholders.

This section presented a comprehensive and transferable framework for the application of data-driven approaches to policy design and evaluation. The integration of data-driven methods with the traditional policy cycle was discussed to enhance the overall policy-making process by providing a more robust, systematic, and evidence-based foundation for decision-making. In the rest of this chapter, the details of the proposed data-driven frameworks will be provided.

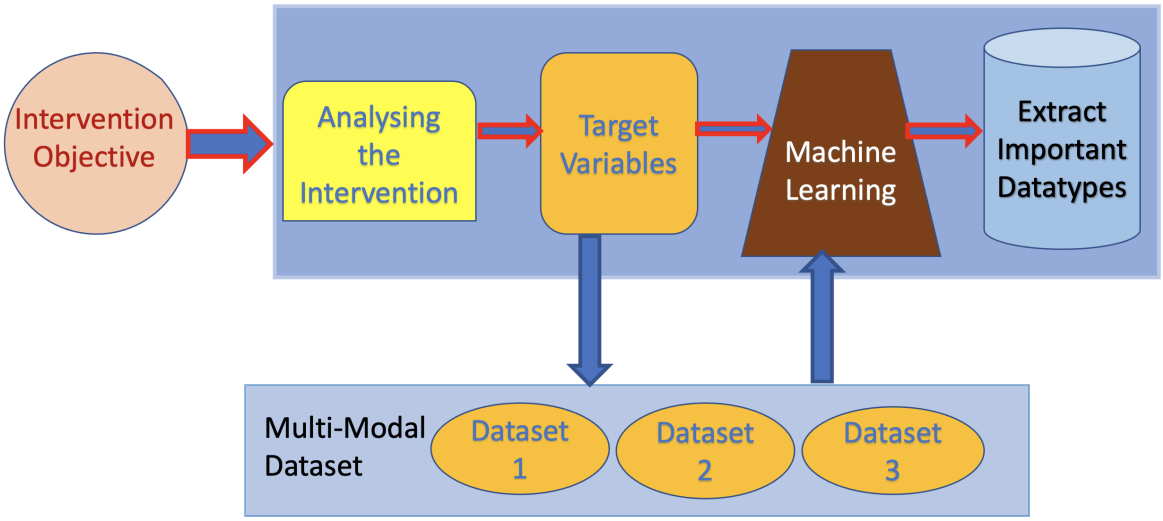

Fig. 3.3 A framework for identifying data types that are relevant to a policy objective

3.5 Framework for Identifying Relevant Data Types

In this section, a framework is presented for identifying data types that are the most relevant to a policy objective. The rationale behind designing this framework is to use well-established machine learning methods that do not require an understanding of physical or chemical properties but need sufficiently rich datasets. This framework sets the steps that need to be taken in order to capture the complex nonlinear relationships between the measured quantities and target variables in machine learning models.

The framework is presented in Figure 3.3. The first step in this framework is to analyse a given policy objective to extract the variables that are important in assessing the successful implementation of the policy intervention. These variables are the ones mentioned explicitly in the intervention documents while specifying quantitatively how much they are expected to change after the implementation of the policy intervention. Once these target variables are identified, the next step is to analyse the data from these variables using machine learning methods to measure their relative importance in predicting and affecting the objective of the policy. This relative importance can then be used to select a subset of datasets that make accurate predictions for assessing the policy objective. In order to compute the relative importance of the data types and extract the important ones, two different data-driven methods are used. The first one is based on *Pearson correlation coefficient*. The correlation takes values in the interval $[-1, 1]$, a value of ± 1 indicates that two variables are dependent linearly, and values closer to 0 means they become independent of each other. A plus sign shows positive relationship (i.e., increase in one variable leads to increase in the other variable) and a

minus sign shows a negative relationship (i.e., increase in one variable leads to decrease in the other variable). The second one is based on *permutation feature importance*, which computes the importance of a feature by first training the model on the train dataset, then permuting the feature and computing the increase in the prediction error of the model. If a feature is important in making predictions, the prediction error should increase after permutation. If a feature is unimportant, the change in the prediction error will be by doing the permutation. In general, feature importance can take any values. The feature importance is normalised to obtain values in the range $[0, 1]$, which show the relative importance of the features.

In this framework, possible datasets must first be identified manually by experts, then reviewed by the developed framework. It is definitely desirable to automate the process of reviewing all possible datasets related to the policy objective rather than an initial manual shortlist of such datasets. This is currently not possible due to the data barriers and concerns described in a report by the Department for Transport (Northhighland worldwide consulting, 2018)[3] and were also encountered extensively in the initial stage of this PhD research. These barriers and concerns about data sharing can be divided into the three categories of external barriers, internal barriers, and cultural barriers (Catapult Transport Systems, 2017):

- External barriers: The organisations are fearful about data sharing because of GDPR concerns. They are worried for breaking the security and safety.
- Internal Barriers: The organisations are worried about cost of open data and data sharing. It may be because of lack of case studies to show them the benefit of open data.
- Cultural barriers: The organisations do not have the right set of skills for standardisation, keeping and maintenance of open data.

In case datasets are used for developing sensitive policy options, such barriers may be created intentionally to preserve confidentiality of the results. Massive data are currently collected from various sources by different public agencies and private sectors which rarely communicate with each other.

The identification of relevant datasets and the effective sharing of data between organisations are critical challenges that require more than just technical solutions. Expert knowledge can navigate these complexities by providing the necessary contextual understanding and strategic insights. Experts can identify the most relevant datasets for a given policy by leveraging their deep understanding of the transport system, its dynamics, and the specific objectives

[3] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730787/local-transport-data-summary.pdf

of the policy intervention. Furthermore, they can facilitate data sharing by addressing legal, ethical, and organisational concerns, such as data privacy, security, and proprietary interests. By advocating for collaboration, experts help build trust between organisations, ensuring that data is shared in a manner that is both secure and aligned with policy goals. This combination of expert knowledge and strategic data management is essential for overcoming barriers and ensuring that data-driven approaches can be effectively applied in transport policy-making.

3.6 Framework for Validating the Objectives of Policy Interventions

In this section, a framework is developed for checking how well the objectives of a policy intervention are achieved. This framework is presented in Figure 3.4. The underlying idea of this framework is to compare the behaviour of the transport system under study in the two cases of with and without the policy intervention. Historical data and new data gathered after applying a policy intervention can be used to construct machine learning models and then make the comparison. This comparison will then be judged against the quantitative objectives of the policy intervention. Since the validation could be part of the initial policy creation, policy refinement, or a post implementation review, two distinguished scenarios can be considered.

First Scenario: The policy intervention is not implemented in the real system yet. In this case, only historical data is available. The target variables and data types relevant to the policy is first obtained using the framework of Figure 3.3. Then, a machine learning model can be trained on the historical data to predict the target variables in the future without the application of the policy intervention (i.e., assuming that no policy intervention is applied, what will happen in the future). It is also essential to understand what will happen if a policy intervention is implemented. For this purpose, various techniques can be used including: multi-agent simulation (Doniec et al., 2008), physics-based modelling (Pielke and Uliasz, 1998; Rastgoftar and Jeannin, 2021), machine learning models (Suleiman et al., 2019), or a combination of these approaches. For example, in air quality modelling, physical meteorological models could be used that include Dispersion Models, Photochemical Models, or Receptor Models (See for example (Pielke and Uliasz, 1998)). These mathematical models are based on the natural behaviour of physical quantities (concentration of the gas/particle, pressure, temperature, etc.), are time consuming to construct, and require iterative tuning of their parameters based on measured data.

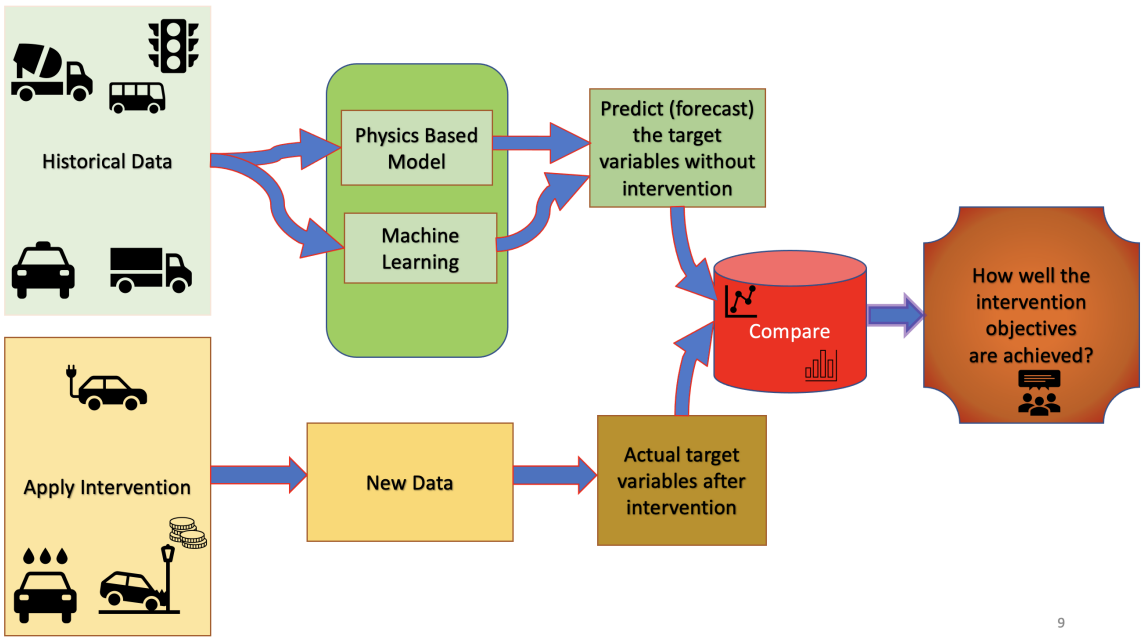

Fig. 3.4 A framework for validating the objectives of a policy intervention and checking how well the objectives are achieved

The underlying principle of these prediction approaches is to raise assumptions on how the policy intervention would affect various features in the system and subsequently make predictions on the target variables. In Chapter 5, the historical dataset is modified based on appropriately justified assumptions, and a second machine learning model is trained for predicting target variables after the application of the policy intervention.

Second Scenario: The policy intervention is already implemented. In this case, data gathered and stored can be divided into two parts: historical data from the transport system before the implementation of the policy intervention, and the most recent data from the system after the implementation of the policy intervention. The target variables and data types relevant to the policy is first obtained using the framework of Figure 3.3. Next, a machine learning model can be trained on the historical data to predict the target variables in the future without the application of the policy intervention (i.e., assuming that no policy intervention was applied, what would have happened in the future). Then, the predicted target variables under no policy intervention are compared with the target variables measured under the policy intervention. The new data obtained under policy intervention could also be used for improving the quality of the models built in the previous scenario, for instance by a better tuning of the model parameters.

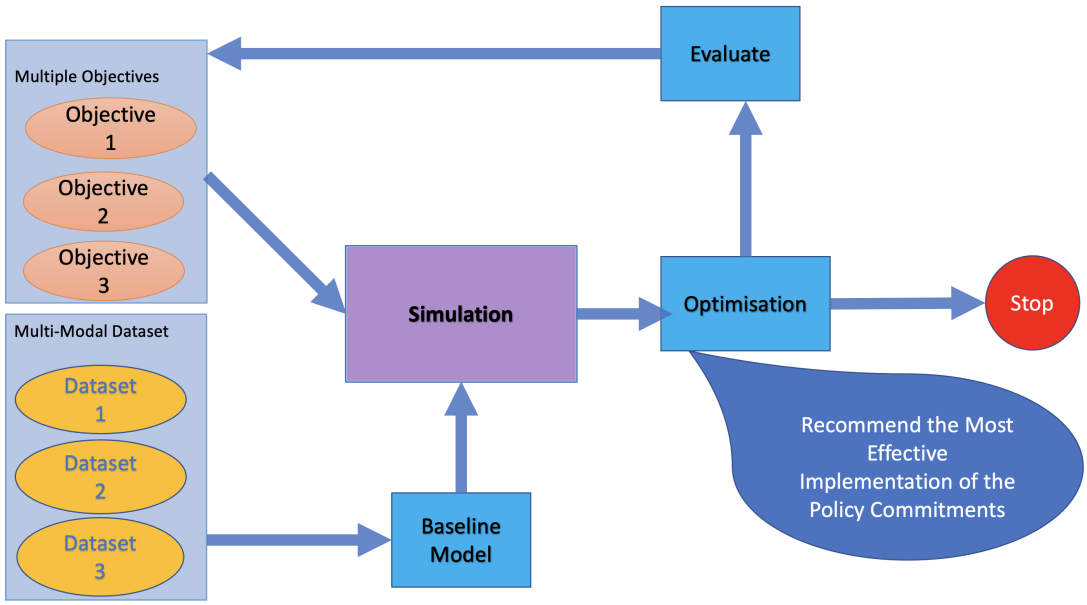

Fig. 3.5 Framework for finding the best implementation of policy commitments in transport systems

3.7 Framework for Optimal Implementation of Policy Commitments

This section presents the designed data-driven framework for finding the best implementation of policy commitments in transport systems. The framework is shown in Figure 3.5. At the heart of this framework is a model that generates the behaviour of the transport network when a policy commitment is implemented. This could be an agent-based model, an implementation of the physics-based model, digital twin of the transport network, or a machine learning model trained properly on a related dataset of the transport system. These models should describe the behaviour of the system if a specific policy commitment is applied to the system. These models are generally constructed starting from a baseline model. The baseline model describes the current situation in the network and it should match the datasets already available from the system. The model is then used in a loop to evaluate different implementations of policy commitments and find the best implementation using optimisation methods.

In order to demonstrate the application of the framework in Figure 3.5 to a case study, a key policy in the transport sector of net zero emission strategy has been selected. The policy will be achieved through various commitments including incentives for individuals and companies to buy electric vehicles, and providing the required EV charging infrastructure. In

order to achieve these commitments, the required number of electric vehicle charging points and stations will be increased in on-street and public places to create a better availability and accessibility of the charging points for the EVs.

A model is built for simulating the EV charging infrastructure to compute the increase in the quantity of charging points with different types. The model includes two distinctive stages of simulation and optimisation. The simulation stage of the baseline model has been constructed in collaboration with the industrial partner of the PhD project, Arup Group Limited (Arup and Limited, 2022). The assumptions used for the construction of this baseline simulation model have been selected and revised according to the available data. The relevant subsets of the output of the simulation model have been used for feeding the optimisation stage of the model.

3.8 Programming Language for the Implementations

For the implementation of the methodology proposed in this thesis, analysing the data, and applying the frameworks to case studies, Python is chosen as the primary programming language due to its extensive support for data analysis and machine learning. Python's simplicity and readability streamline the coding process and its widespread use in the machine learning community provides a robust support system for problem-solving and development. When compared to R, MATLAB, or other popular programming languages for data-driven approaches, Python stands out for several reasons:

- **Flexibility**: Python's extensive library ecosystem, such as NumPy for numerical computing, Pandas for data manipulation, and scikit-learn for machine learning, provides versatile tools that are essential for implementing the methodology of this research and applying to the selected case studies.
- **Community and Support**: Python has a large and active community, providing a wealth of resources, documentation, and forums for troubleshooting, which is invaluable for research and development.
- **Integration**: Python offers excellent integration capabilities, allowing it to work seamlessly with other programming languages and technologies.
- **Open Source**: Unlike MATLAB, Python is open-source, which eliminates the barriers of licensing costs and restrictions, promoting a more collaborative and accessible research environment.

Table 3.2 Advantages of Python programming language used in the thesis

Advantage of Python	Description
Flexibility	Extensive library ecosystem (NumPy, Pandas, scikit-learn)
Community and Support	Large and active community with extensive resources
Integration	Possibility of integration with other languages and technologies
Open Source	No licensing costs, collaborative and accessible

The above reasons are summarised in Table 3.2. Specific libraries selected for their respective strengths and relevance to each chapter of this thesis, summarised in Table 3.3, are as follows:

- **NumPy and SciPy** are used for their efficient numerical computations, which are indispensable for data pre-processing and complex mathematical operations.
- **Pandas** is chosen for its data structures and tools that are designed for quick and easy data manipulation and analysis, particularly when dealing with structured data.
- **Scikit-learn** is employed for its extensive range of machine learning algorithms, providing a comprehensive toolkit for regression and classification tasks.
- **Keras** allows for the construction and training of machine learning models including neural networks with a user-friendly interface for quick prototyping.
- **Matplotlib and TorchMetrics** are incorporated for data visualisation, which aids in the analysis and presentation of results.
- **GDAL** is particularly relevant for spatial data manipulation and analysis, which aligns with the transport and geospatial aspects of this research used for obtaining the results of Chapter 6.

A summary of using these libraries in each technical chapter of this thesis are reported in Table 3.1. Each library's use in this thesis aligns with its capabilities and suitability for the research tasks at hand. This strategic choice of programming language and libraries ensures that the developed methodology is not only efficient and effective but also grounded in the latest technology standards in data science and machine learning.

3.9 Conclusions

This chapter discussed the current state of practice in evidence-based policy-making and the impact of data-driven methods on the policy-making process. Building on this founda-

Table 3.3 Advantages of specific Python libraries used in the thesis

Python Library	Usage in Thesis
NumPy and SciPy	Numerical computations and data pre-processing
Pandas	Data manipulation and analysis
Scikit-learn	Machine learning algorithms for regression, classification
Keras	neural network construction
Matplotlib and TorchMetrics	Data visualisation
GDAL	Spatial data manipulation and analysis

tion, three data-driven frameworks were introduced, each designed to address the research questions of this thesis. These data-driven frameworks include

(a) framework for identifying relevant data types,

(b) framework for validating the objectives of policy interventions, and

(c) framework for optimal implementation of policy commitments.

The chapter illustrated how each framework supports different stages of the policy development process for a generic policy. The integration of the frameworks within the policy cycle was visually represented in the accompanying diagram, which showed

- Framework 1 (Data Analysis) aids in Problem Analysis & Appraisal by identifying relevant data types.
- Framework 2 (Validation) ensures Maintenance, Monitoring & Review and Evaluation by validating objectives and outcomes.
- Framework 3 (Optimisation) supports Implementation & Delivery by optimising policy implementation.

It was discussed that unlike traditional policy-making, which relies heavily on expert judgement and qualitative analysis, the proposed frameworks integrate systematic and quantitative data-driven methods. This approach enhances the robustness and objectivity of the policy-making process.

Finally, the choice of programming language for implementing the methodology and applying the frameworks to case studies in subsequent chapters were presented. Next technical chapters will show how to apply the frameworks of this chapter to the case studies while reporting the obtained results. This will be followed by a discussion chapter before concluding the thesis.

Chapter 4

Machine Learning & Relevant Data-Types for Clean Air Zone Policy

A framework was proposed in the previous chapter for identifying data types that are the most relevant to a policy objective. This chapter shows how to apply the framework to the use case of a clean air zone where the objective is to improve air quality in areas where there are currently nitrogen dioxide (NO_2) exceedances. The intervention implemented in the clean air zone is considered to be in the form of specific charges on vehicles entering the zone dependent on their emission levels. The datasets from the Newcastle Urban Observatory are used.

This chapter tackles the challenge of finding datasets that are relevant to the policy objective. Focusing on the reduction of NO_2 concentrations, different machine learning algorithms are used to build models. The rationale behind using the framework on the use case is that well-established machine learning methods do not require an understanding of air pollutants' physical or chemical properties but need sufficiently rich datasets. These are datasets that have a comprehensive range of features and examples necessary to train effective models. These datasets must be adequately large and diverse to capture the complexities of the quantities they aim to model or predict. The designed framework sets the first steps that need to be taken in order to capture the complex nonlinear relationships between the concentration of air pollutants and meteorological variables in machine learning models.

As part of the framework, suitable machine learning classification models are chosen. The selected models are Decision Tree, Light Gradient Boosting Machine, K-Nearest Neighbour, and Gradient Boosted Decision Tree. The details of these models were presented in Section 2.4.6. The test metrics for comparing the accuracy of machine learning algorithms employed in this research are also given, which can be used to assess the performance of the learned models.

This chapter is organised into the following sections. With focus on the clean air zone and data use, Section 4.1 maps the generic steps of the policy cycle described in Section 3.1 to the steps taken in defining and evaluating the clean air zone intervention. A description of data collection and preparation is provided in Section 4.2 for the geographical area of Newcastle upon Tyne and for the clean air zone case study. Section 4.3 is focusing on the data analysis providing the preprocessing steps performed to prepare and clean the raw data before it is fed into a machine learning model. It also gives the details of applying statistical methods for exploration, analysis, and visualisation of the data gathered from Newcastle Urban Observatory for air quality. Section 4.4 gives the policy objective for the use case of clean air zone considered in this chapter. Sections 4.5-4.6 gives the complete definitions of Pearson Correlation and Feature Importance needed as part of the framework. The test metrics for assessing the performance of the machine learning models are given in Section 4.7. These metrics include *accuracy*, *precision*, *recall*, *F1 score*, and *confusion matrix*. Section 4.8 gives the details of the architecture and implementation of the machine learning models, describes data splitting for model training and evaluation, and presents the results of applying the framework on the dataset from Newcastle Urban Observatory. Finally, the chapter is concluded in Section 4.9. The code for reproducing the results of this chapter is provided in Appendix B.

The content of this chapter is based on the following articles published during the PhD study.

- Farhadi, Farzaneh, Roberto Palacin, and Phil Blythe. "Machine Learning for Transport Policy Interventions on Air Quality." IEEE Access (2023). DOI: https://doi.org/10.1109/ACCESS.2023.3272662
- Farhadi, F., Palacin, R. and Blythe, P., 2022. Machine Learning Methods for Identifying Relevant Data in Transport Policy Interventions. Universities Transport Study Group (UTSG), Edinburgh Napier (July 2022).

4.1 Steps of the Policy Cycle for Clean Air Zone

This section describes how the generic steps of the policy cycle described in Section 3.1 are mapped into the steps taken in defining and evaluating the clean air zone intervention.

1. **Issue Identification:** Local authorities recognise that air quality levels, particularly NO_2, exceed legal limits, primarily due to road traffic emissions. Public health data indicates significant negative health impacts, such as respiratory and cardiovascular

illnesses, associated with poor air quality. The UK Government's "Air Quality Plan for NO_2 in UK" (2017) mandated cities with illegal NO_2 levels to take action. Newcastle was identified as having NO_2 levels above the legal limit, primarily due to road traffic emissions.
Data Use: Continuous air quality monitoring data from stations throughout the city is analysed to identify areas with the highest pollution levels. Epidemiological data linking air pollution to health outcomes is also reviewed to highlight the urgency of the issue. Newcastle City Council used air quality monitoring data to pinpoint problem areas like the Central Motorway and City Centre where NO_2 exceeded legal limits.

2. **Aims:** Authorities establish clear aims for the policy, such as reducing NO_2 levels to meet legal standards within a specified time frame, improving public health, and reducing vehicle emissions. The Air Quality Action Plan for Newcastle and Gateshead set the aim to reduce NO_2 levels to meet legal limits and improve public health.
Data Use: Baseline data on current NO_2 levels and health statistics is used to set specific, measurable targets (e.g., achieving a certain amount of reduction in NO_2 concentrations within a specific time frame).

3. **Problem Analysis and Appraisal:** A thorough analysis of the sources of NO_2 emissions (e.g., diesel vehicles) is conducted to understand the problem's scope. Different strategies are appraised, such as charging high-emission vehicles, improving public transport, or promoting active travel (walking, cycling).
Data Use: Traffic flow data, emission models, and cost-benefit analysis are used to assess the potential impact and feasibility of different options. Health impact assessments evaluate the effects of air quality improvements on public health outcomes. The "Clean Air Zone for Newcastle and Gateshead – Delivery Plan" (2021) provides a detailed analysis of air quality data, sources of emissions, and potential policy measures. The business case included emissions modelling and traffic analysis data to understand the sources of pollution. Various scenarios, such as charging different classes of vehicles or providing exemptions, were appraised for their effectiveness and feasibility.

4. **Preferred Option and Feasible Objectives:** Authorities identify the most viable policy option to achieve the aims set earlier. Newcastle opted for a charging clean air zone targeting non-compliant vehicles.
Data Use: Emissions modelling data is used to predict the impact of different options on air quality. Economic impact assessments help understand the costs and benefits to different stakeholders. These data inputs guide the selection of the preferred policy. The preferred option for Newcastle involved targeting vehicles that do not meet certain

standards. Emission reduction projections and cost-benefit analysis data supported this decision among the available options.

5. **Final Consultation:** Before finalising the policy, a consultation phase is conducted with the public, businesses, and other stakeholders to gather feedback and refine the policy.
   Data Use: Data collected during consultations (e.g., survey responses, public meetings, feedback submissions) is analysed to identify concerns, levels of support or opposition, and potential adjustments needed to the policy. The report on "Clean Air Zone Public Consultation" by Newcastle City Council [1] documented responses from over 19,000 participants through online surveys, public meetings, and stakeholder workshops. Data from this consultation was used to refine the policy, such as adjusting the types of vehicles charged and considering additional support measures for affected groups.

6. **Decision:** Local authorities, in collaboration with central government agencies, make a final decision on the clean air zone policy based on the consultation results, data analysis, and strategic objectives.
   Data Use: Aggregated data from consultation feedback, emissions modelling, and economic impact studies are used to justify and support the decision. Risk assessments ensure potential challenges are identified and managed. The details are documented in the "Minutes of Newcastle City CouncilâĂŹs Cabinet Meeting" (2020) where the decision to implement a clean air zone was finalised.

7. **Implementation and Delivery:** The policy is put into action, including establishing infrastructure (such as signage, cameras, and enforcement mechanisms), informing the public, and launching any supporting schemes including grants or exemptions for certain vehicles (Newcastle Clean Air Zone Implementation Plan, 2021).
   Data Use: Newcastle collects real-time data from monitoring systems (such as cameras) to enforce the policy and track compliance. Traffic flow and air quality data are continuously monitored to adjust implementation as needed.

8. **Maintenance, Monitoring, and Review:** Continuous monitoring of the clean air zone's impact on air quality, traffic patterns, and public health is conducted to ensure it remains effective. Authorities review the policy to make adjustments, such as modifying zone boundaries or updating vehicle exemptions.
   Data Use: Ongoing air quality monitoring, traffic data, public health statistics, and

[1] https://www.newcastle.gov.uk/citylife-news/transport/final-air-quality-consultation-under-way

compliance rates are analysed to assess whether the policy objectives are being met and if adjustments are necessary.

9. **Evaluation:** A comprehensive evaluation is conducted to determine whether the clean air zone has achieved its objectives, such as reducing NO_2 levels and improving health outcomes.
   Data Use: Evaluation involves comparing post-implementation air quality data against baseline levels to measure changes. Data on health outcomes, economic impacts, and compliance rates are also analysed to assess the policy's overall success and inform future policy decisions.

At each step, data is crucial for informing decisions, setting objectives, analysing options, engaging stakeholders, implementing policies, monitoring progress, evaluating outcomes, and refining the clean air zone policy throughout its life cycle.

This chapter is focused on applying the proposed framework for identifying relevant data types proposed in Chapter 3 to the Newcastle clean air zone. This framework can be integrated with steps 3 and 4 of the policy cycle in problem analysis, preferred option and feasible objectives as discussed in Fig 3.2.

4.2 Data Collection and Preparation

As part of the data search for applying the proposed framework to clean air zone, the relevant organisations were contacted. After extensive discussions with the contact persons of the related organisations, the datasets most relevant to the case studies were identified (a list of the contacted persons can be found in Appendix A). For the clean air zone case study in this Chapter and in Chapter 5, the primary data source is the Newcastle Urban Observatory, which has datasets of time-stamped quantities on air quality, weather, and traffic.

4.2.1 Geographical Area: Newcastle upon Tyne, United Kingdom

The city of Newcastle upon Tyne in the United Kingdom has been chosen as the geographical area for this research, which aims to validate the effectiveness of clean air zones. Newcastle upon Tyne is the largest city in the North East of England with approximate population of $300,000$. It is a densely populated urban area with a high level of air pollution, making it an ideal location for testing and implementing measures to improve air quality. Through this research, the impact of clean air zones on air quality can be studied in depth.

4.2.2 Data for Clean Air Zone

The Newcastle Urban Observatory (UO) was developed with a multi-million-pound investment to serve as a large-scale data capturing infrastructure. The UO was initially funded under the UK Collaboratorium for Research in Infrastructure and Cities (UKCRIC).[2] The open dataset of UO is used by community groups, local and national government, and research projects ranging from cyber-security to quantifying the impact of COVID measure and flood forecasting. The data handled by the UO covers a wide range of city metrics including mobility, air quality, climatic variables, and infrastructure.

The volume of data in UO is in the order of billions of data points that are published as anonymous open data (James et al., 2020). The dataset includes 900 million data points measured since 2016, 60 data types, and 2000 observations every minute. Figure 4.1 shows the geographical locations of sensors that measure and send data to the UO. The majority of the sensors are located in Newcastle upon Tyne and its surrounding areas. The map shows clusters of sensors for a better visualisation. Note that the measurements stored in the UO are raw data and needs to be processed for improving the quality of the data. A subset of the measurements relevant to the research of this thesis are used and preprocessed as described in Section 4.3.

4.3 Data Analysis

As highlighted in Figure 2.3, the pipeline of machine learning includes data preprocessing. The goal of preprocessing is to prepare and clean the raw data before it is fed into a machine learning model. Common objectives of preprocessing include

- Data Cleaning: Identifying and handling missing or erroneous data points to ensure the dataset is accurate and reliable.
- Normalisation/Standardisation: Scaling numerical variables to a standard range or standardising them. This helps in preventing variables with larger scales from dominating the learning process.
- Handling Categorical Data: Converting categorical variables into a numerical format that the model can understand.
- Removing Outliers: Visual inspection of data to remove extreme values or anomalies such as negative values for positive quantities and measurements outside the bounds of quantities.

[2] https://urbanobservatory.ac.uk/explore/ukcric

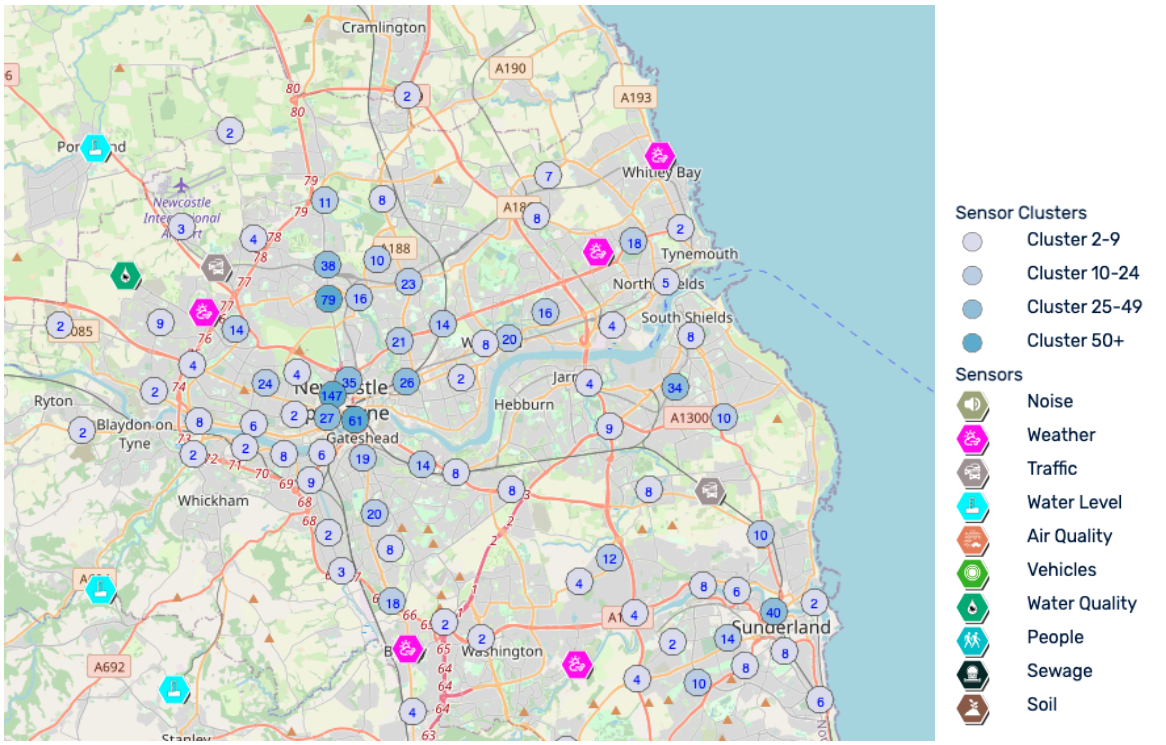

Fig. 4.1 Geographical locations of sensors that measure and send data to the Newcastle Urban Observatory

- Data Splitting: Dividing the dataset into training, validation, and test sets to evaluate the model's performance on unseen data.
- Handling Time-Series Data: Ensuring that the temporal aspect is considered appropriately, and features are engineered accordingly.

In the next subsection, preprocessing and analysing is performed on the datasets gathered from UO.

4.3.1 Preprocessing and Analysing Data from UO

The data being monitored and stored at UO sites include different themes and each theme has different variables. The themes considered for the clean air zone case study include

air quality, *weather*, *traffic*, and *timestamp*.

There are different sensors for monitoring these themes. These sensors are installed in various locations of Newcastle. The considered sensors are close to Newcastle City Centre to have a better understanding of the air quality at the centre of Newcastle. Each theme has different variables. The variables considered for the *air quality* theme include

Table 4.1 Data themes and variables for Clean Air Zone case study

Theme	Variables	Description	Source
Air Quality	CO, $PM_{2.5}$, PM_1, PM_{10}, PM_4, O_3, NO, NO_2	Air pollutants	Newcastle Urban Observatory
Weather	Wind Direction, Wind Speed, Solar Radiation, Solar Diffuse Radiation, Pressure, Rain Duration, Rain ACC, Max Wind Speed	Climatic variables	Newcastle Urban Observatory
Traffic	Traffic Flow, Average Speed	Traffic metrics	Newcastle Urban Observatory, Urban Traffic Management and Control
Timestamp	Year, Month, Day, Hour, Minute, Second	Time of data collection	Newcastle Urban Observatory

CO, $PM_{2.5}$, PM_1, PM_{10}, PM_4, O_3, NO, and NO_2.

The variables considered for *weather* theme include

Wind Direction, Wind Speed, Solar Radiation, Solar Diffuse Radiation, Pressure, Rain Duration, Rain ACC and Max Wind Speed.

The variables considered for the *traffic* theme include

Traffic Flow and Average Speed,

which are both collected by the Newcastle UO and by the Urban Traffic Management and Control at a greater fidelity. The dataset also has *timestamp* theme with variables

Year, Month, Day, Hour, Minute, and *Second.*

This means the measurements are taken in general every second and stored in the database of the UO. A summary of the above themes and variables can be found in Table 4.1.

Reducing NO_2 is the main objective of the policy intervention (cf. 4), thus the availability of NO_2 data in each year is considered. After analysing the NO_2 data gathered and stored by the Urban Observatory, the study is focused on the dataset for the year 2018 that has the largest number of measured values. The dataset of the year 2018 has over one million data entries, while the other years have a substantially smaller number of measurements. Therefore, the year 2018 is chosen for training and validating the policy objective. Table 4.2

shows the statistics of the dataset for the year 2018. The table shows respectively the mean, standard deviation, minimum, 25^{th} percentile (lower quartile), 50^{th} percentile (Median), 75^{th} percentile (upper quartile), and the maximum of each variable.

The preprocessing, computations, and visualisations of this chapter are done using Python programming language. The preprocessing takes into account measurements with obvious errors. For instance, any negative measurement of positive quantities is eliminated from the dataset. Any measurement outside the bounds of quantities are also eliminated (e.g., any outlier sensor reading that is a few times higher that other readings). Any data entry stored in the format of text and is supposed to be number is also eliminated.

The traffic flow measured and stored by Urban Observatory is the *total* the number of buses, coaches, private cars, taxis, vans and heavy goods vehicles. The clean air zone affects these types of vehicles differently. For instance, it is designed to put restrictions and charge commercial vehicles without affecting private cars. In order to make accurate predictions on how the clean air zone reduces the NO_2 concentrations, it is essential to have separate datasets for traffic flow of different vehicle types. The available dataset of the total traffic flow is divided into four different traffic flow for different vehicle types including

1. buses and coaches,
2. heavy goods vehicles (HGVs),
3. cars and taxis, and
4. two-wheeled motor vehicles.

Since the available dataset includes only the total number and does not give separate numbers for four vehicle types, the road traffic statistics published by the Department of Transport (Department for Transport, 2022) is used to get the *percentages* of each vehicle type in Newcastle upon Tyne. According to this report, the percentages are as follows:

- Traffic flow of buses and coaches = 1.24% of the total traffic flow,
- Traffic flow of HGVs = 18.13% of the total traffic flow,
- Traffic flow of cars and taxis = 80% of the total traffic flow,
- Traffic flow of two-wheeled motor vehicles = 0.63% of the total traffic flow.

Table 4.2 Statistics of the dataset from UO

index	mean	std	min	25%	50%	75%	max
O3 (ppb)	269.53	332.38	0.09	23.32	137.87	438.24	2111.90
PM_1 $(\mu g/m^3)$	3.65	4.01	0.18	1.20	2.15	4.43	44.06
$PM_{2.5}$ $(\mu g/m^3)$	6.87	7.86	0.37	2.48	4.11	8.12	92.28
PM_4 $(\mu g/m^3)$	9.99	8.46	0.18	4.78	7.59	12.00	134.13
PM_{10} $(\mu g/m^3)$	13.56	17.62	0.98	5.39	8.38	15.16	323.95
CO (ppm)	98.17	256.44	0.02	5.94	17.56	63.80	8091.09
NO (ppb)	58.99	38.44	7.29	31.07	49.88	77.10	247.87
NO_2 (ppb)	55.22	49.77	0.34	19.32	41.02	74.37	313.73
Rain Duration (h)	9.48	6.41	0.00	4.00	9.00	14.00	50.76
Rain Acc.	3.99	2.99	0.0	0.90	2.87	5.10	15.54
Solar Radiation (w/m^2)	74.54	140.12	0.014	0.24	2.19	92.85	9.4.55
Wind Direction	196.17	36.31	106.33	169.55	207.02	225.28	276.40
Max Wind Speed (m/h)	3.34	1.91	0.00	1.92	2.93	4.402	16.05
Solar Diffuse Radiation (w/m^2)	62.83	91.19	1.01	1.48	11.19	94.08	531.15
Wind Speed (m/h)	9.63	8.69	1.73	6.26	7.99	10.42	152.87
Pressure (mb)	1005.81	13.64	969.07	993.39	1007.22	1017.30	1033.12
Traffic Flow (Passenger Car Units)	12.35	11.19	0.00	1.15	10.08	22.14	44.05
Average Speed (KmPH)	57.16	14.65	26.17	44.64	51.93	72.35	80.00

4.3.2 Data Visualisation

The available datasets in 2018 are visualised to extract useful knowledge and find suitable information. Figures 4.2–4.3 show the time series of air pollutant concentrations (NO_2, NO, O_3, CO, PM_1, $PM_{2.5}$, PM_4, PM_{10}). Figures 4.4–4.5 show the time series of weather conditions (rain duration, solar radiation, solar diffuse radiation, rain accumulation, pressure, max wind speed, wind speed, and wind direction). Figure 4.6 show the time series of traffic metrics (traffic flow and average speed). The statistics of these variables are presented in Figures 4.7–4.9.

Data exploration reveals interesting insights on the trends of the variables. Higher levels of NO_2, CO, and PM are observed during the early months of the year, particularly in January and February. These high levels are likely due to winter heating and lower wind speeds, which reduce pollutant dispersion. As the year progresses, the pollutant levels decrease, reflecting seasonal changes and possibly better air quality measures. O_3 levels peak during the warmer months, showing how weather impacts air quality.

That pressure fluctuates throughout the year due to changing weather systems. Wind speed spikes during certain periods, indicating storms or strong weather events. Rain accumulation and duration increase during mid-year, aligning with the rainy season. Solar radiation peaks in the summer months and decreases in winter, demonstrating the seasonal changes in sunlight exposure.

The average speed graph shows that speeds are generally high but dip during peak hours and congested periods, indicating traffic jams. The traffic flow graph highlights higher vehicle volumes during rush hours, with significant peaks in the early morning and late afternoon, reflecting daily commuting patterns. There is an increase in traffic flow towards the end of the year, likely due to holiday activities.

After this data exploration and visualisation, the dataset is filtered to have the values of variables as time series. The dataset is also unified and transformed into an appropriate format to be compatible with machine learning methods.

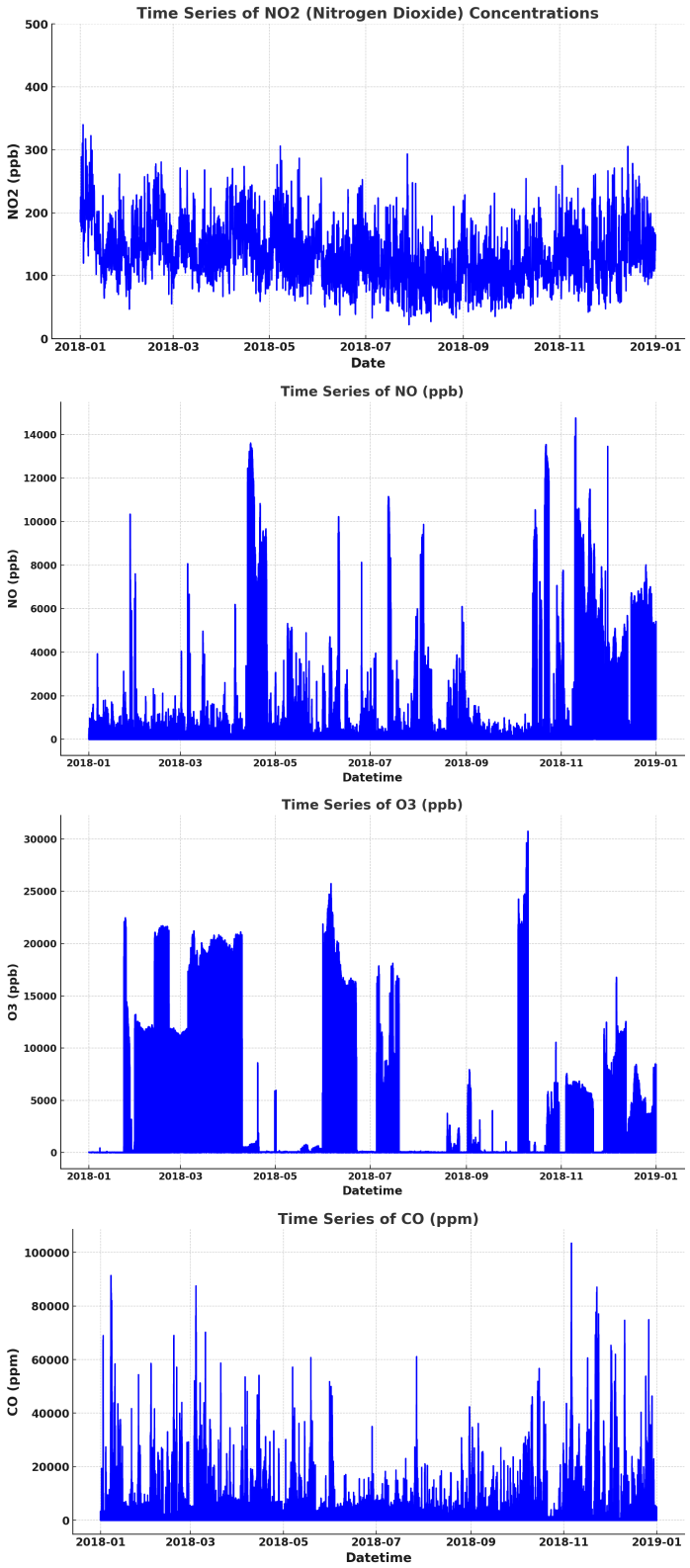

Fig. 4.2 Time series of air pollutant concentrations (NO_2, NO, O_3, and CO)

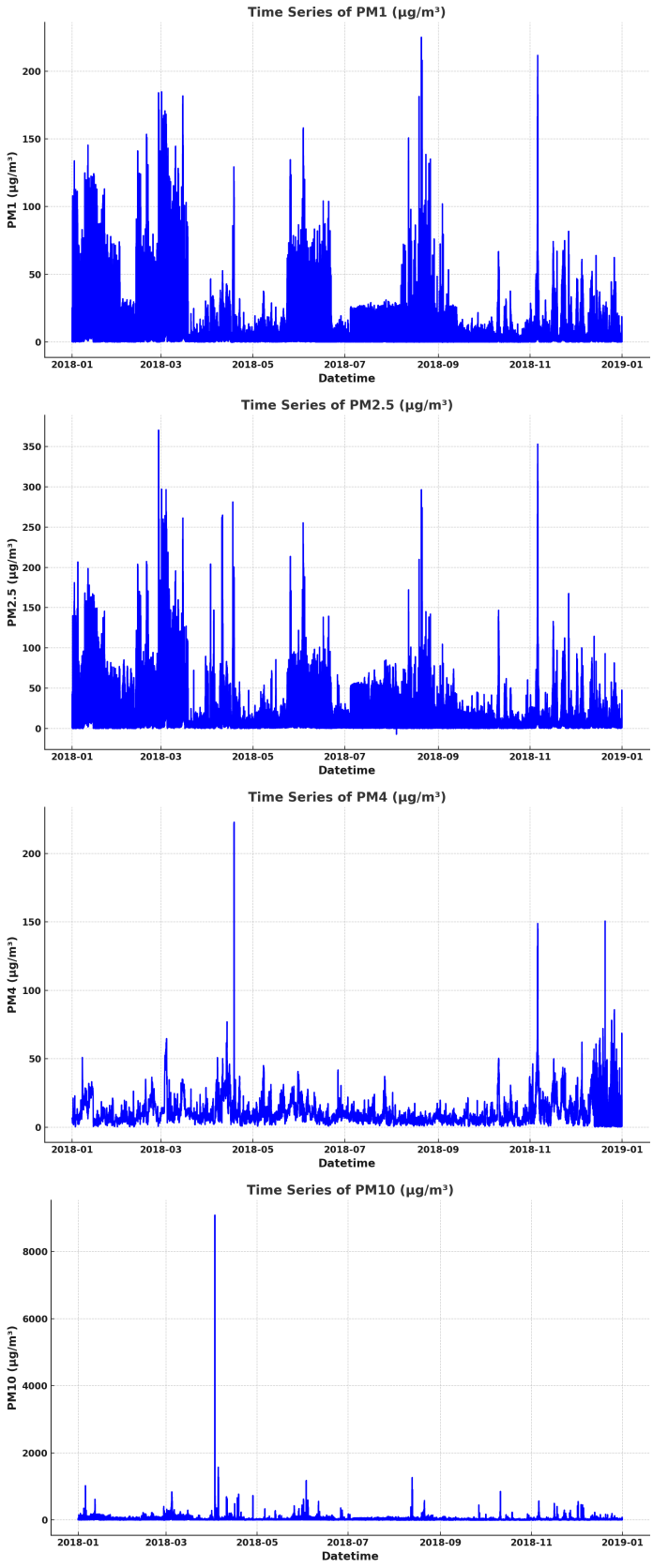

Fig. 4.3 Time series of air pollutant concentrations (PM_1, $PM_{2.5}$, PM_4, PM_{10})

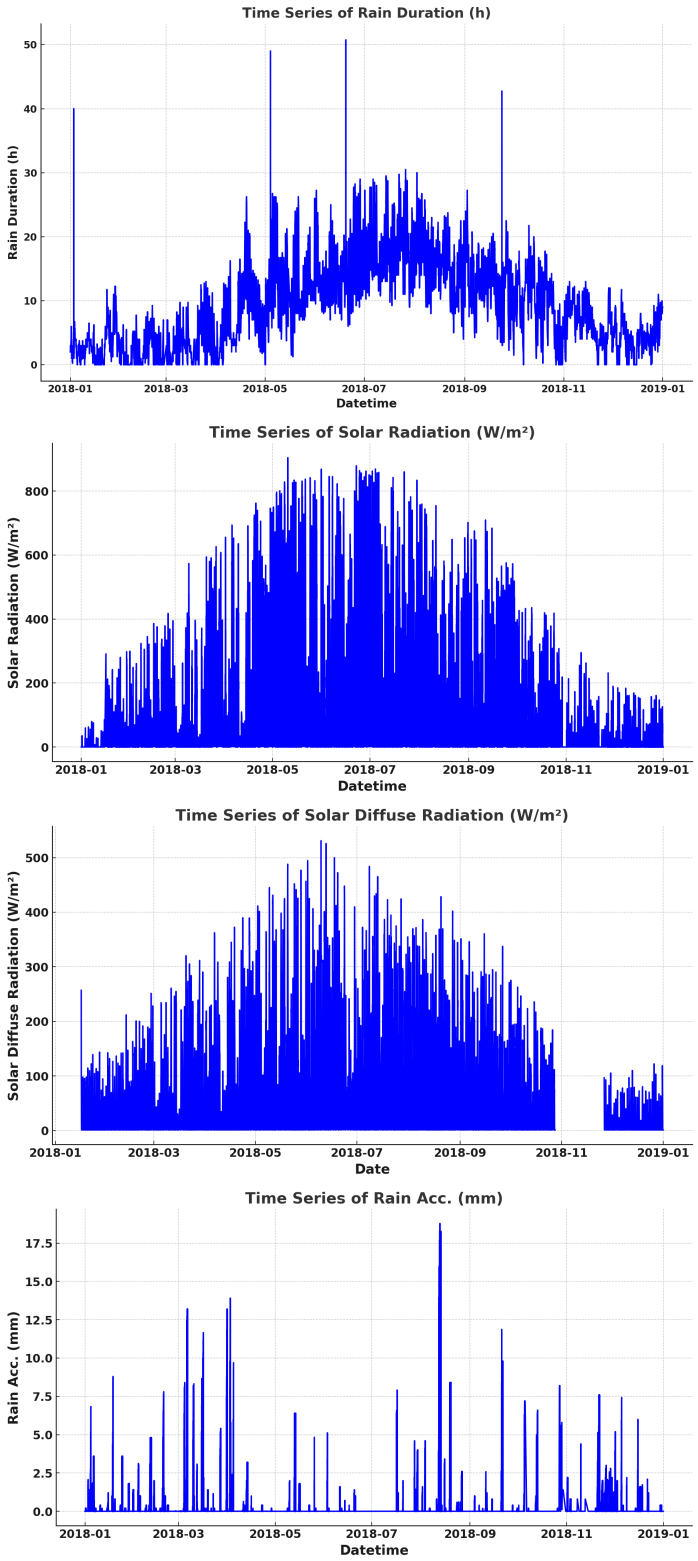

Fig. 4.4 Time series of weather conditions (rain duration, solar radiation, solar diffuse radiation, and rain accumulation)

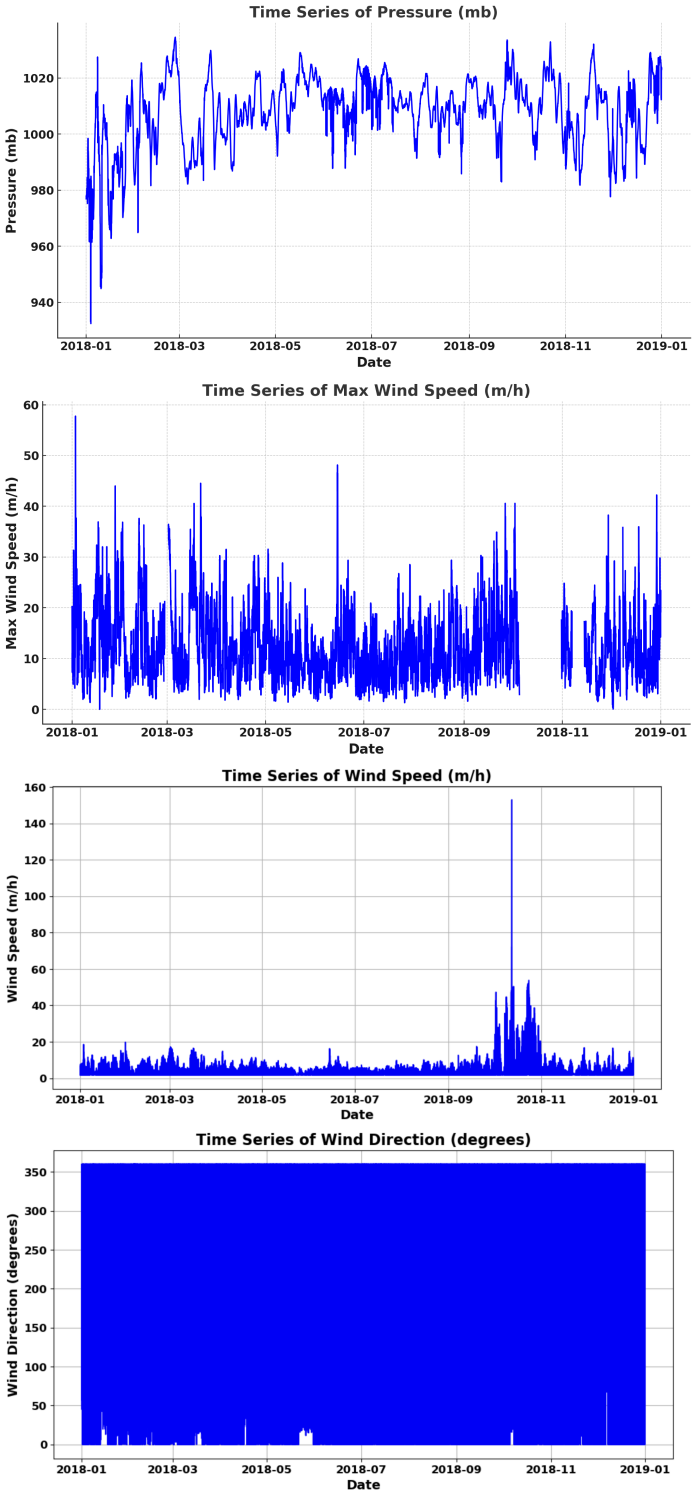

Fig. 4.5 Time series of weather conditions (pressure, max wind speed, wind speed, and wind direction)

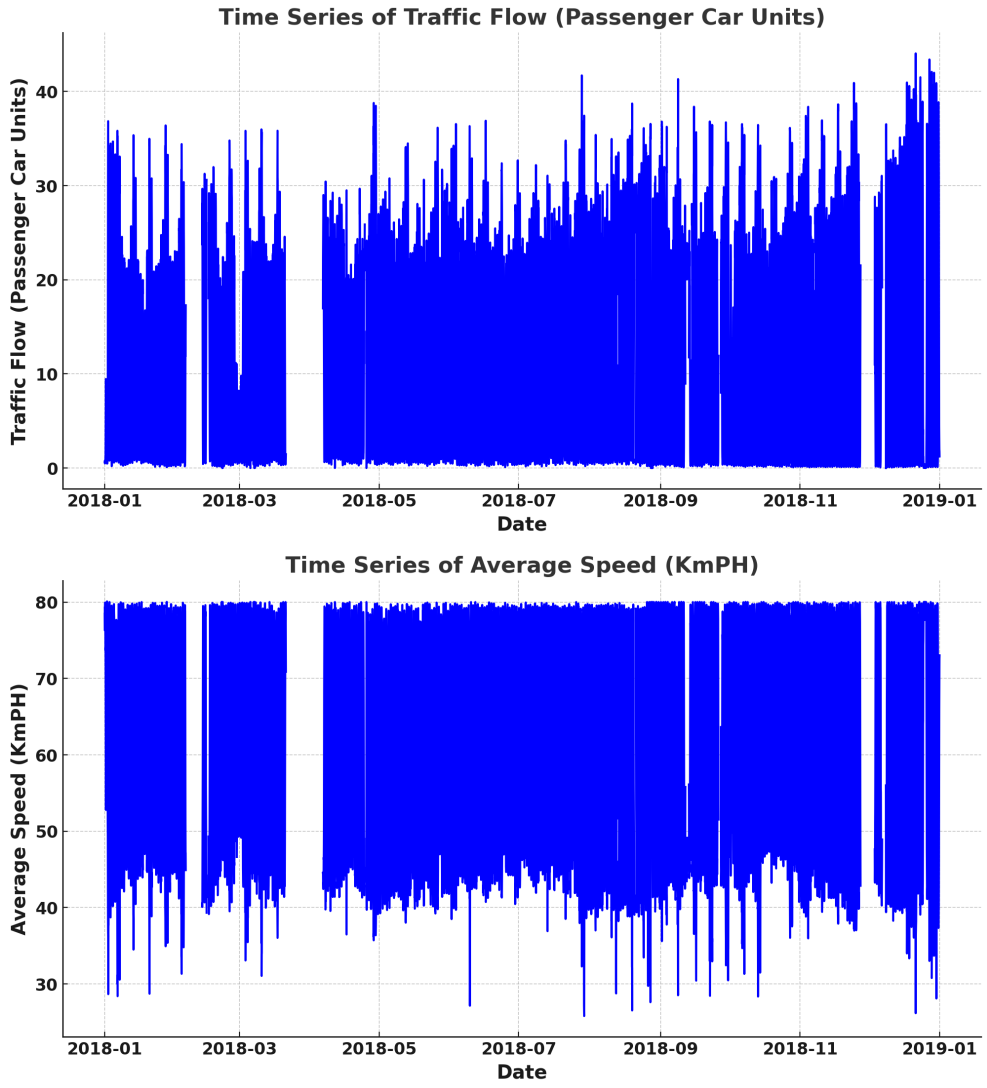

Fig. 4.6 Time series of traffic metrics (traffic flow and average speed)

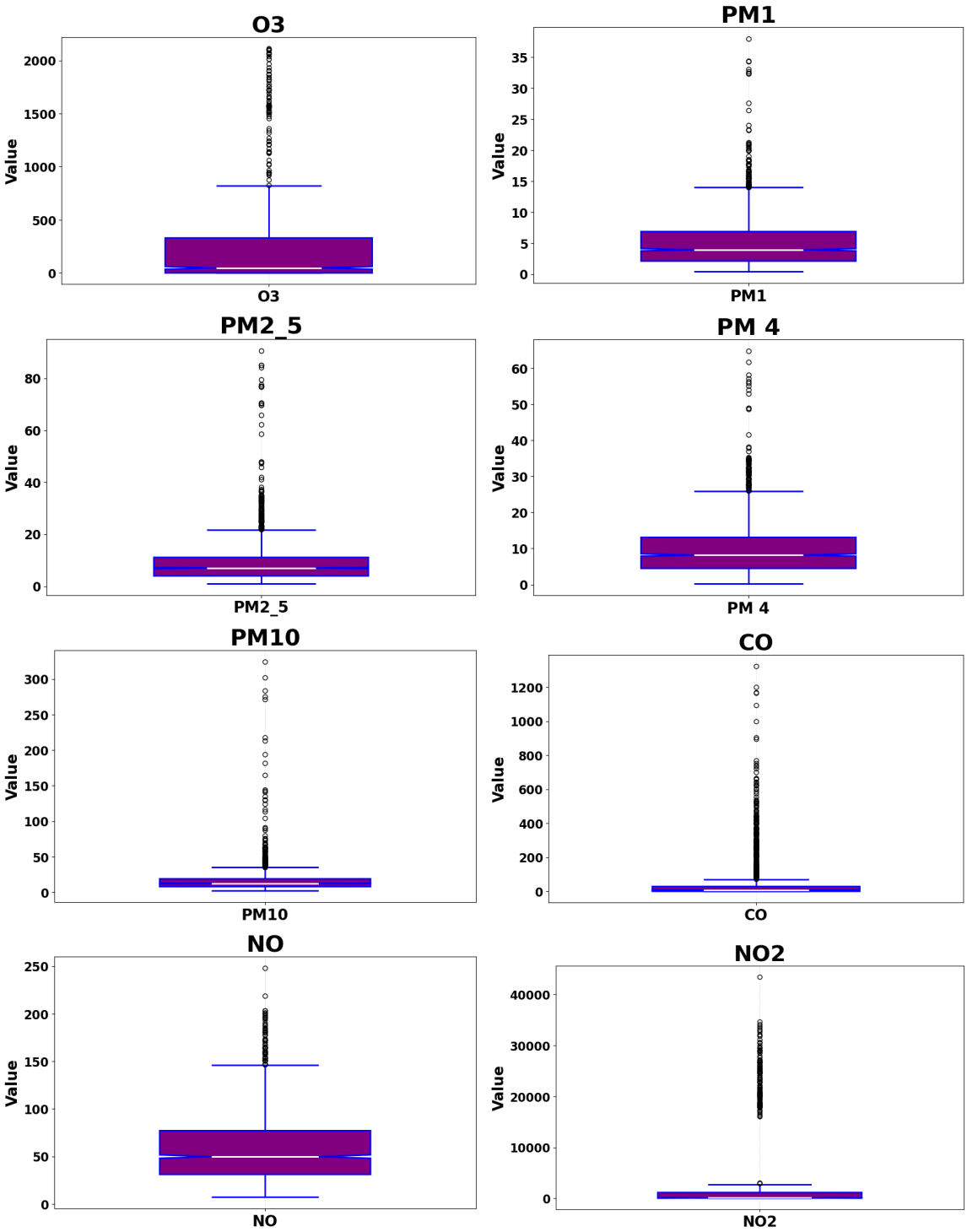

Fig. 4.7 Visualisation of “Air Quality” theme datasets

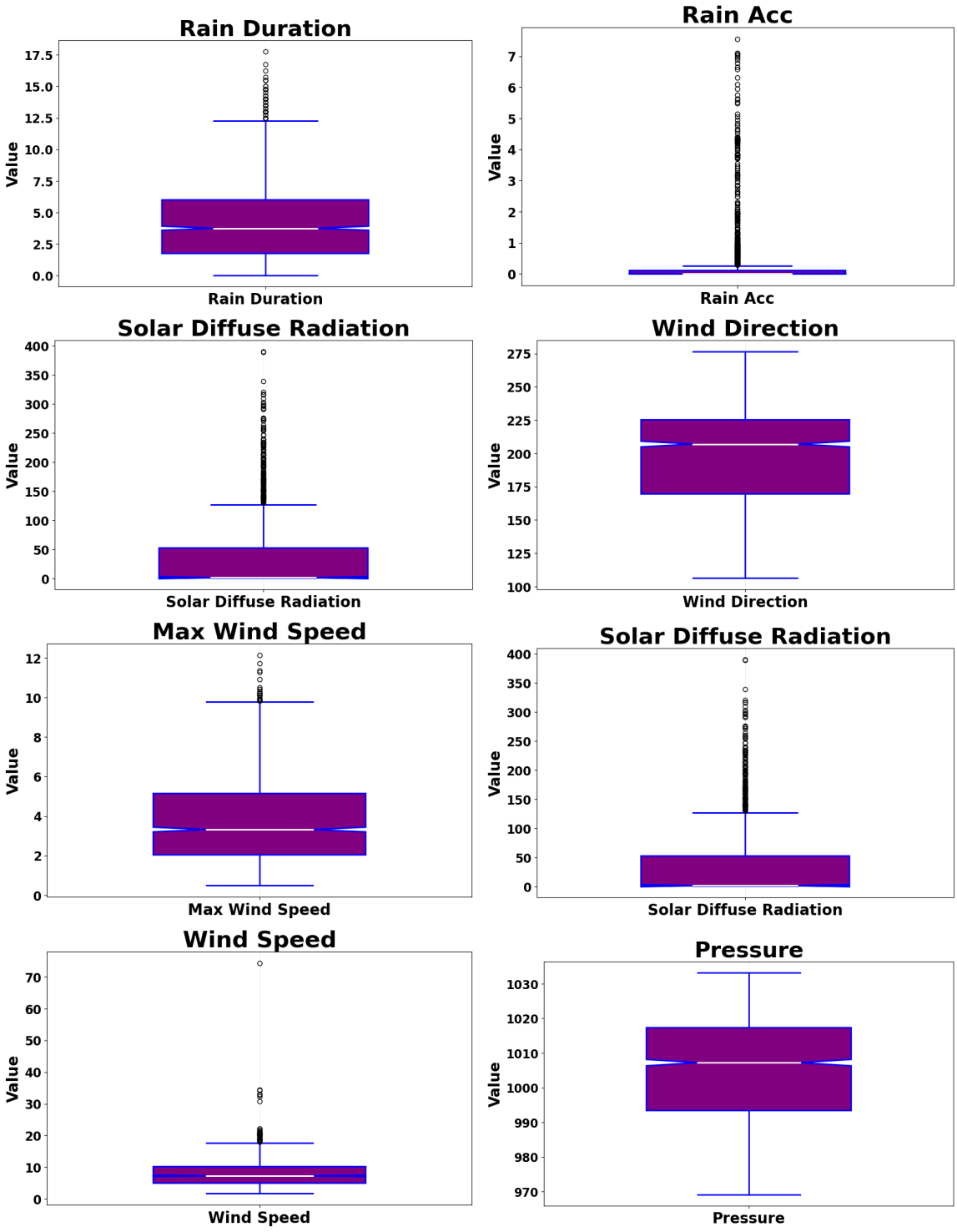

Fig. 4.8 Visualisation of "Weather" theme datasets

4.4 Policy Objective and the Intervention

The cabinet members at Newcastle and Gateshead Councils have confirmed the plans for introducing a clean air zone to operate in Newcastle city centre (Breathe, 2020). The zone

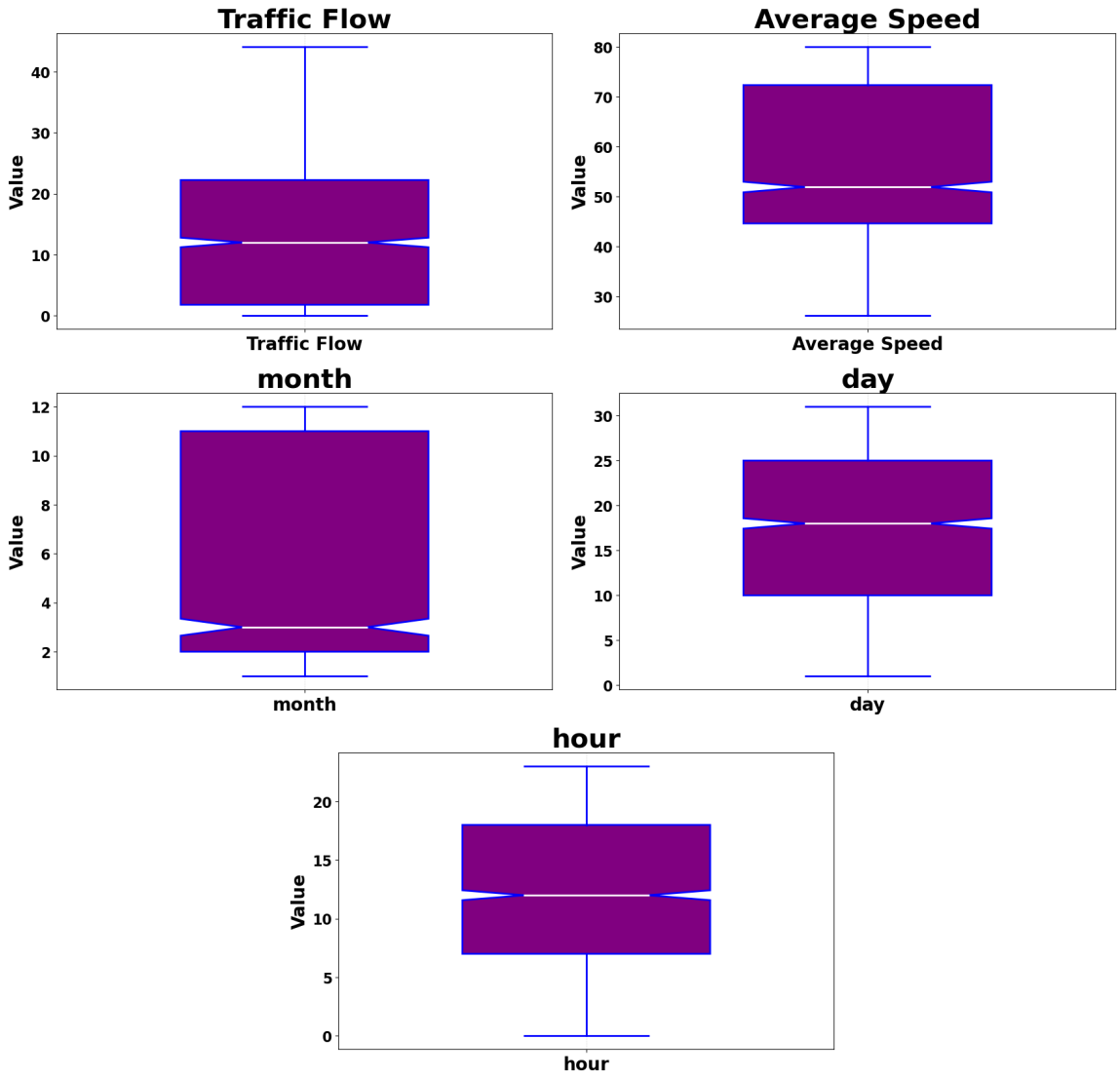

Fig. 4.9 Visualisation of "Traffic" and "TimeStamp" themes datasets

will include the city centre of Newcastle and routes over the Tyne, High Level, Swing and Redheugh bridges. The intervention will impose charges on all buses, taxis, coaches, vans and heavy goods vehicles (HGVs) that do not meet the emissions standards of EURO IV for petrol and EURO VI for diesel vehicles. The primary goal of the clean air zone in Newcastle is to improve the poor air quality. Therefore, the objective of the introduced policy is considered to be the following:

After introducing the clean air zone at Newcastle, the concentration of NO_2 will reduce. More specifically, the time duration when the concentration of NO_2 is unhealthy (NO_2 concentration above 100 parts per billion) will be reduced by 10%.

Note that 10% reduction is selected as an example for a proof of concept to demonstrate the usefulness of the frameworks designed in this chapter. This reduction can be estimated by the Clean Air Zone experts in their technical documents backed up by air quality modelling, operational cost modelling, and behavioural response estimates.[3]

Justification of the selected air quality target. The UK Air Quality Standards Regulations (2010)[4] have set two limit values for NO_2:

(a) the annual mean concentration of NO_2 must not exceed the limit 40 $\left[\frac{\mu g}{m^3}\right]$, and

(b) there should be no more than 18 exceedances of the hourly mean limit value (concentrations above 200 $\left[\frac{\mu g}{m^3}\right]$) in a single year.

The work of the thesis studies NO_2 concentrations as time series with hourly time stamps. Therefore, the hourly limit value of 200 $\left[\frac{\mu g}{m^3}\right]$ and the number of times this limit is exceeded are relevant to the research of this thesis. The considered policy objective is 10% reduction in the time duration when the concentration of NO_2 is unhealthy. This is related to the number of exeedances of the NO_2 concentration limit of $100\,[ppb]$. The value $100\,[ppb]$ is selected based on the divisions proposed by the Office of Air and Radiation (6301A) (2011)[5] with respect to the risk of NO_2 for the general population as described in Section 4.8.

The calculations below show that $100\,[ppb] = 188\left[\frac{\mu g}{m^3}\right]$, which is close to the limit set by the UK Air Quality Standards Regulations (2010). Note that the objective is set to require a certain percentage of reduction relative to the zone without the implementation of policy intervention since this is more appropriate than requiring an absolute reduction in the number of exeedances (i.e., any new intervention is expected to make improvements relative to the current state of the zone).

Translating ppb to $\left[\frac{\mu g}{m^3}\right]$. The thesis reported the concentrations in parts per billion (ppb). The relation between ppb and $\left[\frac{\mu g}{m^3}\right]$ is as follows:

$$\text{Concentration}\left[\frac{\mu g}{m^3}\right] = \text{molecular weight}\left[\frac{g}{mol}\right] \times \text{Concentration}\,[ppb] \div 24.45$$

The molecular weight of NO_2 is $46.01\left[\frac{g}{mol}\right]$. This relation means for NO_2 that $1\,ppb = 1.88\left[\frac{\mu g}{m^3}\right]$ and $100\,ppb = 188\left[\frac{\mu g}{m^3}\right]$.

[3] https://cleanairgm.com/technical-documents
[4] https://www.gov.uk/government/statistics/air-quality-statistics/ntrogen-dioxide
[5] https://www.airnow.gov/sites/default/files/2018-06/no2.pdf.

4.5 Pearson Correlation Coefficient

As part of the framework in Figure 3.4, the correlation between different datasets and target variables can be computed using available data and statistical methods. The correlation coefficient gives a way to assess how much two variables are associated with each other. The correlation coefficient takes values in the interval $[-1,1]$. A value of ± 1 indicates that two variables are dependent linearly. As the correlation coefficient goes towards 0, the relationship between the two variables will be weaker (they become independent of each other). A plus sign in the correlation coefficient shows a positive relationship (an increase in one variable will result in an increase in another variable) and a minus sign shows a negative relationship (an increase in one variable will result in a decrease in another variable). More precisely, for two random variables x, y modelling two datasets, the correlation coefficient is defined as

$$\rho_{x,y} = \frac{\mathbb{E}[(x-m_x)(y-m_y)]}{\sigma_x \sigma_y}, \tag{4.1}$$

where $\mathbb{E}$ is the expectation operator, m_x, m_y are the means and σ_x, σ_y are the variances of x, y. The correlation coefficient is computed using the following formula when a dataset $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ of size n is available for x, y:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{4.2}$$

where $\bar{x}, \bar{y}$ are the empirical means of x, y.

4.6 Permutation Feature Importance

As part of the framework in Figure 3.4, the permutation feature importance computes the importance of a feature by first training the model on the train dataset, then permuting the feature and computing the increase in the prediction error of the model. If a feature is important in making predictions, the prediction error should increase after permutation. If a feature is unimportant, the change in the prediction error will be negligible by doing the permutation. In general, feature importance can take any values. The feature importance is normalised to obtain values in the range $[0,1]$, which show the relative importance of the features.

4.7 Test Metrics

It was discussed in Subsection 2.4.4 that validating the objectives of a policy interventions using data is well-suited for study through supervised learning methods. The choice of supervised learning algorithms will be with respect to their robustness to noise, outliers, and changes in data distribution, which can be influenced by dataset characteristics including size, dimensionality, and noise. Consequently, it is a common practice to employ multiple algorithms to identify those models suitable for specific data properties, assess the robustness and generalisability across diverse datasets, and compare accuracy to evaluate predictive performance and efficiency.

There are five metrics for assessing the performance of the classification machine learning models. It is worth noting that if the performance of the model is high (by appropriate selection of the hyperparameters), this means that the model can capture the essential relations in the dataset and can provide more accurate predictions to be used by policy evaluators and policy makers. A summary of the advantages and disadvantages of these metrics are presented in Table 4.3 with their definitions as follows.

- **Accuracy score:** The first metric is accuracy, which is defined as the total number of correct predictions divided by the total number of predictions. Accuracy can be written as
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- **Precision score:** It is defined for each class as the ratio of true positives to the sum of true and false positives:
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
In other words, for all instances classified positive, what percent was correct.
- **Recall score:** It shows the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives:
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
In other words, for all instances that were actually positive, what percent was classified correctly.
- **F1 score:** This score is the harmonic mean of precision and recall:
$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 4.3 Advantages and disadvantages of different test metrics for machine learning classification models

Metric	Advantages	Disadvantages
Accuracy	Simple and easy to understand, suitable for balanced class distributions.	Can be misleading in imbalanced datasets, may not be appropriate for skewed data.
Precision	Focuses on the positive class, useful when false positives are costly.	Sensitive to imbalanced datasets.
Recall	Emphasises capturing all positives, useful when false negatives are costly.	May result in more false positives, may not be ideal for precision-critical tasks.
F1 Score	Balances precision and recall.	Sensitivity to class imbalance.
Confusion Matrix	Provides detailed breakdown of predictions, Offers insights into true/-false positives/negatives.	Does not provide a single performance metric.

It can take its best value (one) when both precision and recall are equal to one. In worst case, it can be zero.

- **Confusion Matrix:** The precision score defined above is very useful but does not contain all the information needed to judge the performance of a classification model. This is in particular important when the dataset is imbalanced (i.e., some of the labels may appear much less than other labels). The confusion matrix is a matrix that includes the statistics of the correct classes and predicted classes when the trained model is applied on the test dataset to make predictions.

  By definition, a confusion matrix C is a square matrix with dimension n equal to the number of classes, where the entry C_{ij} is equal to the number of predictions known to be in group i and predicted to be in group j. For binary classification (two classes), the count of true negatives is C_{00}, false negatives is C_{10}, true positives is C_{11} and false positives is C_{01}. The performance of a classification is judged to be good if the diagonal entries of the confusion matrix is close to 1 and the off-diagonal entries are close to 0.

Each machine learning model also has its own hyperparameters. For example the Decision Tree classifier has the maximum depth as a hyperparameter. This can be chosen by computing the accuracy of the classifier as a function of maximum depth. The best maximum depth can be selected such that the accuracy of the classifier is maximised.

4.8 Results of Applying the Framework

The dataset is divided randomly into 70% training data for learning the model and 30% test data for assessing the accuracy of the trained model. The 70-30 split is a popular choice in machine learning for a balance between having enough data to train on and enough data to test the model's accuracy effectively. This is done using *train_test_split* function of *sklearn* package in Python. The sensor measurements of NO_2 contain continuous quantities. These values are divided into five different ranges based on their risk for the general population (Office of Air and Radiation (6301A), February 2011). The range of values is based on *parts per billion (ppb)*, which is defined as the number of units of mass of NO_2 per billion units of total mass. These ranges are

1. "Good" for NO_2 concentration between $0-50$ ppb. The NO_2 concentrations in this range are expected to have no impact on health.

2. "Moderate" for NO_2 concentration between $51-100$ ppb. This range of NO_2 is considered to be harmful for people who are sensitive to NO_2. These people should consider limiting extended outdoor exertion.

3. "Unhealthy for Sensitive Group" when the NO_2 concentration is between $101-150$ ppb. This range of NO_2 is considered to be harmful for people with lung disease, for children and older people. They should *limit* extended outdoor exertion.

4. "Unhealthy" for NO_2 concentration between $151-200$ ppb. Children, older people, and anyone with lung disease should *avoid* extended outdoor exertion. Anyone else should limit extended outdoor exertion.

5. "Very Unhealthy" for NO_2 concentration between $201-300$ ppb. Children, older people, and anyone with lung disease should *avoid all* outdoor exertion. Anyone else should limit outdoor exertion.

Figure 4.10 represents the number of NO_2 measurements inside these five classes: Very unhealthy (VU), Unhealthy (U), Unhealthy for sensitive group (US), Moderate (M), and Good (G). The data has been ordered from most to least unhealthy. As it can be seen also from Figure 4.10, Very Unhealthy class has the highest number of measurements between these five classes. Therefore, it is expected that the machine learning models and the training learn the Very Unhealthy class quite well. On the other hand, the number of data entries for Unhealthy class is relatively smaller. Averaging is used to change the time resolution of the measurements from second to hour. This will make the dataset a better representation of the hourly average quantities and make them more robust against measurement noises.

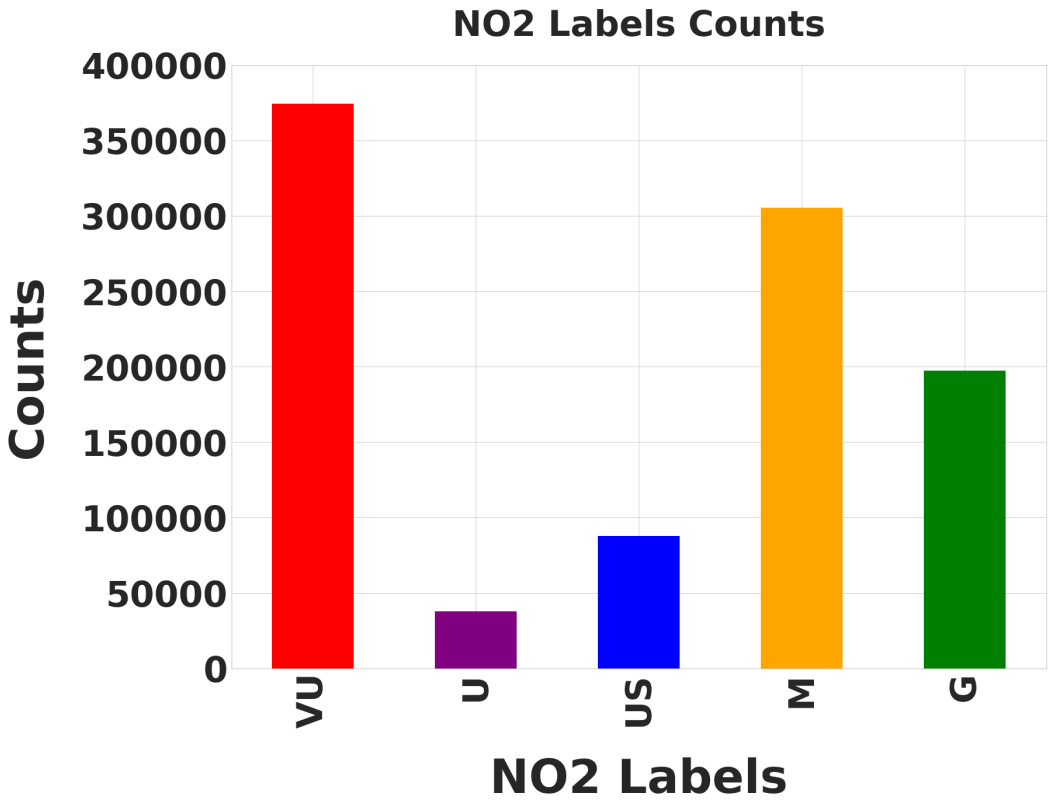

Fig. 4.10 The number of NO_2 measurements in five different classes

The most popular machine learning models are used, which include Decision Tree (DT), Light Gradient Boosting Machine (LGBM), K-Nearest Neighbour (KNN), and Gradient Boosted Decision Tree (LGBM), to build models that can accurately capture the important information in the dataset, make accurate predictions, and help us to extract relevant important data types according to the framework in Figure 3.3. A description of these classification models were provided in Section 2.4.6. Note that the computation of Pearson Correlation for the framework can be done without the need for training a classification model, but the Feature Importance needs constructing a classification model first and then computing the importance values by shuffling the dataset.

4.8.1 Applying the Framework Using Pearson Correlation

The correlations between different features and the NO_2 concentration are computed according to Equation (4.2) to find the most relevant variables related to the NO_2 according to the first framework in Figure 3.3. The result is presented in the right column of Table 4.4 after normalising the values. The feature importance does not have a specific range. To make them comparable, the correlation coefficient and permutation feature importances are

normalised, and the percentages are reported in the table. The numbers in the table are now in the range $[0, 100]$. These numbers show the relative importance of the features. The cut-off value 1% is used as a proof of concept to identify the most important features. In general, this cut-off value should be selected with respect to the size of the dataset and the available computational resources.

The measured variables are written in the left column of Table 4.4 and they are colour-coded to indicate each theme. The variables in the Air Quality Theme are in green colour, the variables in the Timestamp Theme are in red colour, the variables in the Weather Theme are in yellow colour, and the variables in the Traffic Theme are in dark blue colour.

The normalised correlation coefficient reported in the right column show that the two highest coefficients belong to the variables O_3 and the number of cars and taxis. It is also beneficial to look at coefficients across each theme separately. From the Air Quality Theme (green variables), O_3, NO, and CO have the highest correlation with NO_2 but PM_4 has a negligible correlation. In the Timestamp Theme (red variables), Month and Day have the highest correlation. In the Weather Theme (yellow variables) Pressure and Wind Direction have the higher correlation, and in the Traffic Theme (dark blue variables), Cars and Taxis give the highest correlation. Note that although HGVs have a high emission factor, they have a negligible correlation because there is a small numbers of them.

4.8.2 Models Architecture and Implementation for Feature Importance

The selected classification models are Decision Tree (DT), Light Gradient Boosting Machine (LGBM), K-Nearest Neighbour (KNN), and Gradient Boosted Decision Tree (LGBM), to build models that can accurately capture the important information in the dataset, make accurate predictions, and help us to extract relevant important data types according to the framework in Figure 3.3. A description of these classification models were provided in Section 2.4.6. A description of the architecture of these models and their implementation are provided below.

1. **Decision Tree (DT).** The model architecture is based on the following hyperparameters.

   - **Max Depth:** 10
     The maximum depth of the tree, set to 10, to prevent overfitting and ensure the model captures relevant patterns without becoming too complex.
   - **Min Samples Split:** 2
     Minimum number of samples required to split an internal node.

Table 4.4 Normalised Feature Importance computed from different machine learning methods and the normalised correlation

Measured Variables	LGBM	DT	KNN	GBDT	Correlation
O_3	36.3	28.7	45.5	29.1	19.6
NO	1.4	1.8	4.3	2.1	3.1
CO	1.9	2.9	8.6	5.5	5.7
PM_1	1	0.6	0.9	0.8	1.2
$PM_{2.5}$	1.9	1.8	0.9	1.6	1.6
PM_4	1	0.6	0.9	0.8	0.5
PM_{10}	2.9	1.2	0.9	1.6	2.1
Month	19.1	18.7	8.6	15	5.7
Day	4.8	13.5	6.9	8.7	9.4
Hour	1.2	2.6	1.9	2.2	2.4
Pressure	5.7	6.4	4.3	4.7	8.6
Wind Speed	1	1.8	0.9	0.8	0.8
Rain Duration	1	3.5	0.9	1.6	1.2
Max Wind Speed	1.9	1.2	1.4	1.2	4.9
Solar Diffuse Radiation	1	0.6	0.8	0.7	0.9
Wind Direction	2.9	1.2	4.3	0.8	10.6
Solar Radiation	1	1.8	2.6	1.6	1.2
Rain Acc.	1	0.2	0.2	0.1	0.9
Cars and taxis	13.4	11.7	10.3	23.6	16.4
Buses and coaches	1.4	2.1	1.7	2.21	3.1
HGVs	0.1	0.1	0.1	0.1	0.1
Two-wheeled motor vehicles	0.00	0.00	0.00	0.00	0.00
Average Speed	0.1	0.1	0.9	0.2	0.9

- **Min Samples Leaf:** 1
  Minimum number of samples required to be at a leaf node.
- **Random State:** 1234 (to ensure reproducibility of results).

The DT model is implemented according to the following Python code.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

Xs_of_train = trainx18.values
Ys = trainy18.values
Ys_of_train = np.zeros(Ys.shape)
```

```
for i in range(len(Ys)):
    if(Ys[i] == 'Good'):
        Ys_of_train[i] = 0
    elif(Ys[i] == 'Moderate'):
        Ys_of_train[i] = 1
    elif(Ys[i] == 'Unhealthy for sensitive group'):
        Ys_of_train[i] = 2
    elif(Ys[i] == 'Unhealthy'):
        Ys_of_train[i] = 3
    elif(Ys[i] == 'Very unhealthy'):
        Ys_of_train[i] = 4

DT18 = DecisionTreeClassifier(max_depth=10, min_samples_split=2,
                        min_samples_leaf=1, random_state=1234)
DT18.fit(Xs_of_train, Ys_of_train)
DT18_train_res = DT18.predict(Xs_of_train)
dtacc_trainc = accuracy_score(DT18_train_res, Ys_of_train)
print(f'Training accuracy : {dtacc_trainc*100}%')
```

2. **Light Gradient Boosting Machine (LGBM).** The model architecture is developed using the following hyperparameters.

   - **Number of Leaves:** 31
     Controls the complexity of the tree model. A higher number of leaves can increase model accuracy but may lead to overfitting.
   - **Learning Rate:** 0.05
     Determines the step size at each iteration while moving towards a minimum of the loss function. A smaller learning rate requires more trees but can lead to better generalization.
   - **Feature Fraction:** 0.8
     The fraction of features to be used for each tree. Reduces overfitting by introducing randomness.
   - **Bagging Fraction:** 0.8
     The fraction of data to be used for each iteration. Also helps in reducing overfitting.
   - **Max Depth:** -1 (unlimited, but can be adjusted to prevent overfitting)
     The maximum depth of each tree. Limits the growth of the tree to prevent overfitting.
   - **Random State:** 1234 (to ensure reproducibility of the results).

The LGBM model is implemented based on the following Python code.

```
from lightgbm import LGBMClassifier
from sklearn.metrics import accuracy_score

Xs_of_train = trainx18.values
Ys = trainy18.values
Ys_of_train = np.zeros(Ys.shape)
for i in range(len(Ys)):
    if(Ys[i] == 'Good'):
        Ys_of_train[i] = 0
    elif(Ys[i] == 'Moderate'):
        Ys_of_train[i] = 1
    elif(Ys[i] == 'Unhealthy for sensitive group'):
        Ys_of_train[i] = 2
    elif(Ys[i] == 'Unhealthy'):
        Ys_of_train[i] = 3
    elif(Ys[i] == 'Very unhealthy'):
        Ys_of_train[i] = 4

log18 = LGBMClassifier(random_state=1234, num_leaves=31,
                                    learning_rate=0.05,
                                    feature_fraction=0.8,
                                    bagging_fraction=0.8,
                                    max_depth=-1)
log18.fit(Xs_of_train, Ys_of_train)
log18_train_res = log18.predict(Xs_of_train)
com_acc_train = accuracy_score(log18_train_res, Ys_of_train)
print(f'Training accuracy : {com_acc_train*100}%')
```

3. **K-Nearest Neighbour (KNN).** The model architecture is developed using the following hyperparameters.

   - **Number of Neighbours (k):** 5
     Set to 5 to balance between noise reduction and capturing local patterns.
   - **Distance Metric:** Euclidean distance
     Commonly used metric to calculate the distance between points.

The KNN model is implemented using the following Python code.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
```

```
Xs_of_train = trainx18.values
Ys = trainy18.values
Ys_of_train = np.zeros(Ys.shape)
for i in range(len(Ys)):
    if(Ys[i] == 'Good'):
        Ys_of_train[i] = 0
    elif(Ys[i] == 'Moderate'):
        Ys_of_train[i] = 1
    elif(Ys[i] == 'Unhealthy for sensitive group'):
        Ys_of_train[i] = 2
    elif(Ys[i] == 'Unhealthy'):
        Ys_of_train[i] = 3
    elif(Ys[i] == 'Very unhealthy'):
        Ys_of_train[i] = 4

KN18 = KNeighborsClassifier(n_neighbors=5)
KN18.fit(Xs_of_train, Ys_of_train)
KN18_train_res = KN18.predict(Xs_of_train)
knacc_trainc = accuracy_score(KN18_train_res, Ys_of_train)
print(f'Training accuracy : {knacc_trainc*100}%')
```

4. **Gradient Boosted Decision Tree (GBDT).** The model architecture is developed using the following hyperparameters.

   - **Max Depth:** 10
     Set to 10 to prevent overfitting while capturing significant patterns.
   - **Learning Rate:** 0.1
     Controls the contribution of each tree.
   - **Number of Estimators:** 100
     Number of trees in the ensemble.
   - **Random State:** 1234 (to ensure reproducibility of the results).

   The GBDT model is implemented according to the following Python code.

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score

Xs_of_train = trainx18.values
Ys = trainy18.values
Ys_of_train = np.zeros(Ys.shape)
for i in range(len(Ys)):
```

```
    if(Ys[i] == 'Good'):
        Ys_of_train[i] = 0
    elif(Ys[i] == 'Moderate'):
        Ys_of_train[i] = 1
    elif(Ys[i] == 'Unhealthy for sensitive group'):
        Ys_of_train[i] = 2
    elif(Ys[i] == 'Unhealthy'):
        Ys_of_train[i] = 3
    elif(Ys[i] == 'Very unhealthy'):
        Ys_of_train[i] = 4

BRT18 = GradientBoostingClassifier(max_depth=10, learning_rate=0
                                   .1, n_estimators=100,
                                   random_state=1234)
BRT18.fit(Xs_of_train, Ys_of_train)
BRT18_train_res = BRT18.predict(Xs_of_train)
brtacc_trainc = accuracy_score(BRT18_train_res, Ys_of_train)
print(f'Training accuracy : {brtacc_trainc*100}%')
```

4.8.3 Data Splitting for Model Training and Evaluation

In machine learning, the dataset is split into distinct training and evaluation datasets. In order to compare different machine learning models and measure their performance, the following techniques can be used for data splitting:

1. **Fixed Training and Evaluation Sets.** Models are trained on the same training set and evaluated on the same evaluation set to measure their performance.

2. **k-Fold Cross-Validation.** The dataset is divided into k equally sized folds. The model is trained k times, each time using k-1 folds for training and the remaining fold for evaluation.

3. **Leave-One-Out Cross-Validation.** This is a special case of k-fold cross-validation, where k equals the number of data points. Each data point is used once as a test instance while the remaining points form the training set.

4. **Repeated Randomised Splitting.** The dataset is divided randomly into distinct training and evaluation datasets multiple times, a fixed number of models from each class are trained and evaluated on each split, and the average and standard deviation of the performance metrics are reported for each class of models.

For the purpose of this research, *repeated randomised splitting* is selected due to the following reasons:

(a) Fixed splitting may result in training or evaluation sets with disproportionate class distributions in the imbalanced dataset of this research. This will lead to biased performance metrics, and the performance estimate can be highly sensitive to the particular split of the data. This dependence on the initial split can undermine the reliability of model comparisons.

(b) k-Fold cross-validation may not preserve the proportion of classes in each fold, leading to high variance in model evaluation.

(c) Leave-One-Out cross-validation is effective in small datasets, but increases the risk of overfitting for the large datasets of this research and requires excessively large computational time since the number of trained models is equal to the size of the dataset.

In contrast, repeated randomised splitting gives the empirical average of the performance metric for each class of models, which is more robust to disproportionate class distributions. Moreover, the standard deviation of the performance metric across trained models gives a measure of reliability and robustness of computed models.

The accuracy of the models of this chapter will be computed and repeated 10 times. The results reported in the rest of this chapter show that the standard deviation of the performance is less than 1%, which shows a robust and reliable model learning.

4.8.4 Applying the Framework Using Feature Importance

After training the four classification models, the Feature Importance values for all these models are computed. The result is presented in Table 4.4. In each column, the relative importance values more than 1% are highlighted in magenta colour. For LGBM model, O_3 and Month have the highest importance. The following variables have the highest importance value in each theme: O_3 from the Air Quality Theme, Month from the Timestamp Theme, Pressure from the Weather Theme, and Car and taxis from the Traffic Theme.

For DT model, O_3 and Month have the highest importance. The following variables have the highest importance value in each theme: O_3 from the Air Quality Theme, Month from the Timestamp Theme, Pressure from the Weather Theme, and Car and Taxis from the Traffic Theme.

For KNN model, O_3 and Car and taxis have the highest importance. The following variables have the highest importance value in each theme: O_3 from the Air Quality Theme,

Month from the Timestamp Theme, Pressure and Wind Direction from the Weather Theme, and Car and Taxis from the Traffic Theme.

For GBDT model, O_3 and Car and taxis have the highest importance. The following variables have the highest importance value in each theme: O_3 from the Air Quality Theme, Month from the Timestamp Theme, Pressure from the Weather Theme, and Car and Taxis from the Traffic Theme.

Feature importance is generally superior to correlation coefficient for several reasons:

- **Non-linear Relationships**: Feature importance captures non-linear relationships between features and the target variable, whereas correlation only measures linear relationships that may miss complex patterns.
- **Feature Interactions**: Feature importance accounts for interactions between features within the model, providing a more holistic view of their impact on predictions. Correlation, being pairwise, does not consider these interactions.
- **Model-Specific Relevance**: Feature importance is derived directly from the model and reflects how each feature contributes to the model's accuracy. Correlation, on the other hand, is a general statistical measure that does not reflect the model's predictive context.

As Table 4.4 shows for the dataset of the clean air zone case study, all the feature importances and correlation coefficients share common conclusions about the importance of different variables in predicting NO_2 and validating the intervention that involves reduction of NO_2, but also have slight differences. The conclusions of these models could be used in a voting mechanism to decide on the importance of features (similar to ensemble learning (Polikar, 2012)). For this voting mechanism, machine learning models are trained, the relative feature importances are computed, and the features indicated as important by the majority of these models are extracted. For instance, most of these models used in the framework state that O_3, Month, Day, Pressure, and the Number of Cars and Taxis are important. On the other hand, Wind Speed and Number of two-wheeled motor vehicles are less important in building a model for validating the objectives of the intervention. These findings are also confirmed by the general intuitive observations about air quality: there is evidence of high correlation between NO_2 and O_3 (Dimitrievici et al., 2017; Han et al., 2011; Zhao et al., 2020); the month is important as the air quality can get impacted duo to the seasons and weather conditions; the day will impact the air quality as usually traffic volume might be higher during the working days and lower over the weekends. In the dataset of this chapter, the average difference in traffic volume between weekends and the rest of the week is 35%.

There are also differences between the importance values computed using these models. This is mainly due to the nature of these algorithms that elements of randomness in their computations. The randomness is originating from the optimisation used to lean the model (exploring the parameter space of the model randomly), from the random initialisation of the optimisation, and from random splitting of the dataset to train-test data (Witten and Frank, 2002). Therefore, any conclusion taken from the computations should be balanced with the accuracy of the constructed model. for instance, as it will be discussed in the rest of this subsection, the accuracy of the LGBM model is lower compare to other models. Therefore, the results obtained based on this model should be discounted appropriately for making the final decision.

4.8.5 Evaluation of the Classification Models

The metrics Accuracy, Confusion Matrix, F1 score, Precision, and Recall are used to evaluate the obtained classification models (See Subsection 4.7 for the definition of these metrics). Using the repeated randomised data splitting approach, the accuracy of the models are computed and repeated 10 times to account for different training and test split of the dataset. The average accuracy of the LGBM model is 88%, the average accuracy of DT model is 85%, the average accuracy of KNN model is 80%, and the average accuracy of GBDT model is 84%. The standard deviation of these accuracies is less than 1%. Among all these methods LGBM has the highest average accuracy and KNN model has the lowest average accuracy. This shows that the LGBM model can predict the correct class in almost 9 out of 10 cases. While this is an excellent outcome, the accuracy should be considered along other metrics to have the full understanding of accuracy in each class. Therefore, the confusion matrix is also reported.

Figures 4.11–4.14 show the Normalised Confusion Matrix for the four classification models respectively for GBDT, DT, KNN, and LGBM models. The numbers associated with different classes are: 0 for Good, 1 for Moderate, 2 for Unhealthy Sensitive Group, 3 for Unhealthy, and 4 for Very Unhealthy. Recall that for a good classification, the diagonal entry of the confusion matrix should be close to one and the off-diagonal entries should be close to zero. As you can see from these Confusion Matrices, the models have learned the classes $1, 2, 3, 4$ relatively well, but the performance is not good for class number 3. This is mainly due to the fact that the dataset used in this chapter has a different number of data points in each class, and the number of data points in the Unhealthy class is much smaller than in other classes. Note that confusion matrix is used to ensure that the imbalanced classes do not give misleading conclusions.

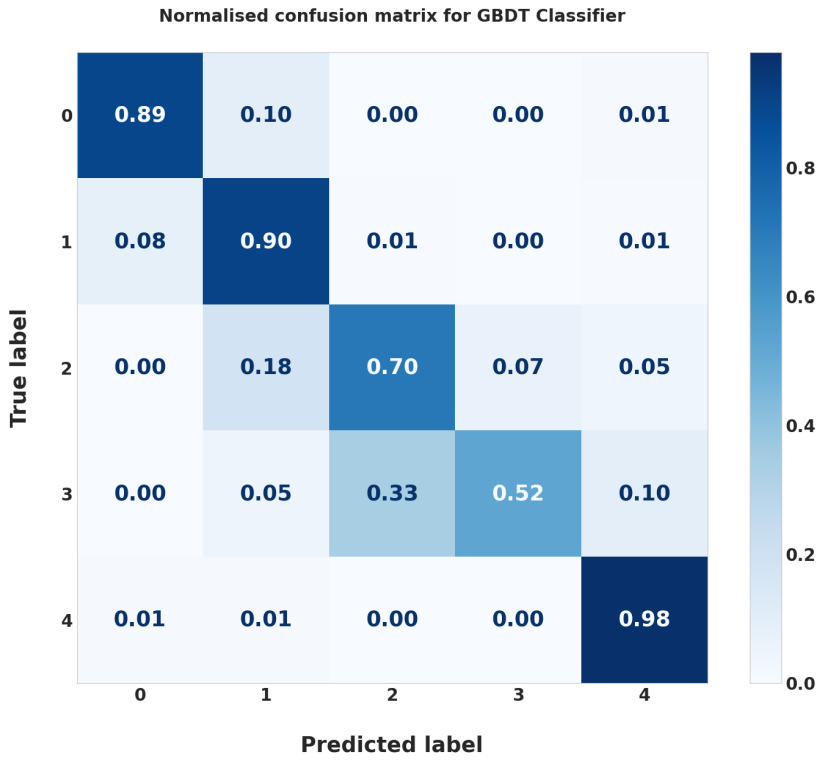

Fig. 4.11 Normalised Confusion Matrix for the GBDT classification model

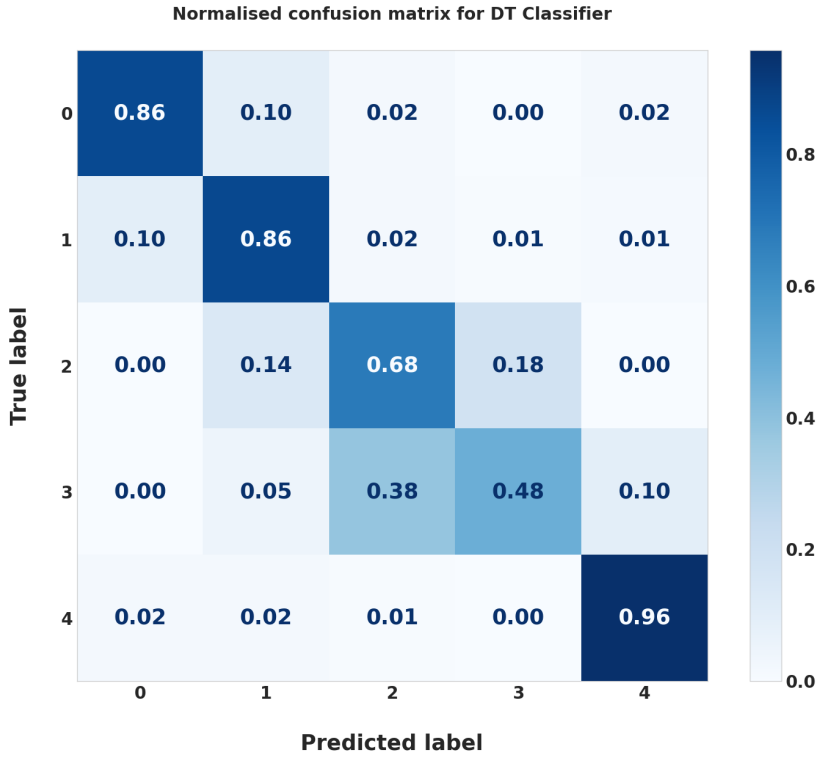

Fig. 4.12 Normalised Confusion Matrix for the DT classification model

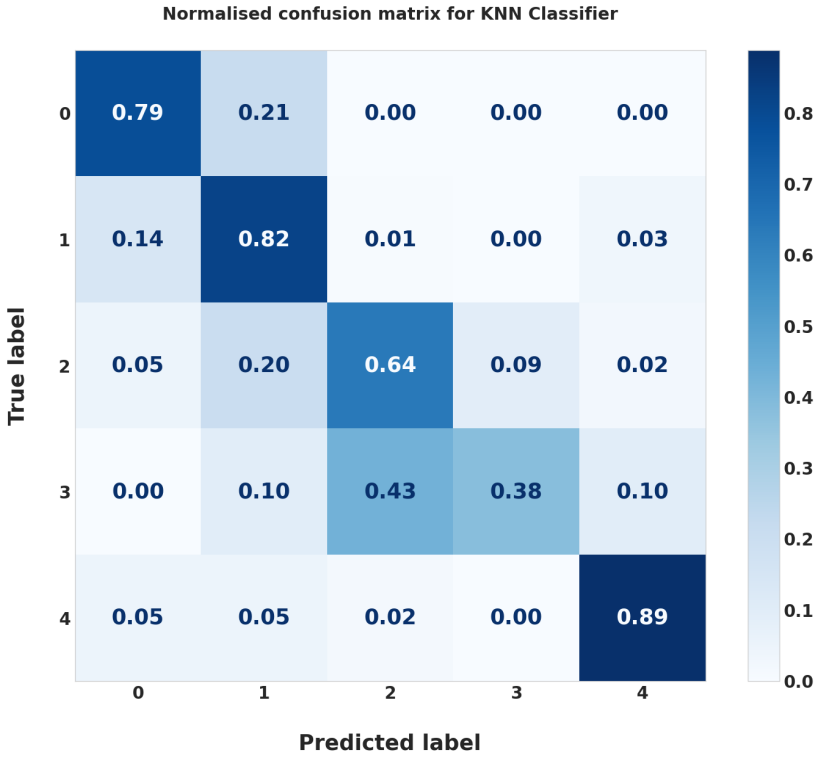

Fig. 4.13 Normalised Confusion Matrix for the KNN classification model

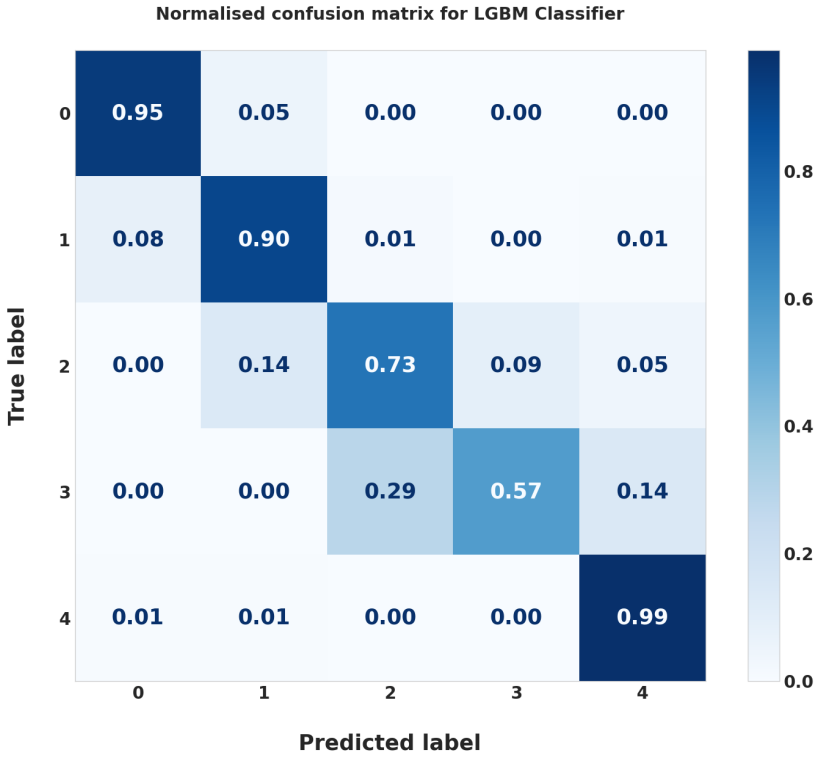

Fig. 4.14 Normalised Confusion Matrix for the LGBM classification model

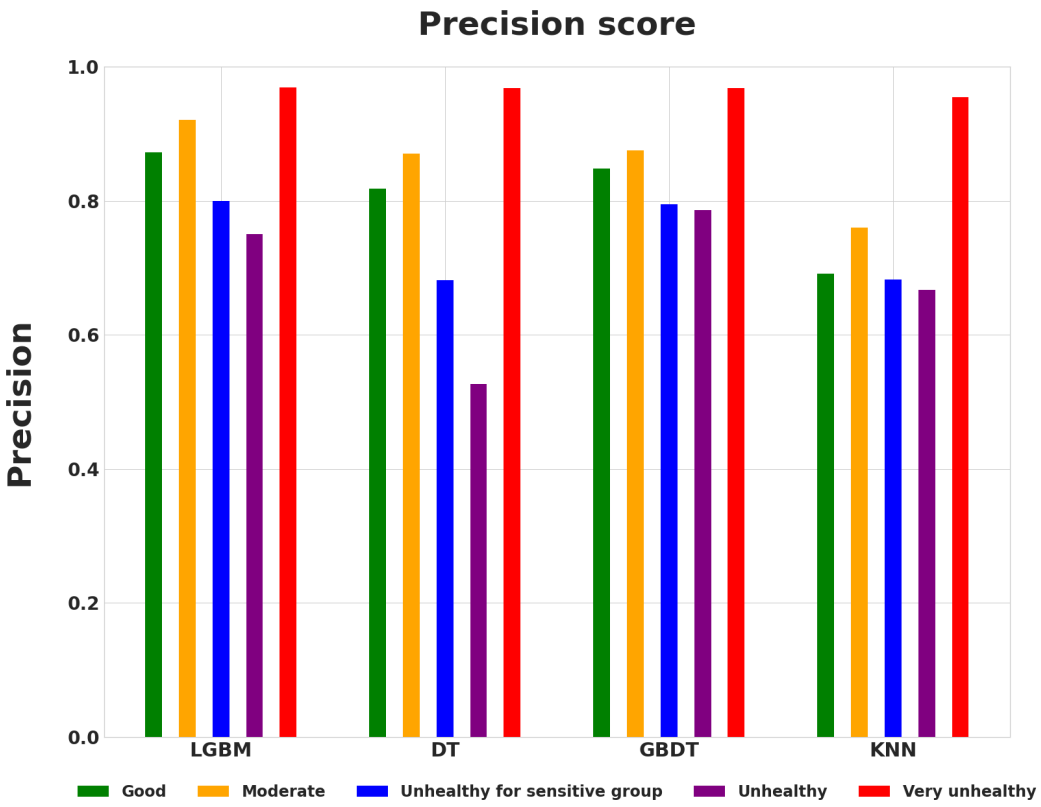

Fig. 4.15 Precision score for each method

Figures 4.15–4.17 shows the comparison between the trained models using Precision, F1 score and Recall score. The figures provide these metrics for all five classes separately. The number of data points of classes are different as reported in Figure 4.10, the machine learning algorithms have different performances in capturing the relations in the data: the class Unhealthy with the smallest data points has the lowest score and the class Very Unhealthy with the largest data points has the highest score.

4.9 Conclusions

This chapter described how the generic steps of the policy cycle are mapped into the steps taken in defining and evaluating the clean air zone intervention with reference to the use of data. Throughout this chapter, the proposed framework for identifying the relevant data types was applied to a policy objective on air quality that reduces NO_2 concentrations through the implementation of a clean air zone. The chapter demonstrated how machine learning can effectively aid in finding and understanding the critical data types that align with the objective of this policy intervention.

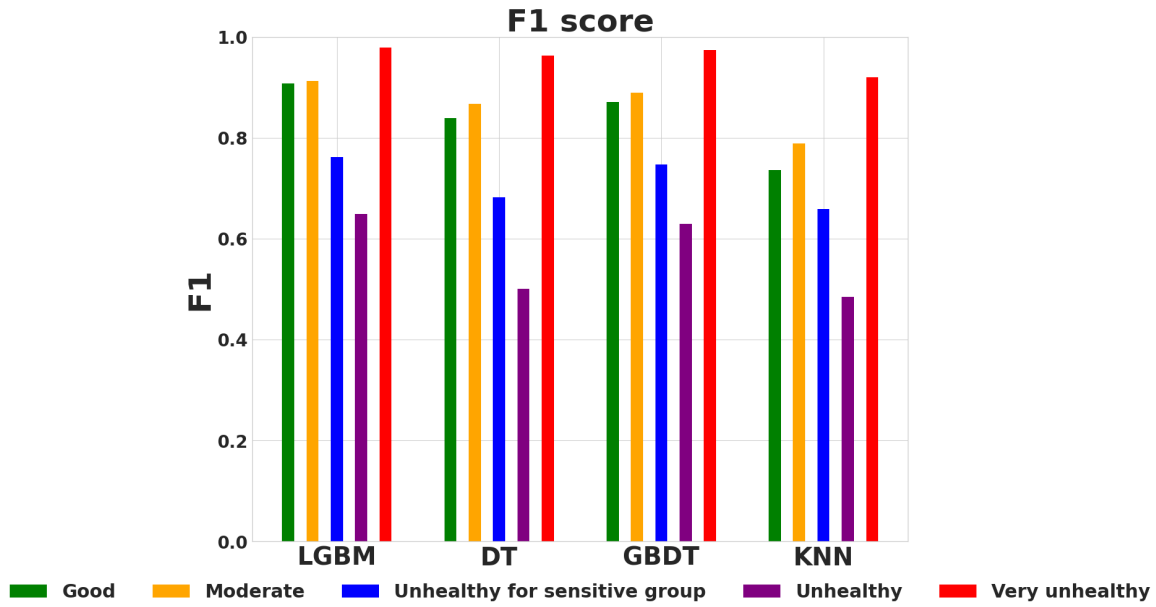

Fig. 4.16 F1 score for each method

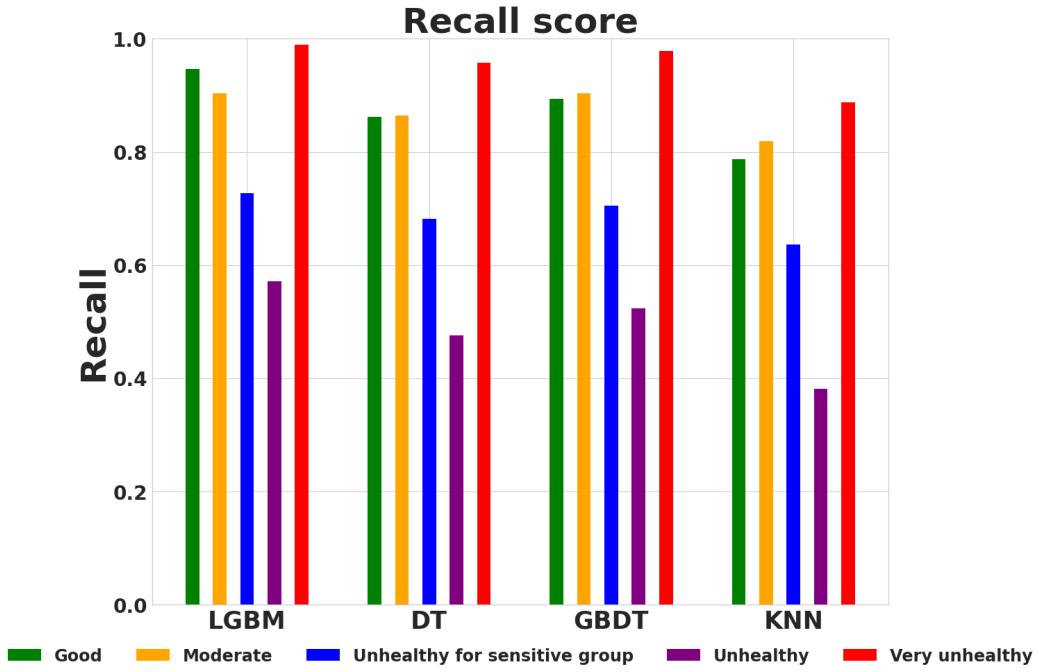

Fig. 4.17 Recall score for each method

To demonstrate applying the framework, data from the Newcastle Urban Observatory was utilised. This chapter described the datasets collected and prepared for applying the methodology of the thesis and the designed frameworks to clean air zone case study in the geographical area of Newcastle upon Tyne. Furthermore, data analysis provided the preprocessing steps performed to prepare and clean the raw data before it is fed into a machine learning model. It also gave the details of applying statistical methods for exploration, analysis, and visualisation of the data gathered from Newcastle Urban Observatory for air quality.

The application of machine learning classifiers including Decision Tree, Light Gradient Boosting Machine, K-Nearest Neighbour, and Gradient Boosted Decision Tree, provided valuable insights into the importance of different features in predicting NO_2 levels. Among these models, Light Gradient Boosting Machine stood out with its highest accuracy of 88%. While this is an excellent outcome, the accuracy should be considered along other metrics to have the full understanding of accuracy in each prediction class. Therefore, the confusion matrix was also reported.

The main advantage of using these machine learning models is that their training does not require an understanding of air pollutants' physical or chemical properties. The structures and properties of machine learning models allow us to incorporate complex nonlinear relationships between the concentration of air pollutants and meteorological variables.

The next chapter will utilise the relevant data types identified through this framework to validate the objective of the intervention. This validation will be accomplished using a time series method.

Chapter 5

Machine Learning & Validating the Objective of Clean Air Zone

A framework was proposed in Chapter 3 for validating the objectives of policy interventions. This chapter shows how to apply the framework to the use case of a clean air zone where the objective is to improve air quality in areas where there are currently nitrogen dioxide (NO_2) exceedances. The framework addresses the challenge of validating the policy objective by comparing the NO_2 concentrations of the zone in the two cases of with and without the intervention.

A time series machine learning model is developed for predicting the NO_2 concentrations using dataset from the Newcastle Urban Observatory. As part of the framework, a suitable time series model is chosen. The selected model is Long Short-Term Memory with it details presented in Section 2.4.6. The test metric for assessing the performance of the LSTM model employed in this research is also given.

This chapter is organised into the following sections. The details of the policy objectives is discussed in Section 5.1. Time series prediction using LSTM is presented in Section 5.2. The details of the architecture and implementation of the LSTM model are given in Section 5.3. The test metric for assessing the performance of the LSTM model is also given in this section. Section 5.4 presents the results of applying the framework on the selected policy objective and the dataset from the Newcastle Urban Observatory. Finally, the chapter provides the conclusions in Sections 5.5. The code for reproducing the results of this chapter is provided in Appendix C.

The content of this chapter is based on the following articles published during the PhD study.

- Farhadi, Farzaneh, Roberto Palacin, and Phil Blythe. "Machine Learning for Transport Policy Interventions on Air Quality." IEEE Access (2023). DOI: https://doi.org/10.1109/ACCESS.2023.3272662

- Farhadi, Farzaneh, Roberto Palacin, and Phil Blythe. "Data-driven framework for validating policies: Air quality case study." 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022. Published. DOI: https://doi.org/10.1109/ITSC55140.2022.9922587

- Farhadi, F., Palacin, R. and Blythe, P., 2023. Machine Learning Methods for Policy Interventions. ECTRI Young Researchers, Lisbon, Portugal (May 2023).

5.1 Policy Objective and the Intervention

For the objective of the intervention, it is considered to be 10% reduction in the time duration when the concentration of NO_2 is unhealthy (NO_2 concentration above 100 ppb). The limit of 100 ppb is selected based on the divisions proposed by the Office of Air and Radiation (6301A) (2011)[1] with respect to the risk of NO_2 for the general population as described in Section 4.8. The objective is equivalent to 10% reduction in the number of NO_2 exeedances of 188 $\left[\frac{\mu g}{m^3}\right]$, which is close to the limit value of 200 $\left[\frac{\mu g}{m^3}\right]$ set by the UK Air Quality Standards Regulations (2010).[2]

Since the clean air zone of Newcastle is at the very early stage of being implemented, only historical data before the implementation of the intervention is available. To develop a machine learning model for predicting the concentration of NO_2 after the implementation of the clean air zone, the historical dataset is modified based on the assumptions mentioned next.

- Implementation of the clean air zone of Newcastle will result in at least 20% reduction in the number of cars and taxis, 10% reduction in the number of buses and coaches, and 20% reduction in the number of HGVs.

- The implementation of the zone will result in an average reduction in the air pollution concentrations. To make this assumption quantitative and realistic, the Emissions Factors Toolkit (EFT) published by the Department for Environment, Food & Rural Affairs (Defra) of the united Kingdom is used to estimate the average concentrations based on the traffic flow before and after the intervention. The difference between

[1] https://www.airnow.gov/sites/default/files/2018-06/no2.pdf.
[2] https://www.gov.uk/government/statistics/air-quality-statistics/ntrogen-dioxide

these two concentration estimations is taken and is deducted from the original dataset. This means the particles from the Air Quality Theme are modified as follows: CO – 18, $PM_{2.5}$ – 16, PM_1 – 10, PM_{10} – 21, PM_4 – 23, O_3 – 29, NO – 18, NO_x – 24, and NO_2 – 25.

These assumptions are used to demonstrate the applicability of the developed framework. The next subsections discuss how to apply the framework presented in Section 3.6 to this clean air zone intervention.

5.2 Time Series Prediction and Evaluation Metrics

The dataset of air quality from the Newcastle Urban Observatory includes measurements of the features as a function of time. These measurements can be seen as a sequence of data points that are sequential and are dependent along the time axis. As discussed in Chapter 4 of this thesis, the structures and properties of machine learning models allow us to incorporate complex nonlinear relationships between the concentration of air pollutants and meteorological variables.

The Long Short-Term Memory (LSTM) model is chosen for predicting NO_2 concentrations. Selecting LSTM over other types of models is due to the following reasons:

- Temporal dependencies: LSTM networks are specifically designed to handle time-series data and capture long-range dependencies (Van Houdt et al., 2020), which are crucial when working with air quality data. Air quality variables including NO_2 concentrations are influenced by past values and trends, making LSTMs more suitable for than deep neural networks, which do not explicitly model temporal dependencies.

- Handling of vanishing and exploding gradients: LSTMs are designed to overcome the vanishing and exploding gradient problems often encountered in training standard recurrent neural networks (Graves, 2012). These issues make it difficult for recurrent neural networks to learn long-range dependencies in time-series data. LSTMs, with their gating mechanisms, can efficiently learn long-range dependencies without the gradient problems, making them more appropriate for predicting NO_2 concentrations.

- Robustness to missing data (Tian et al., 2018): In real-world air quality datasets, missing data is a common issue. LSTMs are more robust to missing data than deep neural networks due to their ability to maintain hidden states over time. This allows LSTMs to better handle gaps in the data and still provide accurate predictions.

The accuracy of time series models in making correct predictions on the dataset is evaluated by a metric called "Root Mean Square Error (RMSE)", defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|^2} \tag{5.1}$$

where y_t is the true output (vector of labels) and $\hat{y}_t$ is the output predicted by the time series model. Two time Series Models are trained in this chapter:

1. The first LSTM model is developed to predict NO_2 concentrations in the future *without* the application of the intervention. This model is trained on the historical time series data. The RMSE in (5.1) is computed with y_t being the true NO_2 concentration levels in the dataset and with $\hat{y}_t$ being the NO_2 concentration levels predicted by the trained model.

2. The second LSTM model is developed to predict NO_2 concentrations in the future *with* the application of the intervention. Since the clean air zone of Newcastle is at the early stage of implementation and the dataset under the intervention is not available yet, the historical dataset is modified based on the assumptions mentioned in the previous section. The RMSE in (5.1) is computed with y_t being the NO_2 concentration levels obtained under assumptions of Section 5.1 and with $\hat{y}_t$ being the NO_2 concentration levels predicted by the model trained on the modified dataset.

5.3 Model Architecture and LSTM Implementation

The details of LSTM model were described in Section 2.4.6. The model is developed using the following hyperparameters.

- **Units:** 200
  Number of LSTM units in the layer. This parameter determines the dimensionality of the output space.

- **Input Shape:** (trainx18.shape[1], trainx18.shape[2])
  Defines the shape of the input data. The model expects data with this shape for training.

- **Loss Function:** Mean Squared Error (MSE)
  The loss function used to measure the difference between predicted and actual values.

- **Optimizer:** Stochastic Gradient Descent (SGD)
  Optimization algorithm used to update the weights of the network.

- **Epochs:** 500
  Number of times the learning algorithm will work through the entire training dataset.
- **Batch Size:** 1
  Number of samples per gradient update.

The LSTM model is implemented according to the following Python code.

```
# Model definition
model18 = Sequential()
model18.add(LSTM(200, input_shape=(trainx18.shape[1], trainx18.shape[
                                    2])))
model18.add(Dense(1))
model18.compile(loss="mean_squared_error", optimizer='sgd')

# Model training
history18pre = model18.fit(trainx18, trainlab18, epochs=500,
validation_data=(testx18, testlab18), batch_size=1, verbose=2,
                                    shuffle=True)

# Plot training history
plt.plot(history18pre.history['loss'], label='Train')
plt.plot(history18pre.history['val_loss'], label='Test')
plt.xlabel('Epoch')
plt.xlim(xmin=0)
plt.ylabel('MSE')
plt.legend()
plt.show()

# Model evaluation
from sklearn.metrics import mean_squared_error

yhat18 = model18.predict(testx18)
rmse18 = np.sqrt(mean_squared_error(testlab18, yhat18))
total_rms = np.sqrt(mean_squared_error(testlab18, np.zeros(len(
                                    testlab18.values.tolist()))))
rmse18per = 100 * rmse18 / total_rms

print(f'Test RMSE for 2018: {rmse18:.3f}')
print(f'Test root mean square for 2018: {total_rms:.3f}')
print(f'Prediction root mean square for 2018:
        {np.sqrt(mean_squared_error(yhat18,
        np.zeros(len(testlab18.values.tolist())))):.3f}')
print(f'Test RMSE percent for 2018: {rmse18per:.3f}')
```

5.4 Results of Applying the Framework

Since the clean air zone of Newcastle is at the early stage of implementation, the framework in Figure 3.4 is used under the first scenario, where only the historical data before the implementation of the intervention is available. The LSTM model is used for training on the historical time series data and predict NO_2 concentrations (or its level) in the future without the application of the intervention. An LSTM model is also developed for predicting the concentration of NO_2 with the implementation of the clean air zone using the dataset modified according to Section 5.1. The dataset of the year 2018 is divided into the first 10 months for training and evaluation, and the next two months for prediction.

Figure 5.1 shows the percentages of each class of NO_2 predicted by the two LSTM model for the year 2018 with and without the intervention. According to Figure 5.1, the NO_2 concentrations of this year under no intervention will be 37.34% Very Unhealthy, 3.77% Unhealthy, 8.74% Unhealthy for Sensitive Group, 30.46% Moderate, and 19.69% Good. Figure 5.1 also shows that the NO_2 concentration of this year under intervention will be 37.16% Very Unhealthy, 0.84% Unhealthy, 7.84% Unhealthy for Sensitive Group, 15.02% Moderate, and 39.14% Good. As it can be seen, the number of NO_2 concentrations that are Very Unhealthy, Unhealthy, or Unhealthy for Sensitive Group and Moderate is reduced and the number of NO_2 concentrations that are Good is increased. The largest reduction is in the Moderate class (15.44% reduction) and the smallest reduction is in the Very Unhealthy class (0.18% reduction).

Figures 5.2–5.3 shows the LSTM prediction of NO_2 concentrations with and without the intervention in the available data points of months 11 and 12 (not used for training the models). The horizontal axes of these figures indicate the available time points ordered in a sequence. The vertical axes are the classes of NO_2 concentrations for these time points predicted by the LSTM models. Figure 5.4 shows the Cumulative Distribution Function (CDF) in the LSTM predictions with and without the intervention. According to the CDF, the classes 0 and 1 (good and moderate) have more amount than other three classes compared with the case of no intervention.

Note that Figure 5.1 shows a consistent result on NO_2 reduction: the number of NO_2 concentration levels that are Good is increased under the intervention, and the number of NO_2 concentration levels in the other four classes (Very Unhealthy, Unhealthy, Unhealthy for Sensitive Group, and Moderate) is reduced.

Figures 5.2–5.3 show that overall the NO_2 concentration levels are reduced in most cases under the policy intervention (the green curve is below the red curve in most time points). There are a few exceptions to this reduction, which is quite natural since the NO_2 concentration has a complex behaviour that may depend on many variables in the dataset

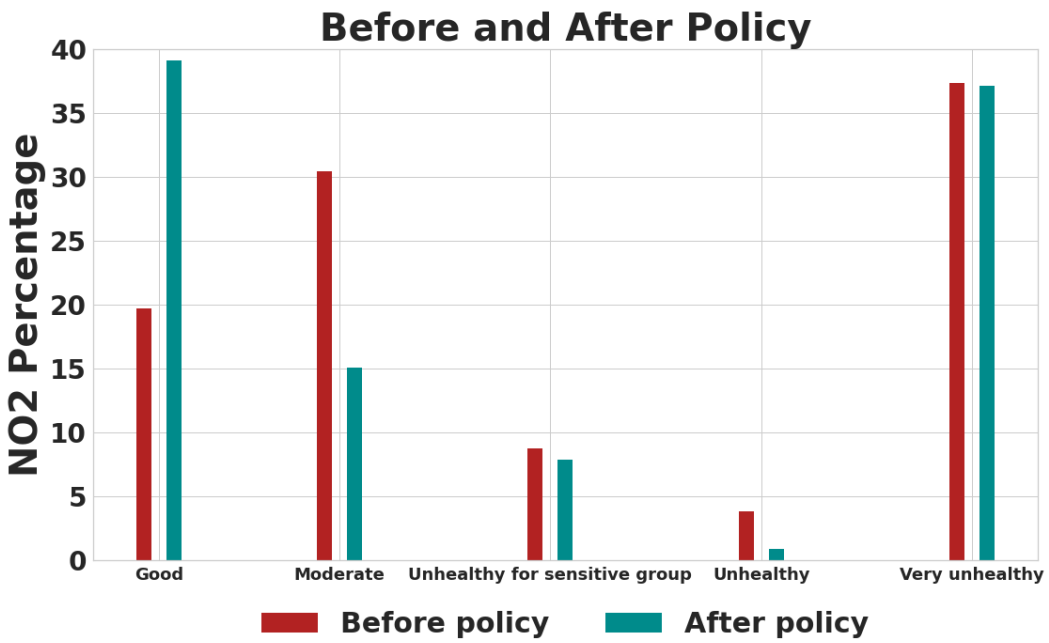

Fig. 5.1 Percentages of each class of NO_2 (number of hourly NO_2 measurements in each class divided by the total number of hours)

from the previous time instances, thus the two curves for NO_2 concentration levels before and after the policy intervention should not be compared is a point-by-point fashion (they are the outputs of two LSTM models trained on different datasets). The air quality is expected to be improved in average and not throughout a time period. This is backed by all three Figures 5.1–5.3 (concentration levels may increase in some time instances but will reduce in many other time instances).

While a clean air zone can lead to substantial improvements in air quality, achieving consistent year-round pollution reduction requires the clean air zone to be complemented by broader measures including addressing various pollution sources and promoting sustainable transportation alternatives.

5.4.1 Outcome of the Framework Applied to the Dataset

The number of NO_2 concentrations that are unhealthy is the sum of three classes: Very Unhealthy, Unhealthy, and Unhealthy for Sensitive Group. According to Figure 5.1, this is $37.34\% + 3.77\% + 8.74\% = 49.85\%$ without implementing the intervention and is $37.16\% + 0.84\% + 7.84\% = 45.84\%$ with the implementation of the intervention. This shows the total reduction of $49.85 - 45.84 = 4.01\%$ and relative reduction of $4.01/49.85 = 8\%$. Therefore,

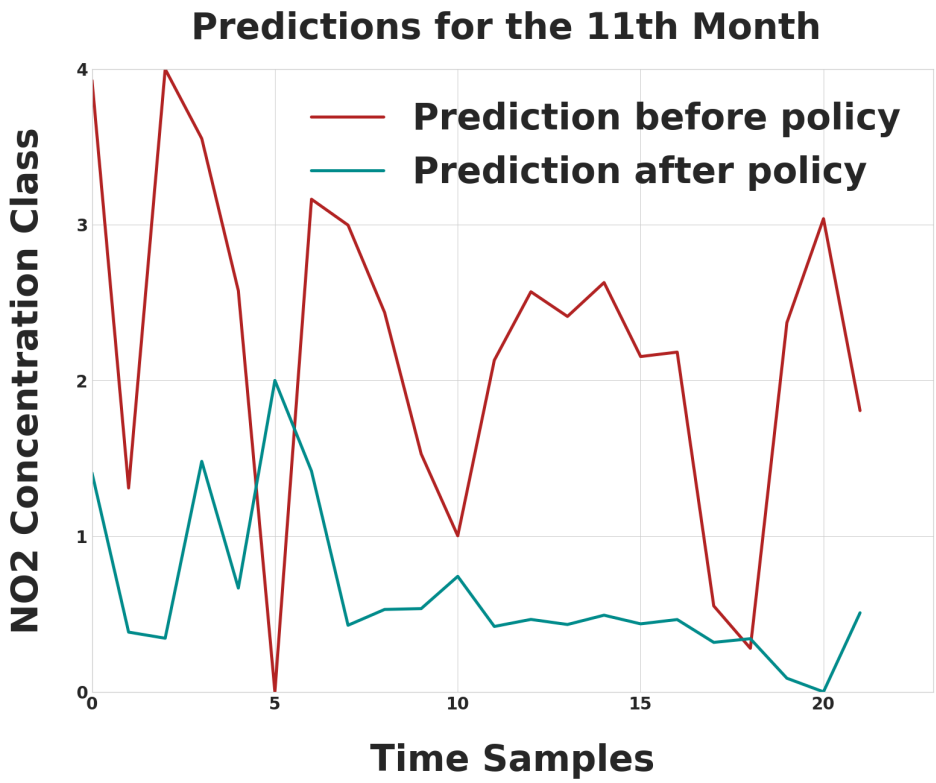

Fig. 5.2 LSTM prediction for NO_2 concentrations for the 11^{th} month

the objective of the intervention which is 10% reduction will not be achieved and further adjustment of the intervention may be needed to reach the 10% reduction.

5.4.2 Time Series Model Evaluation

The RMSE from Equation (5.1) is used to assess the performance of the learned LSTM models. The value of the RMSE for the model without the intervention is 0.946, while the RMSE for the model with the intervention is 0.850. This shows that the two models are performing similar to each other in terms of capturing the behaviour of data. Figure 5.5 shows the loss value of the training with and without the intervention as a function of epoch number (i.e., the number of times that the learning algorithm will work through the entire training dataset for updating the model). The loss starts from a high value and gradually decreases until converging to a fixed value while the learning algorithm finds the best parameters for the LSTM model.

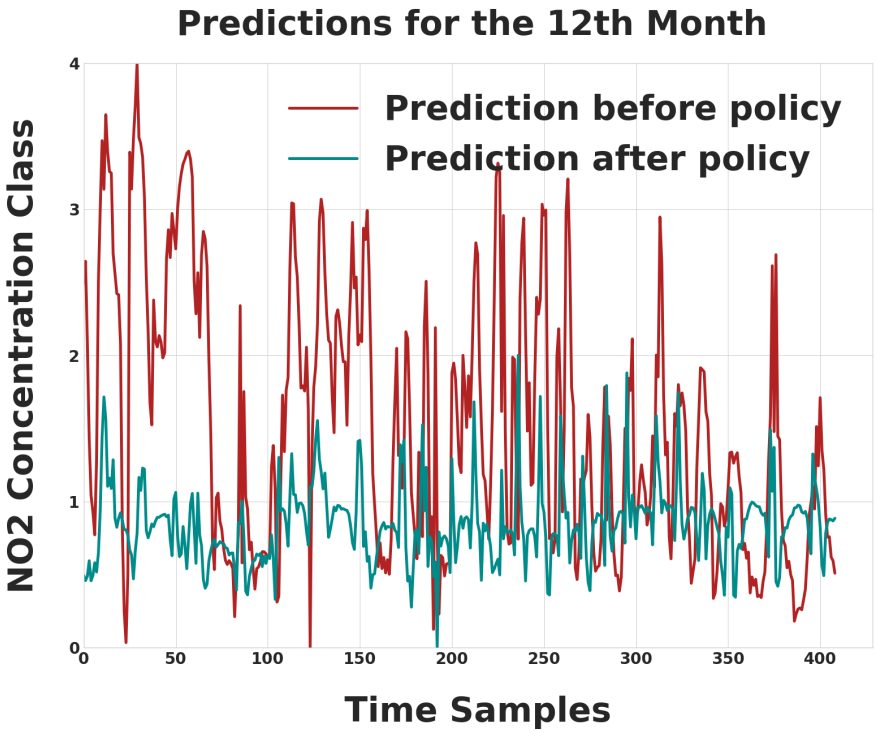

Fig. 5.3 LSTM prediction for NO_2 concentrations for the 12^{th} month

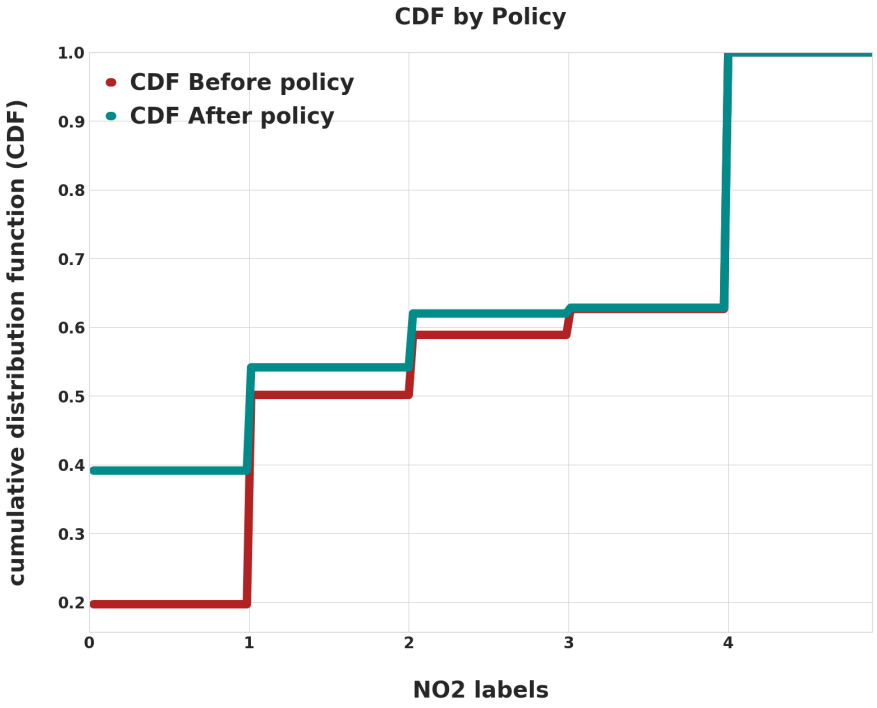

Fig. 5.4 CDF of the LSTM method with and without the intervention

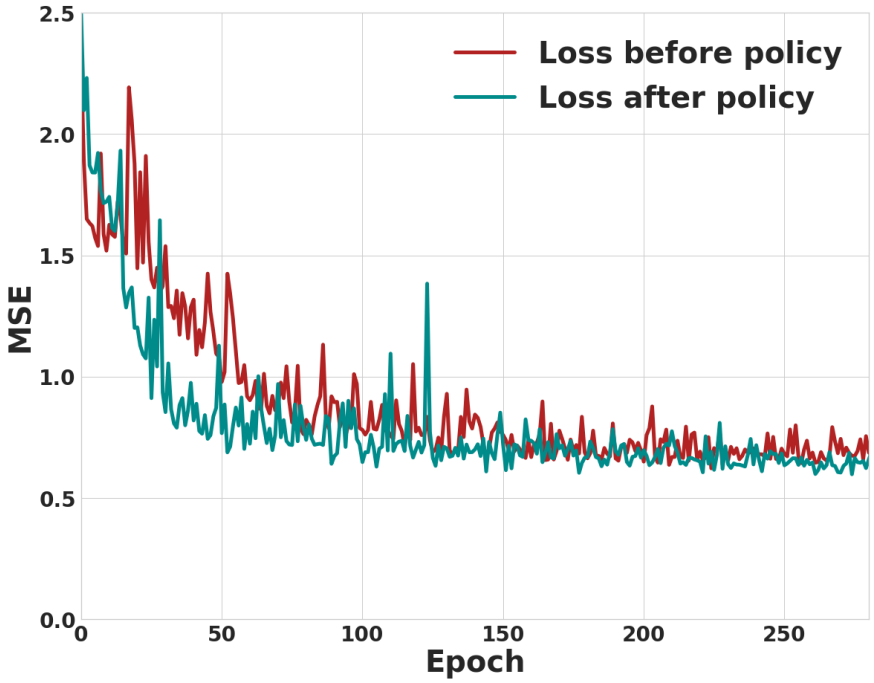

Fig. 5.5 LSTM loss with and without the intervention as a function of epoch number

5.5 Conclusions

In this chapter, the potential impacts of a clean air zone intervention were investigated using a time series analysis method, specifically Long Short-Term Memory (LSTM). The aim of the proposed intervention was to reduce the NO_2 concentrations. This chapter showed how to apply the framework in Figure 3.3 and presented in Section 3.5 to this intervention to tackle the specific challenge of validating the policy objective using machine learning techniques. However, since the intervention is in the early stage of implementation at Newcastle, an LSTM model was used to predict the NO_2 levels under appropriate assumptions.

The LSTM model was employed to analyse the historical data from 2018 and to predict NO_2 concentrations under existing conditions, simulating a scenario without the clean air zone intervention. Subsequently, using a set of assumptions to represent the effects of a clean air zone, the LSTM model was used again to forecast NO_2 concentrations in the presence of the intervention. This comparison provides a valuable understanding of the potential effectiveness and impacts of such a clean air zone policy.

The historical data from the first 10 months were used to build and evaluate the LSTM model, and the predictions were made for the last two months. It was observed that the LSTM model can successfully predict the NO_2 concentrations with root mean square error of 0.95. This result indicates that time series forecasting with LSTM is a powerful method

for predicting air pollution levels in a given location over time. By analysing historical data on air pollution levels and other relevant factors, LSTM models can identify patterns and trends that can be used to predict future levels of pollution with a high degree of accuracy. This can help policymakers to plan and implement policies that are better aligned with the actual trends and patterns of air pollution in their area.

The framework used in this chapter is flexible and can be adapted to different policy objectives. It does not incorporate information from physical or chemical models of pollutants but offer proof-of-concept in situations where post-policy implementation data is not yet available. The presented results show the use of machine learning methods in analysing and validating interventions in transportation systems. The role of machine learning can be summarised as predicting what is going to happen in the future if the policy is not implemented (using available historical data), and predicting the air quality and other related variables using transport behaviour changes in response to the implemented policy.

The next chapter will focus on applying the data-driven framework proposed in Figure 3.5 of Chapter 3 for the implementation of the policy commitment on use case of expanding the electric vehicle charging infrastructure.

Chapter 6

Multi-Objective Optimisation & Electric Vehicle Charging Infrastructure

This chapter applies the framework of Chapter 3 that combines simulation and multi-objective optimisation to efficiently implement transportation policy commitments, using as a case study the electric vehicle (EV) charging infrastructure in Newcastle upon Tyne. The framework uses a baseline simulation model developed by the industry partner, Arup Group Limited, to estimate EV demand and quantities from 2020 to 2050. A multi-objective optimisation approach is then employed to determine the optimal types, locations, and quantities of charging points, along with the corresponding total capital and operational expenditures (CapEx and OpEx) and charging point operating hours.

Four future energy scenarios are considered, providing predictions of EV quantities and energy demand for Newcastle upon Tyne. The optimisation results highlight the benefits of diversifying charging point types, demonstrating reduced total expenditure while maintaining satisfactory performance in meeting charging demand. Sensitivity analysis further confirms the importance of charging point diversity in optimising the charging infrastructure.

Quantitatively, the optimal solutions recommend installing higher number of fast charging points to reduce the percentage of slower charging points from the current 60% to around 25% in the four scenarios. The optimal solutions still put priority on the slower charging points (around 25%), with faster charging points having smaller portions each around 10%-13%. The optimisation shows that while 7kW charging dominates the market currently, it is more beneficial to improve charging efficiency and reduce investment costs with other types of charging points in the future installations. Moreover, in the leading scenario for the year 2042 with 134,606 EVs, a total of 4,753 charging points are recommended, resulting in an average operating time of 7.57 hours per charging point. The results also illustrate the spatial distribution of charging points, with higher concentrations in urban areas and near major

roads. Note that the variations of the portions of different types of charging points for the four scenarios are relatively small and within 3% range of the total number of charging points.

The methodology's robustness is demonstrated through sensitivity analysis, examining variations in EV numbers and charging point types. The analysis reveals the adaptability of the optimisation framework to different scenarios, providing valuable insights for decision-making and highlighting potential areas for improvement.

Overall, the results of this chapter offer a scientific and comprehensive approach to support the implementation of transportation policy commitments, particularly in the context of achieving net-zero emissions. It serves as a valuable tool for optimising EV charging infrastructure and can be applied to other regions and datasets to inform evidence-based decision-making.

This chapter is organised as follows. After an introduction in Section 6.1, With focus on the expansion of EV charging infrastructure and data use, Section 6.2 maps the generic steps of the policy cycle described in Section 3.1 to the steps taken in defining and evaluating the plans for expansion of the EV charging infrastructure. A scheme for applying the third framework of this thesis to the case study on expansion of the EV charging infrastructure is provided in Section 6.3. A description of data collection and preparation is provided in Section 6.4 for the geographical area of Newcastle upon Tyne. Regarding the geographical area, the area of Newcastle upon Tyne is divided into smaller unified regions, called LSOAs, and are used to perform a refine analysis of the EV charging infrastructure. The EV charging simulation model is discussed in Section 6.5 together with the details of the estimation methods for finding the future number of EVs and their total energy demand. Section 6.6 discusses the novel multi-objective genetic algorithm and apply it to the EV infrastructure planning. The results of applying the framework based on the optimisation are discussed in Section 6.7. Sensitivity analysis of the results of the multi-objective optimisation are provided in 6.8. Further considerations and concluding remarks are provided in Sections 6.11–6.12. The code for reproducing the results of this chapter is provided in Appendix D.

The content of this chapter is based on the following articles published during the PhD study.

- Farhadi, Farzaneh, Shixiao Wang, Roberto Palacin, and Phil Blythe. "Data-driven multi-objective optimization for electric vehicle charging infrastructure." iScience (2023). DOI: https://doi.org/10.1016/j.isci.2023.107737

- Farhadi, F., Wang, S., Palacin, R. and Blythe, P., 2023. Efficient electric vehicle charging infrastructure planning using data-driven optimization. Institution of Engineering and Technology (IET). DOI: https://doi.org/10.1049/icp.2023.3116

- Farhadi, F., Palacin, R. and Blythe, P., 2023. Data-Driven Framework for Implementing Policy Commitment, Universities Transport Study Group (UTSG), Cardiff (July 2023), Nominated for Smeed prize.
- Farhadi, F., Palacin, R. and Blythe, P., 2023. Multi-Objective Optimisation Method for Electric Vehicle Charging Infrastructure, 8th Annual Electric Vehicle Conference, Edinburgh Napier (June 2023).

6.1 Introduction

In order to help local authorities and other governmental organisations plan for the EV charging infrastructure, this chapter applies the framework proposed in Chapter 3 using the scheme described in Section 6.3 to the case study on expansion of the EV charging infrastructure. As part of the scheme, an optimisation method based on a modified genetic algorithm is used. The goal of the optimisation is to consider and optimise the following factors needed to design and expand the EV charging infrastructure

1. Charging point type,
2. Charging point location,
3. Charging point quantity,
4. Total capital and operational expenditures, and
5. Operating hours of charging points.

The open-access statistics published by the official authorities are used in the modelling and optimisation developed in this chapter. Section 6.4 gives the references for data collection and preparation. The geographical divisions and vehicle data are presented and analysed in Chapter 6.4. LSOAs of Newcastle are selected as geographical divisions that are small enough for capturing essential details in an accurate simulation model, and at the same time large enough for reducing computational complexity of the developed simulation model. In order to better investigate multiple aspects of EV infrastructure planning at the same time, this research uses genetic algorithm, which is improved based on the concepts of Long Short-Term Memory (LSTM) networks and fuzzy logic. The main contributions of this chapter are as a follows.

1. By drawing on LSTM networks from machine learning literature, the traditional genetic algorithm is extended and combined with fuzzy logic to design a multi-purpose decision model for multi-objective optimisation problems.

2. The developed new optimisation framework is used to optimise multiple objectives such as total capital and operational expenditures, and charging point efficiency.

3. The designed model removes the need to compress multiple objective functions into a single objective function. Instead, the underlying simulation environment is modelled using actual data, enabling a transition from function-driven to data-driven optimisation and evaluation.

4. An implementation of the computations are provided using vector and matrix representations. Matrices give a compact way of handling large volumes of data and updating values efficiently.

6.2 Steps of the Policy Cycle for Expansion of the EV Charging Infrastructure

This section discusses how the generic steps of the policy cycle described in Section 3.1 are mapped into the steps taken in defining and evaluating the expansion of the EV charging infrastructure.

1. **Issue Identification:** The need to reduce carbon emissions from the transport sector is identified as a critical goal to meet the UK's net zero targets by 2050. The "North East Combined Authority (NECA) Transport Manifesto" (2016) and the "North East Transport Plan" (2021) identified the need to reduce carbon emissions from road transport to meet the UK's climate goals and improve air quality. Increasing the adoption of EVs is seen as a key solution, but insufficient charging infrastructure is identified as a barrier to EV uptake. The NECA identified that the limited availability of public EV charging infrastructure was a barrier to the uptake of EVs.
   Data Use: Data showing the correlation between EV adoption rates and the availability of charging stations highlights gaps in current infrastructure. Local transport data is also used to identify areas with high potential demand for EV charging.

2. **Aims:** The aim is to expand EV charging infrastructure to support the transition to electric vehicles, reduce greenhouse gas emissions, and improve air quality. Specific objectives may include increasing the number of public chargers by a certain percentage or ensuring that every resident is within a reasonable distance of a charging station. "Go Ultra Low North East Programme Outline" (2016) set out specific aims to support EV adoption in the North East by expanding charging infrastructure and increasing

public awareness of EV benefits.
Data Use: Baseline data on the current number and distribution of charging stations is used to set specific, measurable targets (e.g., installing a specific number of new chargers by a certain time frame). For example, the objective mentioned in the Go Ultra Low North East Programme was to create a network of public charging points that would support the region's goal of becoming a leader in sustainable transport. The plan aimed to install charging hubs in key locations, such as urban centres and major travel routes.

3. **Problem Analysis and Appraisal:** Authorities analyse the existing EV charging network, user demand patterns, grid capacity, and the costs and benefits of various expansion strategies. This involves assessing different types of chargers (rapid, fast, slow) and suitable locations (e.g., residential areas, highways, shopping centres).
Data Use: Geographic Information System (GIS) data, traffic flow analysis, and user surveys are used to identify optimal locations for new chargers. Data on grid capacity and power demand forecasts help appraise the feasibility of different expansion options. "Electric Vehicle Charging Infrastructure Strategy for Newcastle" (2024) analysed existing charging infrastructure and current and projected demand. The analysis emphasises using data and transport modelling to identify gaps in the charging network and prioritise areas for new charging points.

4. **Preferred Option and Feasible Objectives:** A preferred strategy for expanding the EV charging network is selected. This might include a mix of public and private investment, partnerships with businesses and energy providers, and prioritising specific types of chargers or locations.
Data Use: Cost-benefit analysis data is used to evaluate different expansion options. Market data on EV adoption trends and charging behaviour informs decisions on the mix and placement of chargers.

5. **Final Consultation:** A public consultation phase is conducted to gather input from key stakeholders, such as local communities, businesses, energy providers, and EV users, on the proposed expansion plans.
Data Use: Data from surveys, public meetings, and feedback submissions is analysed to gauge support, address concerns, and refine the policy. Feedback is also sought from technical experts on infrastructure requirements and grid impacts.

6. **Decision:** Local authorities, in collaboration with national government bodies and private sector partners, make a final decision on the EV charging infrastructure expansion

plan.
Data Use: Aggregated data from consultations, market analysis, and technical assessments is used to justify and support the decision. Risk assessments ensure any potential challenges, such as grid constraints or public opposition, are managed.

7. **Implementation and Delivery:** The chosen strategy is implemented, including securing funding, installing chargers, and integrating with the existing grid and transport infrastructure. A roll-out plan is created to guide the deployment of chargers over time.
Data Use: Real-time data from smart chargers is used to monitor usage patterns and identify areas of high demand. Data from grid operators is also used to ensure sufficient electricity supply and manage load distribution.

8. **Maintenance, Monitoring, and Review:** The infrastructure is regularly maintained to ensure reliability and performance. Authorities continuously monitor usage patterns, grid impacts, and public feedback to ensure the network meets demand.
Data Use: Data from smart charging systems, user feedback, and grid operators is analysed to identify maintenance needs, optimise charger placement, and adapt to changing demand patterns. Regular reporting ensures transparency and accountability.

9. **Evaluation:** A comprehensive evaluation is conducted to determine whether the expansion has met its objectives, such as increasing EV adoption rates and reducing emissions.
Data Use: Evaluation involves comparing post-implementation data (e.g., number of EVs, charging station usage, emissions levels) with baseline data. Surveys and user feedback are also collected to assess user satisfaction and identify areas for further improvement.

Throughout the expansion of EV charging infrastructure and at each step, data is crucial for informing decisions, setting objectives, analysing options, engaging stakeholders, implementing policies, monitoring progress, evaluating outcomes, and refining the expansion policy throughout its life cycle.

This chapter is focused on applying the proposed framework for optimising policy implementation proposed in Chapter 3 to the expansion of EV charging infrastructure in Newcastle. This framework can be integrated with steps 6 and 7 of the policy cycle in Decision and Implementation/Delivery as discussed in Fig 3.2. The framework was described in Figure 3.5 that requires building a simulation model. This simulation model will be described in the next section.

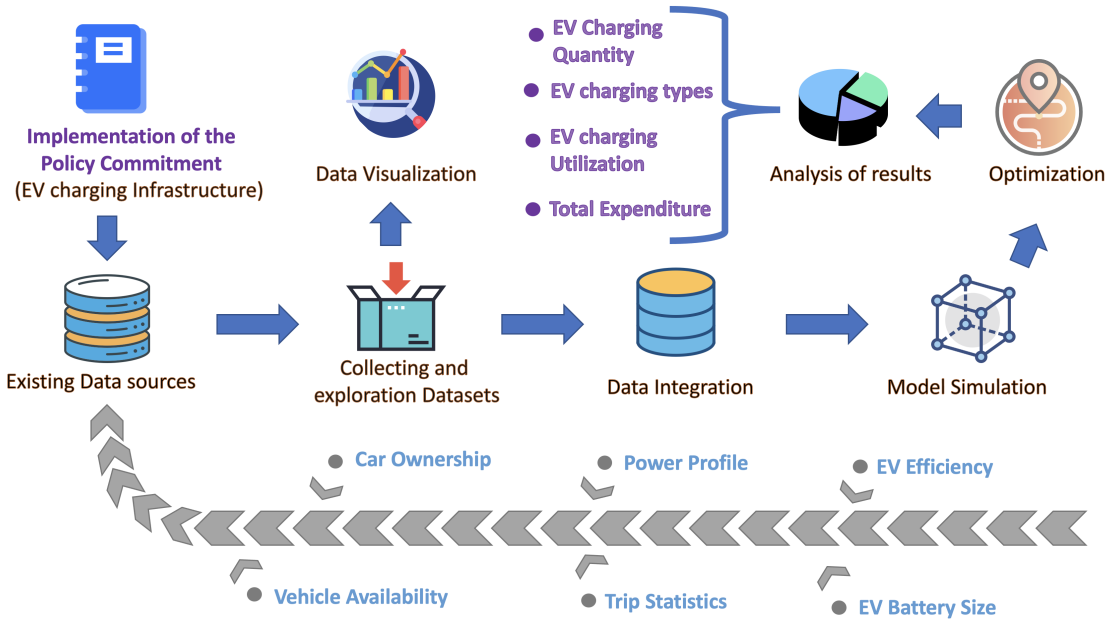

Fig. 6.1 A scheme for applying the framework of Figure 3.5 to the case study on expansion of the EV charging infrastructure

6.3 Scheme for Building the Simulation Model

The scheme for building the simulation model and integrate it with optimisation is presented in Figure 6.1. The scheme shows the process of gathering the suitable datasets relevant to the policy commitment and build a model that can simulate future scenarios for optimisation purposes to achieve a better implementation of policy commitments in transport systems. The following steps are taken to build the scheme.

1. Selecting transport policy commitment from net zero emission strategy as a case study.
2. Analysing the policy commitment of providing EV charging infrastructure by performing a suitable literature review in order to understand the objectives and outcomes of the policy commitment.
3. Analysing assumptions, datasets, and calculations in the construction of the baseline simulation model.
4. Analysing future energy scenarios developed by National Grid and used for constructing the simulation model.
5. Computing the estimated power demand and EV quantities for the years 2020–2050 for four future energy scenarios.

6. Building an optimisation approach using the simulation model for providing an efficient expansion of EV electrical infrastructure with respect to the charging point quantities, their types, locations, costs, and operating hours.

In developing the optimisation solution for this research, several innovative approaches have been employed to address the limitations of traditional methods, enhancing the accuracy, flexibility, and overall effectiveness of the solution. These innovations are as follows:

- Integration of data with a simulation model: To overcome the limited capabilities of traditional genetic algorithms, the optimisation method incorporates available data with a simulation model. This integration not only enhances the decision-making process but also produces more accurate and realistic solutions tailored to the specific context of electric vehicle charging infrastructure planning.

- Fuzzy logic for objective ranking and combination: When handling multiple objectives in optimisation problems, the proposed method employs fuzzy logic to combine and rank the various objectives. This approach is flexible and intuitive, eliminating the need to compress multiple objectives into a single one. As a result, the optimisation solution delivers more comprehensive and balanced outcomes that accurately reflect real-world priorities and trade-offs.

- Inspiration from Long Short-Term Memory (LSTM): Recognising that traditional optimisation methods often struggle to remember useful information from previous generations, the optimisation solution incorporates concepts from LSTM in machine learning. This innovation enables the optimisation method to retain important information over time, thereby improving its performance by leveraging knowledge from previous generations.

These key innovations demonstrate the methodological advancements made in the optimisation solution, resulting in a more effective and robust approach to electric vehicle charging infrastructure planning. By integrating data with simulation models, employing fuzzy logic for objective ranking and combination, and incorporating the concept of LSTM, the proposed optimisation solution addresses the challenges faced in traditional methods and delivers more reliable and practical results.

The overview of the genetic optimisation solution inspired by the LSTM model and using fuzzy logic has been summarised in Figure 6.2. The optimisation uses a simulation environment that simulates the EV energy demand, EV quantity, and EV locations. The EV charging points are considered as genes in the genetic algorithm. Each gene encodes variables including the location and type of the charging point. These genes are generated

randomly. The performance of the genes in the simulation environment is assessed based on the utilisation and costs of the charging points. Then some of the genes with high performance are kept and new genes are generated randomly again. The iteration is continued until converging into a solution.

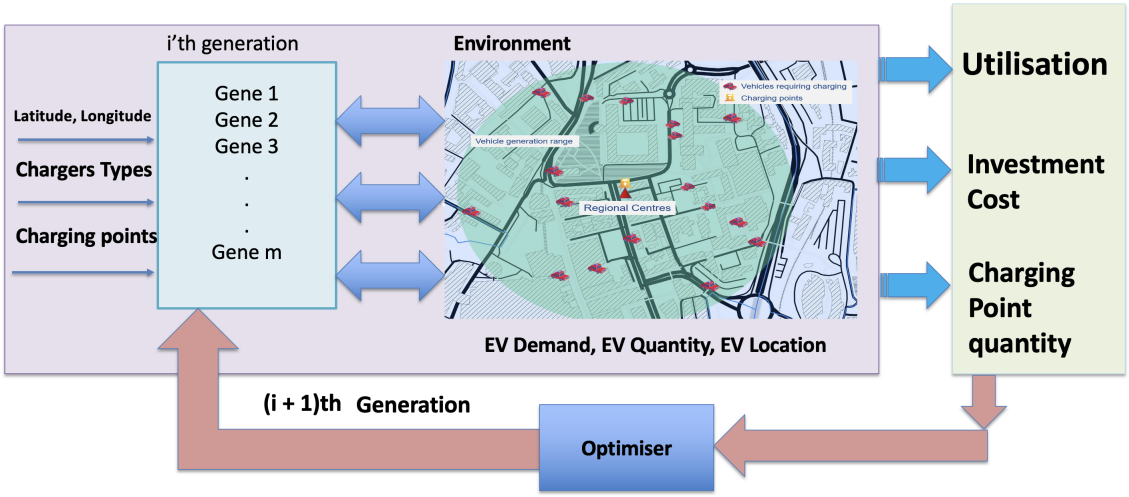

Fig. 6.2 Overview of the genetic optimisation solution inspired by the LSTM model and using fuzzy logic

In conclusion, the integration of fuzzy logic and LSTM with genetic algorithm has shown promising results in solving multi-objective optimisations by handling complex and non-linear problems and improving the convergence speed and accuracy of genetic algorithms.

6.4 Data Collection and Preparation

The city of Newcastle upon Tyne in the United Kingdom has been chosen as the geographical area for this research, which aims to expand the electric vehicle charging infrastructure. The UK government has set a target of phasing out petrol and diesel cars by 2030 and has designated Newcastle upon Tyne as one of the cities to receive funding for expanding electric vehicle charging infrastructure. The city's existing infrastructure, including its public transport system and electric vehicle charging points, provides a good basis for the expansion of the electric vehicle charging network. Through this research, the expansion of electric vehicle charging infrastructure can be studied in depth.

There are three geographical divisions used for statistical purposes. These are called *Output Areas* (OAs), *Lower Layer Super Output Areas* (LSOAs), and *Middle Layer Super Output Areas* (MSOAs). The number of people and households in any of these areas is stated in Table 6.1 (data from Office for National Statistics). The OAs are the smallest division.

The LSOAs have an average of 1,600 residents and 670 households. Currently, there are 34,753 LSOAs in England (32,844) and Wales (1,909). MSOAs have an average population of 7500 residents or 4000 households.

Table 6.1 Geographical divisions used for statistical purposes (data from Office for National Statistics)

Area type	Lower threshold		Upper threshold	
	People	Households	People	Households
Output Areas	100	40	625	250
Lower Layer Super Output Areas (LSOAs)	1,000	400	3,000	1,200
Middle Layer Super Output Areas (MSOAs)	5,000	2,000	15,000	6,000

In order to divide the geographical area of Newcastle upon Tyne to smaller unified regions, LSOAs are used to perform a refine analysis of the EV charging infrastructure. Figure 6.3 shows the LSOAs of Newcastle upon Tyne on its map. This map contains 175 LSOAs. The Table 6.2 shows the data of the LSOAs with 175 entries and with latitudes and longitudes of the centre points of each LSOA.

For this chapter, a data-driven baseline model is used from the industrial partner of this PhD project, Arup Group Limited. Datasets have been collected from different sources such as the Department for Transport, National Grid, Census, Bloomberg Professional Services, Open Charge Map, Geotab and Met Office. These datasets provide valuable information on electric vehicle charging infrastructure simulation for demand estimation. The outputs of the simulation have been used for making the simulation environment as part of the proposed framework. These datasets are described in the next subsection.

6.4.1 Data for EV Charging Infrastructure

A summary of the datasets used in the baseline demand prediction is as follows.

1. **Department for Transport**:
    - National Trip End Model (NTEM) (Transport, 2022) is a travel generation model designed by the Department for Transport. It can predict travel conditions in England and Scotland using official population, household and employment data.
        - Car Ownership (Transport, 2022): Forecast registered vehicles at the MSOA level.
        - Region Trip End Data By Availability (Transport, 2022): Forecast quantity of trips ending at the MSOA level by purpose, mode and car availability.

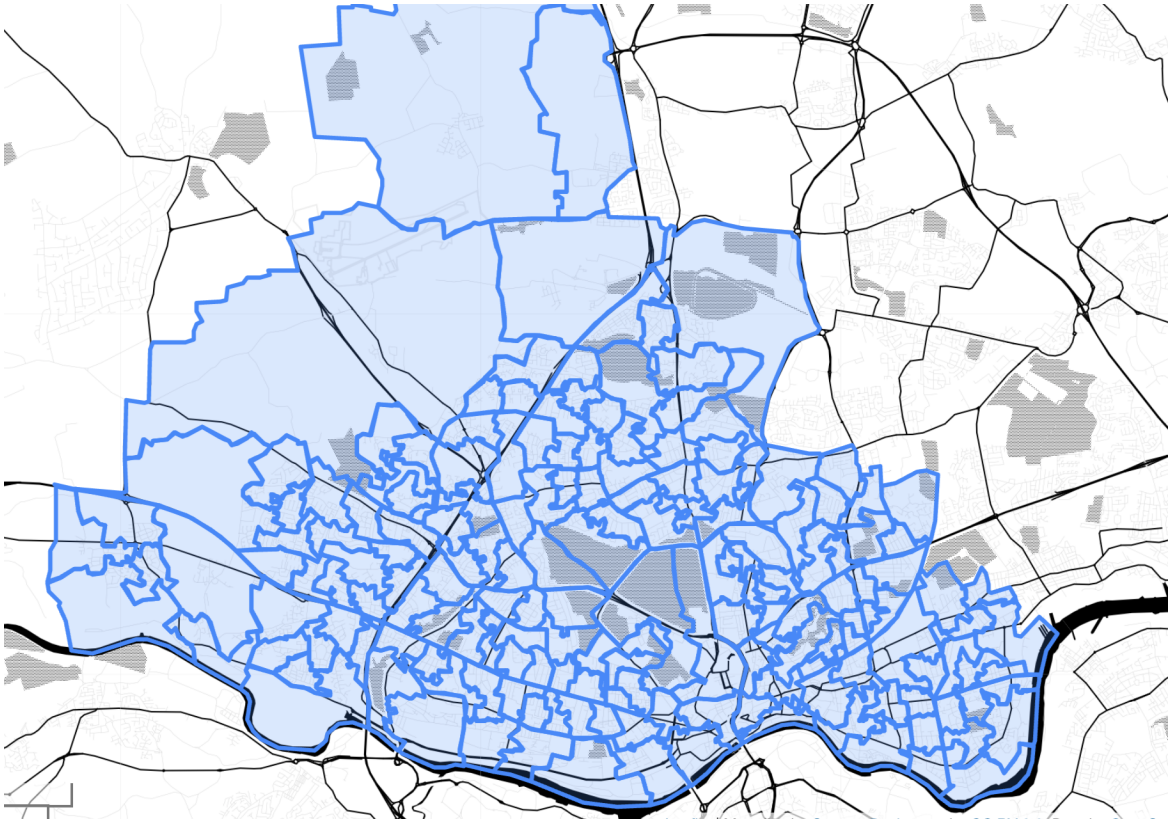

Fig. 6.3 LSOAs of Newcastle upon Tyne with 175 LSOAs in the map as geographical divisions

Table 6.2 LSOAs of Newcastle upon Tyne with latitudes and longitudes of the centre points of each LSOA

	LSOA/DZ centre point latitude	LSOA/DZ centre point longitude
0	54.9765	-1.67682
1	54.9756	-1.66885
2	54.9684	-1.66731
3	54.9669	-1.65558
4	54.9794	-1.68239
...	...	...
170	55.0026	-1.70740
171	54.9974	-1.68732
172	54.9973	-1.70406
173	54.9967	-1.71593
174	55.0059	-1.71914

  - Region Trip End Data By Direction(Transport, 2022): Forecast quantity of trips ending at MSOA level by purpose, mode, time period and trip type.
  - Zones (Transport, 2022): Matching Zone IDs, MSOA names and Local Authority.

- Charger Profiles
  - Electric Charge Point Analysis 2017 Domestics Tables (Transport, 2018a): Quantity of domestic home charging point plug-in events against time of day.
  - Electric Charge Point Analysis 2017 Public Sector Fasts Tables (Transport, 2018c): Quantity of public fast charging point plug-in events against time of day.
  - Electric Charge Point Analysis 2017 Rapids Revised (Transport, 2018b): Quantity of public rapid charging point plug-in events against time of day.
- National Travel Survey
  - Average trips (Planning, 2013): Data on number of trips made and distance travelled, produced by Department for Transport.

  - Annual mileage of cars by ownership (Driving and Transport, 2013a): Data on vehicle mileage and occupancy, produced by Department for Transport.
  - Average distance travelled by rural urban classification (Driving and Transport, 2013b): Data about travel by region and mode of transport, produced by Department for Transport.

2. **National Grid** (NationalGrid, 2022b): The report by National Grid titled “A Net Zero Future" presents a range of different, credible ways to decarbonise the energy system. In another report titled “Future Energy Scenarios" (NationalGrid, 2022a), the National Grid provides forecasts of the quantity of battery electric vehicles (BEV) on the road and total vehicles assuming different scenarios.

3. **Census**:

   - Office for National Statistics (Statist, 2021) provides the quantity of households in each LSOA by dwelling type: detached, semi-detached, terrace, flat, apartment and mobile home.
     - Nomis is a service provided by the Office for National Statistics (Nomis, 2013b). It provides the quantity of households in each data zone by dwelling type: detached, semi-detached, terrace, flat, apartment and mobile home.
     - Nomis provides the statistics of car and van availability (Nomis, 2013a): Quantity of cars or vans in households for each data zone. The dataset has the following classification of households: no car, 1 car, 2 cars, 3 cars, 4 or more cars.
     - Nomis provides the statistics of movement of people and distance between local authorities (LAs) (Nomis, 2013c). Data is showing the living location, working time, and working location of people.
   - Central Statistics Office (Census, 2016): Quantity of households in each SA by dwelling type: detached, semi-detached, terrace, flat, apartment and mobile home.

4. **Bloomberg Professional Services** provides a detailed dataset of battery electric vehicles and Plug-in hybrid electric vehicles in the market. The dataset contains the battery size, year, make, model and price of the cars (Services, 2022).

5. **Open Charge Map** (OpenChargeMap, 2021): A detailed list of registered charging points.

Table 6.3 Datasets used in EV charging baseline model

Source	Details
Department for Transport	– National Trip End Model: Forecast registered vehicles and trips at the MSOA level (Transport, 2022). – Charger Profiles: Quantity of charging point plug-in events (Transport, 2018a,b,c). – National Travel Survey: Data on trips, mileage, and distance travelled (Driving and Transport, 2013a,b; Planning, 2013).
National Grid	– Reports on decarbonising the energy system and forecasts of BEVs (NationalGrid, 2022a,b).
Census	– Office for National Statistics: Quantity of households by dwelling type (Nomis, 2013a,b,c; Statist, 2021). – Central Statistics Office: Household data by dwelling type (Census, 2016).
Bloomberg Professional Services	– Dataset of battery and hybrid electric vehicles (Services, 2022).
Open Charge Map	– List of registered charging points (OpenChargeMap, 2021).
Geotab and Met Office	– EV efficiency against temperature (Geotab and Office, 2022).

6. **Geotab and Met Office** (Geotab and Office, 2022): Data defining EV efficiency against temperature.

To have a reliable prediction of the number of EVs and simulate a realistic environment, the above datasets are used. The general approach that is taken in this chapter is to predict the average vehicle mileage demand in each region and then obtain the EV charging demand using the datasets from the number of regional vehicle registrations and regional vehicle miles travelled. The trips are then considered to be either *home based* or *non-home based*. A home-based trip is a trip that starts and ends in the same LSOA. A non-home based trip is a trip that starts and ends in different LSOAs. Table 6.3 shows a summary of the above datasets.

In the next subsection, the above datasets are analysed.

6.4.2 Analysing Data for EV Charging Simulation Model

As the first step in the framework of Figure 6.1, the relevant historical data is analysed to gain an understanding of the current behaviour in the EV infrastructure and its usage. The key statistics of the current charging provisions in the UK are as follows:

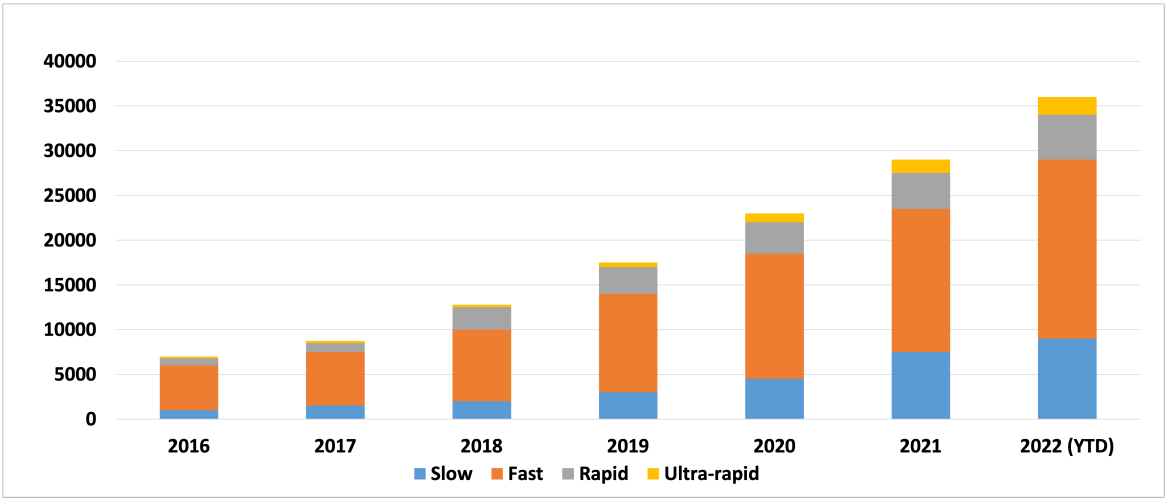

Fig. 6.4 Number of public UK charging points between 2016 and 2022 (data from Zap-Map)

- Figure 6.4 shows the number of public UK charging points from 2016 until now (data from Zap-Map retrieved in October 2022). The charging points are divided into four categories based on their speeds or power ratings: Slow (3-6kW), Fast (7-22kW), Rapid (25-99kW) and Ultra-rapid (100kW+). The number of these charging posts across UK in 2022 are: 8,702 Slow points, 20,568 Fast points, 4,373 Rapid points, and 2,135 Ultra-rapid points. Therefore, there are 35,778 Slow to Fast charging points, compared to 6,508 Rapid and above charging devices.

- Figure 6.5 shows the number of public UK charging points per 100,000 of population by UK country and region (data from DfT on January 2022). On average there are approximately 35 charging devices per 100,000 population, excluding London as an outlier.

- Considering the two categories of Slow to Fast and Rapid to Ultra-Rapid points, the average split of chargers between these two categories is 77 and 33.

Figure 6.6 shows the distribution of charging points by geographical area in the UK (data from Zap-Map retrieved in October 2022). Statistics similar to the above can be obtained for Newcastle upon Tyne, and gives results that show the same trend normalised with the respective population of Newcastle upon Tyne. The key statistics of the current charging provisions in this local authority is as follows. The detailed data is provided on the DfT website.[1]

- The total public charging points in Newcastle upon Tyne is 146.

[1] https://maps.dft.gov.uk/ev-charging-map/index.html

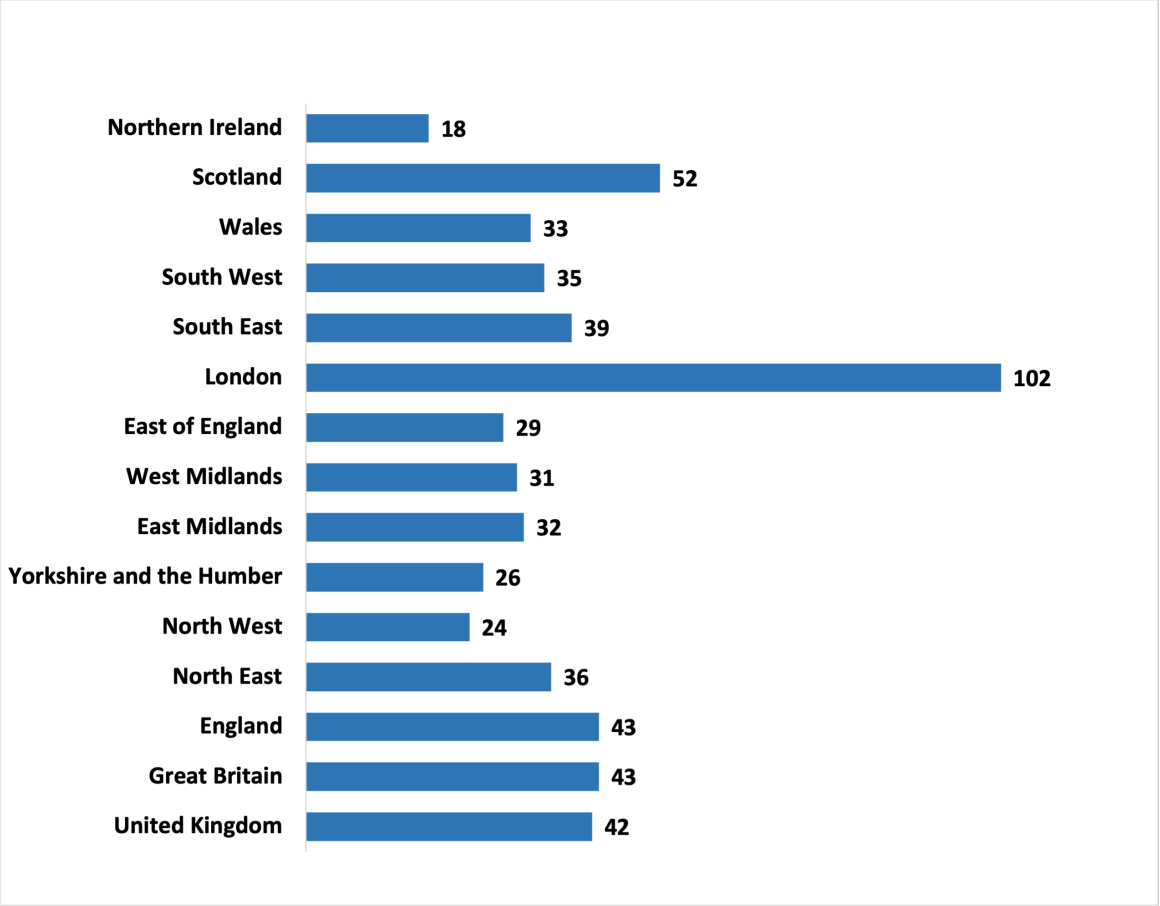

Fig. 6.5 Number of public charging points per 100,000 of population by UK country and region (data from DfT)

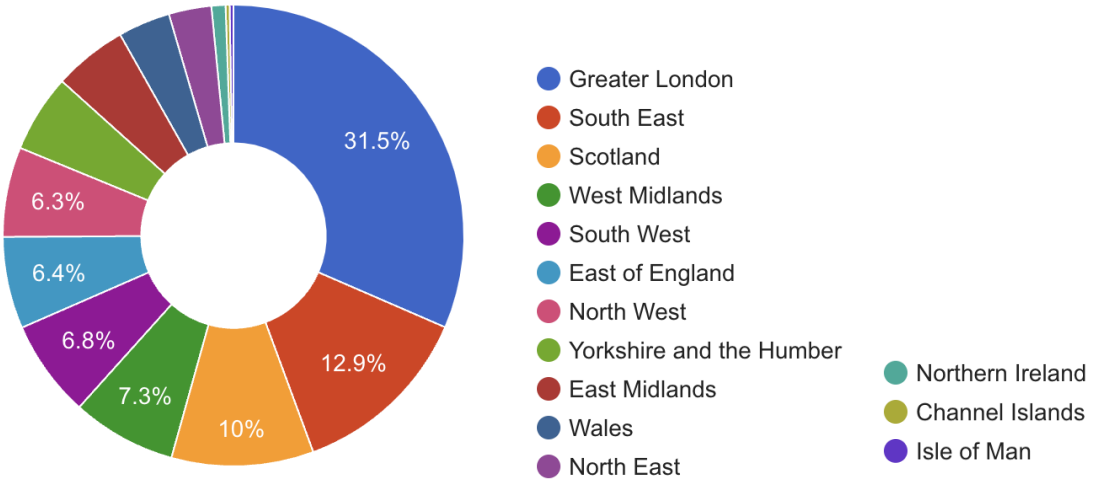

Fig. 6.6 Distribution of charging points by geographical area in the UK with the total charging points of 35,778 (data from Zap-Map)

- Total public rapid charging devices is 28.
- The number of charging points per 100,000 population is 47.6.

Note that the charging points installed privately at home or at workplace locations are not included in the above statistics, which are estimated to be more than 400,000 (ZapMap, 2022a). This is due to the fact that the focus of this study is on the location and quantity of charging points that will be installed by local authorities under the related policy commitments. The new regulations set by the UK government should see up to 145,000 extra charging points installed across England each year in the run up to 2030.[2] These statistics show the current status of the available EV charging infrastructure, which should be expanded as the number of EVs increases in the future.

To properly predict the requirements for expanding the EV charging infrastructure, the current demand for road use by EV's are analysed. The Department for Transport (DfT) has provided the statistics of the road use in a report (Department for Transport, 2016). The report integrates the information and data sources on vehicles, travel, and traffic. Relevant to the work of this chapter, the report identifies the percentage of car trips by mileage. The distribution of travel mileage has stayed the same (number of trips with the same mileage divided by the total number of journeys has remained roughly the same). Short trips less than 5 miles account for 56% of total trips. These statistics also show that 94% of all car trips were less than 25 miles. This portion has not changed since 2002. Therefore, it is essential to

[2] https://www.zap-map.com/electric-vehicle-charging-2022/

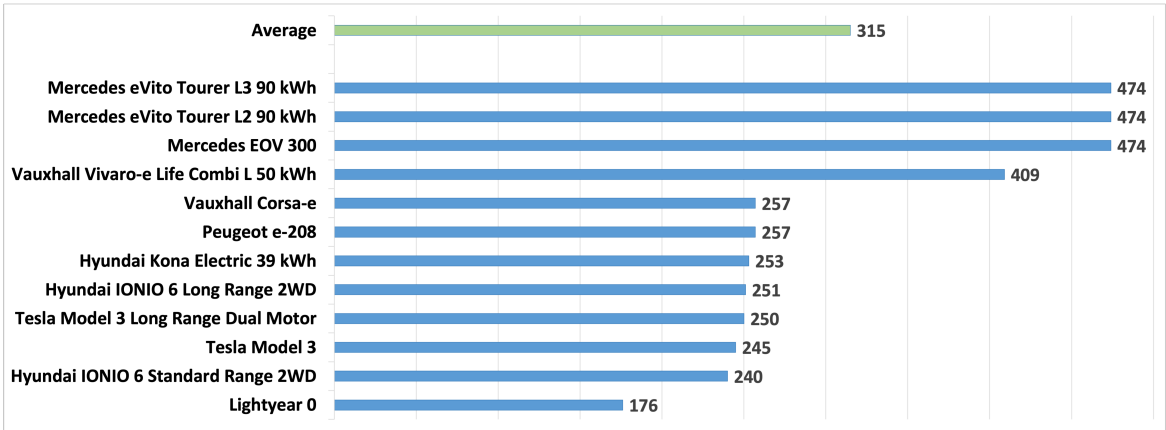

Fig. 6.7 Energy consumption of sample EVs available in the market (from the UK Electric Vehicle Database)

adapt the future EV charging infrastructure to the aggregate behaviour of users with respect to the statistics of the length and frequency of their travels.

EVs available in the market have different battery capacities with different efficiencies. Figure 6.7 shows the energy consumption of some of the EVs available in the market in terms of [Wh/mile], with the full list available in the UK Electric Vehicle Database.[3] The average consumption over 189 type of EVs is 315 wh/mile, obtained by taking the average over the EVs in the mentioned database. Considering a 30kWh battery capacity on the low end of the spectrum and the average consumption of 315 wh/mile, such a battery capacity can service the vast majority of trips in a day, assuming the vehicle completes less than 4 trips a day each less than 25 miles. These computations will be used in the sequel sections for estimating the required energy consumption and the charging points of EVs.

Locations of charging points. Charging points will serve users differently. In the following, four types of location for charging points are described.

- **Home Location Type.** This location type is associated with Charging points installed for off-street parking. The charging points with this location type typically have 3kW or 7kW rates. The dwell times in these charging points are relatively long (overnight with 10 hours or longer).

- **Destination Location Type.** charging points associated with attractions, such as supermarkets, shopping centres and work. Charger ratings will typically range between 7kW and 150kW, with dwell times expected to be between 1 hour and 8 hours.[4]

[3] https://ev-database.uk/cheatsheet/energy-consumption-electric-car
[4] https://www.rac.co.uk/drive/electric-cars/charging/electric-car-charging-speeds/

- **On-Street Residential Location Type.** Charging points associated with on-street deployment or similar, meeting the needs for those who cannot charge at home due to limited off-street parking access. These charging points are likely to be 7kW and 11kW rated, dependant on the site electricity circuit. Examples of these charging points could include cable gullies, retrofitting existing lighting columns and bollard/pillars. Each of which would be dedicated bays on-street or in car parks close to residential housing.

- **On-Route Location Type.** Chargers that are Rapid and above are the closest charging equivalent to the ICE vehicle petrol/diesel refuelling. These charging points are most appropriately placed at fuel courts and service stations in close proximity to the SRN/Highway. Dwelling times are expected to be less than 1 hour.

In this study, two broad categories are considered for the location types to make a better connection with the trips estimated from road use statistics:

- **Off-street Location Type.** charging that is related to Home Charging above.

- **Public Location Type.** charging that is an aggregation of destination, on-street residential charging, and on-route charging mentioned above.

The chargers of electric vehicles are categorised by the charger ratings (KW). The current popular charger ratings are summarised below (ZapMap, 2022a). Note that the charging point technology might change in the future, therefore more charger ratings can be introduced to the EV charging infrastructure, and the below definitions and names might change.

- **Slow charging points**: These types are 3 kw chargers. The length of charging with these charging types usually takes around 6-12 hours. This type of charger is mostly used for household purposes.

- **Fast charging points**: These types are 7 kw, 11 kw and 22 kw. It can take around 3-4 hours to fully charge some models.

- **Rapid charging points**: These types are 50 kw. These chargers usually can charge to 80% in 30-50 minutes.

- **Untra-Rapid charging points**: are 100 kw and up to 150 kw. These chargers are added in order to charge quickly as possible within 20-30 minutes.

These types of charging points can be further classified based on the underlying charging power method that could be either *Alternating Current* (AC) or *Direct Current* (DC). The

Table 6.4 Priorities of charging types computed by combining multiple factors

Charger Priority							
Chargers Types	AC				DC		
	Slow	Fast			Rapid	Ultra-Rapid	
	3	7	11	22	50	100	150
Priority weights (%)	68	72	80	76	56	32	32

power that comes from the power grid is always AC. Electronic devices can have a converter built into the plug to store the power in the battery using DC.

In order to get a measure of priority for these types of charging points, six key factors for each of the charger ratings (KW) are considered. These factors include infrastructure and installation costs, utilisation, impact on the grid, future proofing, and tariff (Nicholas, 2019). Table 6.4 shows the average prioritisation computed as a weighted sum of these factors. A priority weight of 100% means that the charger is the most suitable option, whereas a priority weight of 0% indicates the charger is not as suitable.

The above analysis based on relevant datasets gives an understanding of the current demand for road use by EV's and the current EV charging infrastructure. This analysis will be used on this chapter to predict the requirements for expanding the EV charging infrastructure based on the future likely demand and building a simulation model that generates the future demand to be used as part of applying the framework in Figure 3.5.

6.5 Electric Vehicle Charging Simulation

To plan and develop the transport infrastructure including EV charging points, it is essential to develop scenarios that clarify how users might engage and interact with the transport infrastructure and how the current set of policies affects this interaction at large scale. The future travel scenarios are constructed based on a range of factors that affect the future of the transport as elaborated in the rest of this section. It is important to emphasise that such scenarios are not predictions as there are large uncertainties around the user behaviours and technological developments, but they provide plausible futures. These scenarios are not intended to be good or bad statements, but are aimed at formulating the most plausible combinations of uncertain factors, and find possible actions that need to be taken or adapted to these scenarios as more certainties are revealed over time.

Despite the recent advances in the EV technology and the EV charging facilities (Aruna and Vasan Prabhu, 2021; Hutchinson et al., 2019), the number of EVs and the charging points are relatively low at the moment. As of March 2023, there are only 735,000 EVs on UK

roads, which is less than 2% of the total licensed vehicles.[5] This leads to large uncertainties in the future user charging behaviours. The requirements on the EV charging infrastructure depend on charging behaviours, the availability of the off-street parking, and mobility trends. This section describes scenarios considered by National Grid in making predictions on the number of EVs in the future. It then discusses the scenarios considered by Arup that is based on the predictions of the number of EVs provided by National Grid.

National Grid has considered four future energy scenarios to forecast the EV numbers. These four scenarios are based on the set of policies, the speed of decarbonisation and the level of societal changes. The four scenarios are as follows.

1. **Steady Progression:** a pessimistic scenario with a slow speed of decarbonisation of the energy vectors and low level of societal change, slow adoption of EVs and slow installation of charging points. The ban on the sale of new petrol and diesel vehicles is achieved in 2035 by cars and in 2040 by vans.

2. **System Transformation:** a scenario with a moderate speed of decarbonisation and middle level of societal change. Charging points for EVs are installed ahead of the need. The ban on the sale of new petrol and diesel vehicles is achieved in 2032.

3. **Consumer Transformation:** a scenario with a moderate speed of decarbonisation and higher level of societal change. Drivers adopt EVs ahead of charging provisions. The ban on the sale of new petrol and diesel vehicles is achieved in 2030.

4. **Leading The Way:** an optimistic scenario with a fast speed of decarbonisation and highest level of societal change. The ban on the sale of new petrol and diesel vehicles is achieved in 2030.

Arup has developed four scenarios to forecast the number of EVs. These scenarios are adapted from the National Grid ones, and are as follows. **(1) Baseline:** A baseline set of assumptions that relies on the behaviour of consumers to date to forecast EV energy demand and charging point quantities. **(2) Consumer Efficiency:** A scenario that assumes EV purchases are done mainly for every day short distance use. EV owners use more lower charger speeds and operate between 20%-80% to optimise their battery life. **(3) Government On-Street:** Public residential chargers are made available through appropriate government schemes. **(4) Rapid Dominant:** Rapid and above charging points are made available to reduce consumer dwell times, which are 50kW charging points and above.

Figure 6.8 shows the estimation of the number of EVs for Newcastle upon Tyne, which is obtained using the future energy scenarios of the National Grid proportioned to the number of

[5] https://www.zap-map.com/ev-stats/ev-market/

cars Newcastle. These estimations are also used in the scenarios developed by Arup. Three of the four curves in Figure 6.8 show a reduction in the number of EVs on the road between 2042 and 2048. Steady Progression scenario shows a peak in 2050. The fall in EV numbers in the three scenarios meeting net zero by 2050 is due to the National Grids assumptions surrounding other forms of propulsion, automated self-driving vehicles and public transport. According to the Figure 6.8 the peak year for Consumer Transformation is in year 2046, for Leading the Way is in year 2042, for Steady Progression is in year 2050 and for System Transformation is in year 2048. The maximum estimated number of EVs and the year in which the maximum occures are reported in Table 6.5.

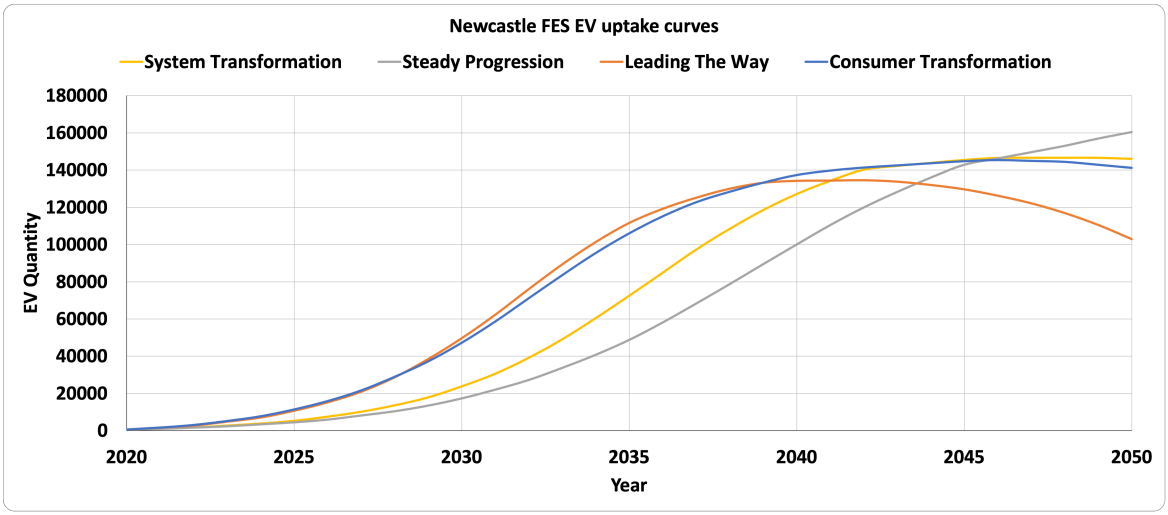

Fig. 6.8 Estimation of the number of EVs for Newcastle upon Tyne using the future energy scenarios of national grid

Table 6.5 The maximum number of estimated EVs and the year in which the maximum occurs based on future energy scenarios of National Grid for Newcastle upon Tyne

Future Energy Scenarios	Maximum Number of Estimated EVs	Peak Year
Steady Progression	160,403	2050
System Transformation	146,617	2048
Consumer Transformation	145,345	2046
Leading The Way	134,606	2042

6.5.1 Estimating the Number of EVs in Each LSOA

Next it is discussed how to estimate the number of EVs in each LSOA. Define the EV uptake percentage as the total number of forcasted EVs divided by the total registered vehicles from

the National Trip End Mode (NTEM). The NTEM also provides the *total registered vehicles* for each MSOA. This total registered vehicles for MSOA is split into the number of vehicles for LSOAs using a proportional split determined by the 2011 Census. The forecasted number of EVs in each LSOA is the number of registered vehicles in the LSOA multiplied by the EV uptake curve percentage:

$$\mathrm{EV}_{\mathrm{LSOA}} = \mathrm{EV}_{\mathrm{uptake}} \times \mathrm{Veh}_{\mathrm{MSOA,NTEM}} \times \frac{\mathrm{Veh}_{\mathrm{LSOA,cen}}}{\sum_{\mathrm{LSOA}} \mathrm{Veh}_{\mathrm{LSOA,cen}}} \tag{6.1}$$

Note that this proportional split can be further refined by considering demographics and household income that may mean ownership in some LSOA will grow in different ways to others. The optimisation methodology of this chapter is general and can be adapted to take into account such factors. Section 6.6 will discuss the results of the optimisation for Base Scenario from Arup in a combination of the estimated number of EVs for Newcastle upon Tyne from Future Energy Scenarios in peak years for *public* locations.

6.5.2 Construction of the Simulation Model

In order to simulate the EV charging infrastructure of Newcastle upon Tyne, the datasets of Section 6.4.1 and the assumptions and calculations of this section have been used to estimate the total EV travel mileage, the energy demand, and the power demand. A summary of the calculations for estimating the average daily EV mileage is as follows. The national trip end model has been used for the Origin/Destination data. This dataset is open source and is provided by the Department for Transport. The dataset defines two types of origin: Home based and Non-home based. Each LSOA is assigned an urban-rural classification as defined by the Census data, for which the average annual mileage and the average number of car/van driver trips are used to define an average trip mileage for each of the four urban-rural classes. For all home-based trips, the average trip mileage for each rural and urban classification is applied. The non-home based trips is split into non-home based trips that start within the Local Authority area and non-home based trips that start outside of the given Local Authority area, using a census Local Authority origin destination dataset. For non-home based trips that start within the given Local Authority, the average trip mileage determined in the previous stage is applied. For non-home based trips that start in another Local Authority area, the travelled distance is calculated using spatial mapping. The EV travel mileage is computed by considering the total travel mileage, the trend in the growth of registered EVs and the total number of registered vehicles, as well as to the EV battery range. The general overview of the simulation is shown in Figure 6.9. The simulation has a Destination Model

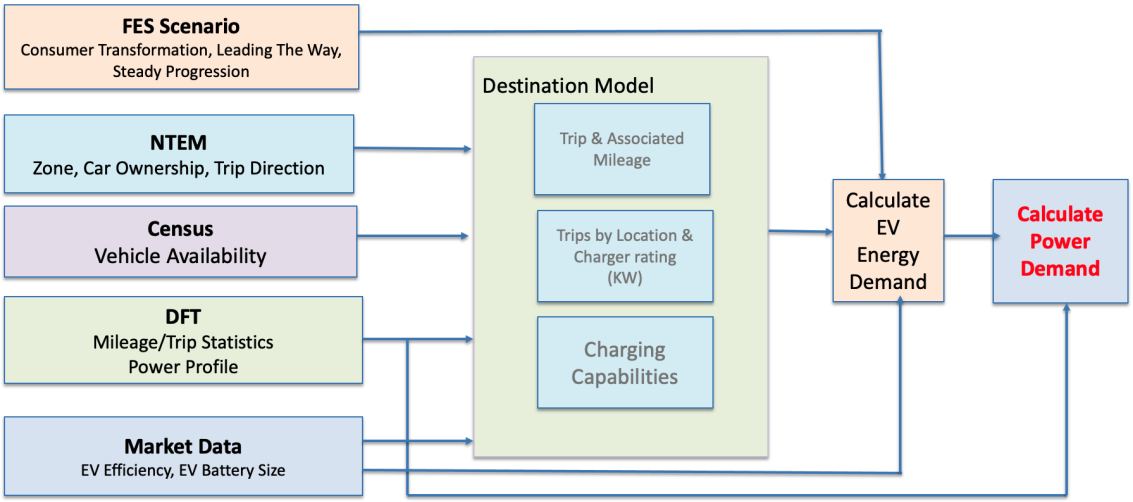

Fig. 6.9 An overview of the Destination Model: connecting data-driven inputs to estimate EV energy demand and power demand

that seamlessly integrates multiple data sources to estimate EV energy demand and power demand at a granular geographical level.

6.6 Optimisation Approach

This section investigates multi-objective optimisation methods in order to inform on the future EV charging infrastructure and find the best quantity of charging points, their types and locations whilst minimising the total capital and operational expenditures. Based on the existing optimisation methods, a new optimisation solution is proposed in this section to address the limitations of the standard single-objective optimisation methods. More specifically, the optimisation solution of this section is different from previous ones in the following directions: (a) it is guided by Long Short-Term Memory (LSTM) models from the Machine Learning literature; (b) it does not compress the multiple objectives into a single objective, but combines and ranks multiple objectives through fuzzy logic; (c) it extends the capabilities of the traditional genetic algorithms by integrating available data with a simulation model.

6.6.1 Optimisation Steps for EV Charging

The optimisation model of this chapter is designed to adopt the most central idea of genetic algorithms, which is to simulate nature for selection, evolution and reproduction. By constructing a genotype complete with the main parameters possessed by a charging point, traits

and expressions are designed to describe the characteristics of a charging point and how it will behave in a given environment. The main idea of the genetic algorithm for solving the problem of finding the best location of charging points is as follows. The EV power demand generated from the EV users will be responded to by the EV chargers depending on the location of EVs and the charging points. The number of EVs that need to be charged will be different over time (different generations or iterations). The algorithm deploys different types of charging points at various locations to select the best locations and types.

Next it is discussed how to apply the optimisation for the EV charging infrastructure planning. The following assumptions are made based on the processed data: (1) EVs will be charged according to EV charging demand, which means that users will charge according to their needs and will not necessarily wait for the EV battery to run out of charge completely; (2) Users prefer more powerful charging points, which will allow them to have shorter waiting times for charging, but users will consider a combination of distance and time; (3) Projected based on data, the initial number of EVs is selected according to Table 6.5. (4) EVs requiring charging are randomly generated within the LSOAs; (5) After comprehensive consideration, six types of charging points are selected for optimisation: 7kW, 11kW, 22kW, 50kW, 100kW and 150kW.

The steps for constructing the new optimisation model are as follows.

1. Analysis and processing of data to determine the subject and environment of optimisation.

2. Construct gene vectors, determine genotypes and score types, and construct the initial solution vector space, i.e. construct the initial population.

3. Through the processing of real data, a resource vector is designed through mapping and an environment matrix is constructed.

4. Study the specific expression of the design score type through the influencing factors of the real problem, and complete the construction of multiple objective functions.

5. Use the elite strategy of fuzzy logic to rate the score types.

6. Memorise and eliminate the initial solution vector matrix based on the ratings by constructing a transformation matrix. This is similar to the memory gate and forgetting gate of LSTM models.

7. Generate the next generation of children based on the rating content and child generation matrix, and add them to the solution vector matrix space.

8. The environment matrix is regenerated according to the mapping rules.
9. Rating by calculating the score type and then calculating the crowding by the inner product for the exemplar elite, so that the crowding is in the right zone to ensure species diversity and also the right direction for optimisation and search.
10. Repeat steps 5 to 9 until the generational requirements are satisfied.

6.6.2 Constructing the Solution Vector Spaces

Based on the optimisation method proposed in the previous section, gene vectors are constructed for each EV charging point as

$$\vec{X}_j^{[i]} = \left(a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3\right), \tag{6.2}$$

where a_1 is the gene ID, a_2 represents the latitude, a_3 represents the longitude, a_4 is the charging type, a_5 is the charging point total capital and operational expenditures, b_1 represents the number of charges score, b_2 represents the charging point operating hours score, b_3 represents the generation of birth in the optimisation algorithm. The relationship between these values will be discussed after introducing the environment matrix.

The total investment is computed according to the following equation: (Jia et al., 2012; Zhou et al., 2022)

$$a_5 = C_j \frac{r_0(1+r_0)^{n_{year}}}{(1+r_0)^{n_{year}} - 1} + M_j, \tag{6.3}$$

where r_0 is the inflation rate which is set to 10%, C_j is the construction cost of each charging point that includes the acquisition cost, the land price, and the cost of replacement of the charging point at the end of its life span. M_j is the maintenance costs and n_{year} is the planned life span set to 15 years (Leone et al., 2022).

After designing the genetics of the charging point, the 175 LSOAs of Newcastle upon Tyne are considered with random selection of charging points of any type in the centre of each LSOA. These gene vectors are then added to the space of solutions.

6.6.3 Generating the EV Power Demand

A Heat Map of the EV charging power demand is shown in Figure 6.10 for the peak year (2042) of the Leading The Way scenario. The initialisation of EVs in one LOSA is presented in Figure 6.11. The behaviour of the EVs is simulated using the EV peak power demand as described in the previous section. The required number of EVs in each LSOA is generated

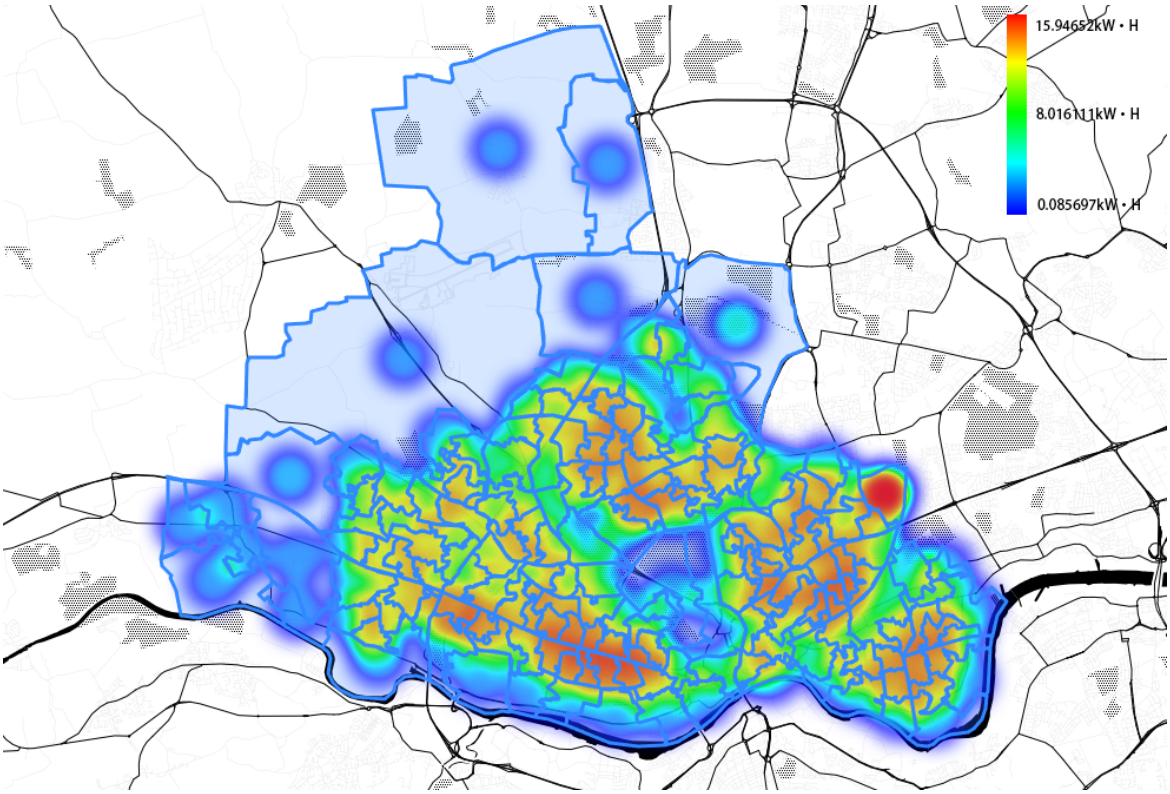

Fig. 6.10 Heat map of EV charging power demand in the peak year (2042)

Fig. 6.11 Initialisation and EV charging power demand

randomly within a circle around the centre points of the LOSA. Normal distribution is used to generate the location of EVs according to the following formulas

$$\begin{aligned}
&\mu_1, \mu_2 \sim \mathscr{N}(0,1), \quad A \sim \mathscr{U}(-\pi, \pi) \\
&\text{EV Latitude} = CL_a + \mu_1 R\cos(A) \\
&\text{EV Longitude} = CL_o + \mu_2 R\sin(A) \\
&R = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(Ma_i - CL_a)^2 + (Mo_i - CL_o)^2]}.
\end{aligned} \tag{6.4}$$

The above formulas mean that the scaling factors μ_1, μ_2 are selected randomly according to the normal distribution with zero mean and standard deviation equal to one $\mathscr{N}(0,1)$. The angle A is selected according to uniform distribution from the range $[-\pi, \pi]$ to cover the whole circle. The latitude and longitude of the EV are then computed from the centre point of the LSOA (CL_a, CL_o). These formulas ensure a normal distribution of EV numbers from the centre to the edges of LSOAs.

The radius R is calculated as the average distance of the (Ma_i, Mo_i) to the centre points. Here, (Ma_i, Mo_i) are the latitude and longitude of the i^{th} marginal point. Since regional centres are mostly commercial centres or transport hubs, this means that EVs are likely to be there more often using normal distribution to simulate this behaviour. EVs may engage in cross-region charging behaviour for a number of reasons, e.g., when the charging points in other region are full.

The following equations are used to generate the total power demand, which is then used for creating the charging requirements of EVs.

$$\begin{aligned}
&\eta \sim N(0,1), \\
&P = (1 \pm \eta) \times \text{EV Peak Power Demand} \\
&0 \leq P \leq 2 \times \text{Max Power Demand}.
\end{aligned} \tag{6.5}$$

In the above equations, η is the deflation factor. Adopting a normal distribution means that EV users rarely charge when their battery is empty or full (due to anxiety factor), but rather when it is below a threshold. A limit is placed on the total EV power demand P, which is two times the maximum power demand.

Now the resource vector $\vec{Y}_j^{[i]}$ and environment matrix $\mathbf{E}^{[i]}$ are constructed using the location, demand and quantity as described above. c_{f1} represents the latitude, c_{f2} represents

the longitude, and c_{f3} represents the power demand of EVs:

$$\vec{Y}_j^{[i]} = \left(c_{j1}, c_{j2}, c_{j3}\right), \quad \mathbf{E}^{[i]} = \begin{pmatrix} \vec{Y}_1^{[i]} \\ \vec{Y}_2^{[i]} \\ \vdots \\ \vec{Y}_n^{[i]} \end{pmatrix} = \begin{pmatrix} c_{11}, c_{12}, c_{13} \\ c_{21}, c_{22}, c_{23} \\ \vdots \\ c_{n1}, c_{n2}, c_{n3} \end{pmatrix}. \tag{6.6}$$

The shortest route of the EV to be charged, the total expenditures and the electrical load are considered in the optimisation. The goal of the optimisation is that the EV charging points be used for as many EVs as possible, the EV charging points work for as long as possible, and the least amount of total expenditures is invested to place the right type of charging points in the right locations. Therefore, these factors need to be optimised at the same time as a multi-objective optimisation problem. This is shown in Figure 6.12.

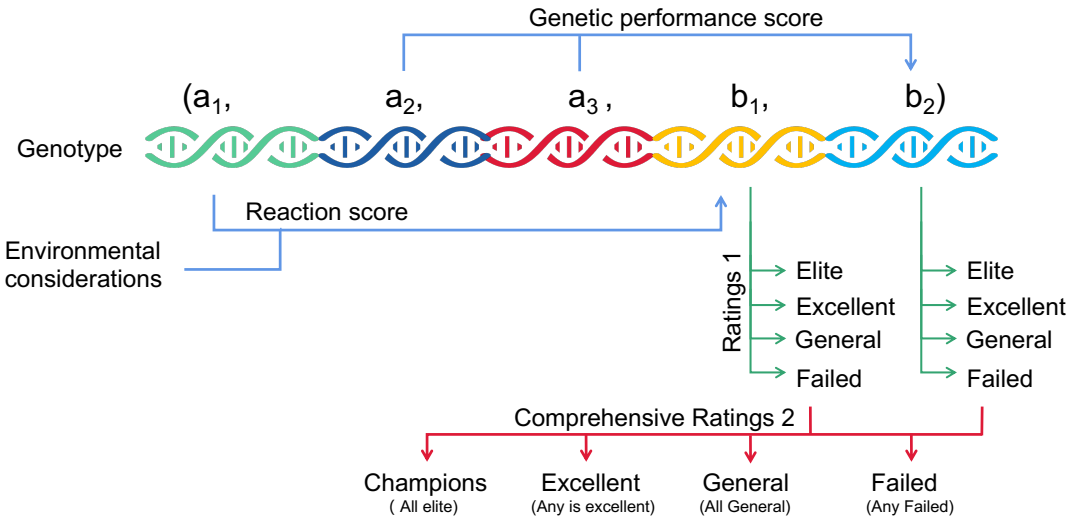

Fig. 6.12 Example of a single gene in the genetic optimisation

6.6.4 Selecting the Elite

An elite strategy is set up for each score through fuzzy logic as in Table 6.6. The multiple objective functions are integrated through such a combination of elite strategies to obtain a solution set for the optimisation. The percentages in the Elite Strategy Rating depend on the specifics of the problem under study. The percentages are set heuristically and then are tuned based on the behaviour of the gene vectors in the solution set of the optimisation. In the case study, the bottom 20% are considered as unhealthy individuals that will be eliminated during

the iterative process. The top 10% of individuals will generate 1 or 2 offspring individuals at random, and the 10% to 50% of individuals will only generate 1 offspring individuals, and the 50% to 80% of individuals will generate 0 or 1 offspring individuals.

Table 6.6 Elite Strategy Rating

Rank	Elite	Excellent	General	Failed
Rating interval	[0%,10%]	(10%,50%]	(50%,80%]	(80%,100%]
Number of offspring	1 or 2	1	0 or 1	0

6.6.5 Utilisation Rate

The utilisation rate is expressed in the following equation:

$$\text{Utilisation rate} = W_1 \left(\frac{\text{Number of Feeded EVs}}{\text{Total Number of EVs}} \right) + W_2 \left(1 - \frac{1}{n} \sum_{i=1}^{n} \frac{(b_2^i - T_{re})^2}{T_{re}^2} \right). \quad (6.7)$$

The rate is formed as the weighted sum of two terms. The first term indicates the portion of EVs with their charging requested being satisfied. The working time of the charging points are expected to be around T_{re}. The quantity T_{re} represents the recommended working time for the charging points, which is $[10, 15]$ hours (de Mattos Affonso and Kezunovic, 2018). The second term shows the deviation of the charging point working hours from T_{re}. This term decreases by larger deviations. Here, $T_{re} = 10$ hours is chosen with the weights $W_1 = 0.6, W_2 = 0.4$.

6.7 Results of Applying the Framework

The optimisation approach is applied for obtaining the charging locations in Newcastle upon Tyne considering the peak years of the four different scenarios reported in Table 6.5. Figure 6.13 shows the cluster map of the results for *Leading The Way scenario*. The numbers written on the circles represent the number of EV charging points within the related areas. The clustering is done by setting the map size to be 15,000 smaller than the actual size. The colours of the circles represent the proportion of the total number of charging points in the area to the total number of charging points (red colour for higher portions). Figure 6.14 shows the spatial distribution of the genes in the optimisation representing EV charging points. Colour of the dots in the figure shows the type of the charging points. The optimisation results indicate that by taking 4753 charging points and arranging them in the areas shown in

Figure 6.14, a more appropriate efficiency of the charging points can be gained with lower total capital and operational expenditures. The two Figures 6.13–6.14 indicate that in rural areas where demand is low, fewer charging points are placed and close to the main roads. In urban areas, where there is a high demand, the number of charging points is high and they are located close to residential areas, shopping malls and major roads.

The relatively large number of charging points is computed by the optimisation under the current and predicted road use behaviour. In case policies are deployed that encourage a major shift to public transport and deter substantially the wide spread of private car usage, it is necessary to revise the scenarios, assumptions of the baseline model, and the predicted energy demand. This in turn affects the required number of charging points.

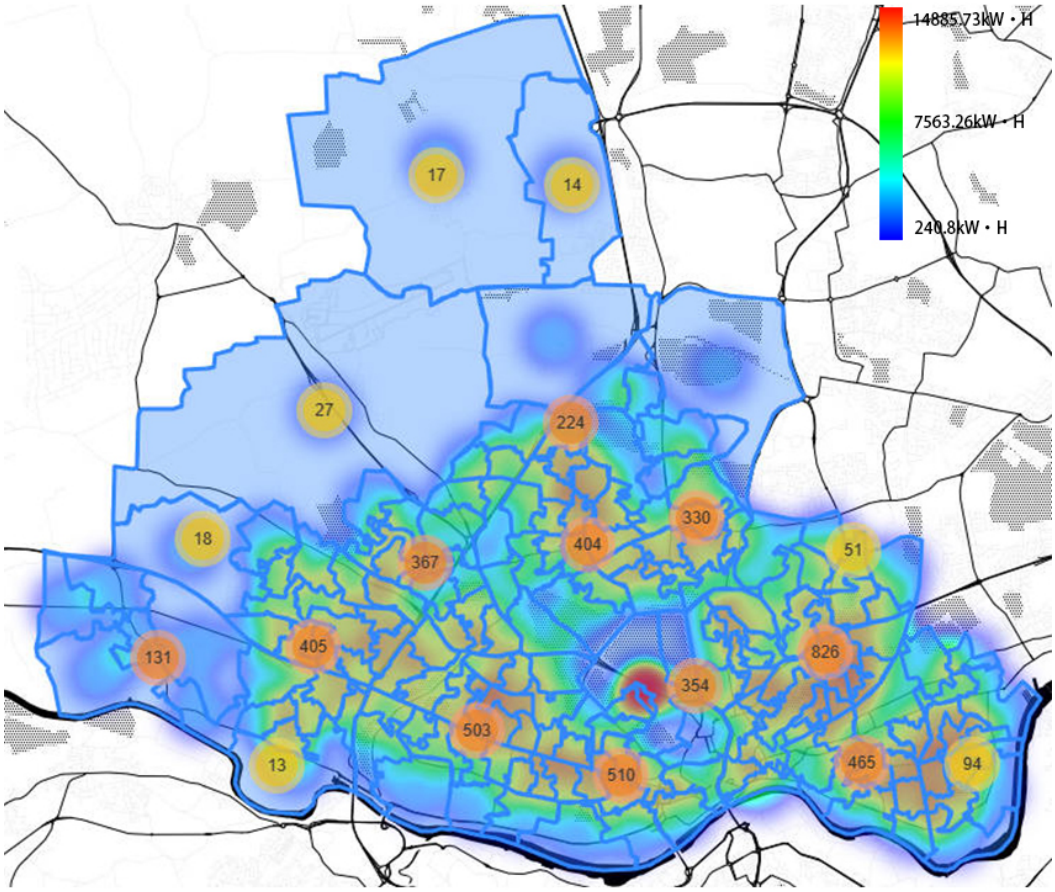

Fig. 6.13 The cluster map of the results for Newcastle upon Tyne

Table 6.7 gives the optimisation results for the peak years of the four scenarios. The total number of charging points, average operating hours, and the total costs are reported. As can

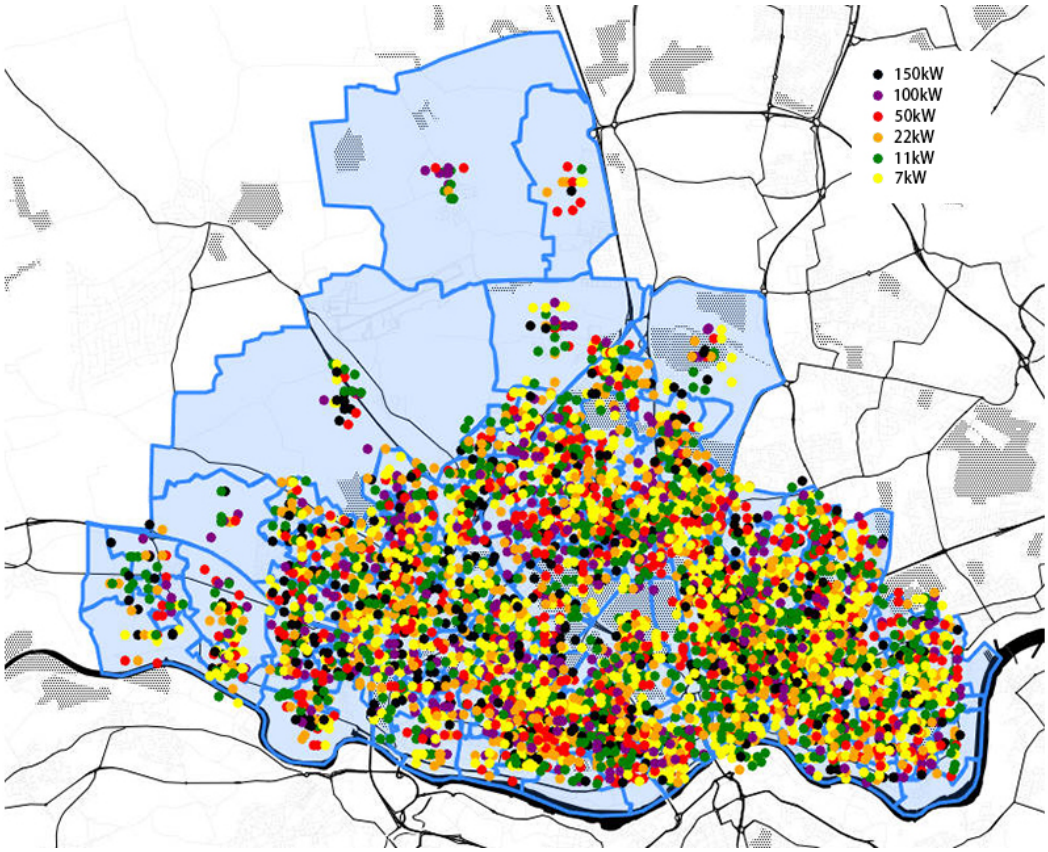

Fig. 6.14 Spatial distribution of the genes in the optimisation representing EV charging points

Table 6.7 Optimisation results for peak years in different scenarios

Scenario	Leading The Way	Consumer Transformation	System Transformation	Steady Progression
Peak year	2042	2046	2048	2050
Quantity of EVs	134,606	145,345	146,617	160,403
Total number of charging points	4753	5167	5386	5817
Total Cost (£)	8,195,000	8,859,400	9,117,900	9,880,200
Average operating hours of charging points (h)	7.57	7.77	7.81	7.95

be seen from the table, the required total number of charging points will increase with their total expenditures increasing as well. In contrast, the average operating hours of charging points is in the range [7.5,8], and goes up from 7.57 to 7.98 hours.

Figure 6.15 shows the portion of 6 types of charging points for the four scenarios. The variations of the portions of different types are relatively small (within 3% range of the total number of charging points). It is found that regardless of the chosen initialisation of the optimisation process, the optimal solution puts priority on the slower charging points (respectively 7kW, 11kW, and 22kW). The faster charging points (150kW, 100kW, and 50kW) have smaller portions each around 10%-13%. This means that while 7kW charging dominates the market currently, it is more beneficial to improve charging efficiency and reduce the total investment costs by installing more from faster types of charging points.

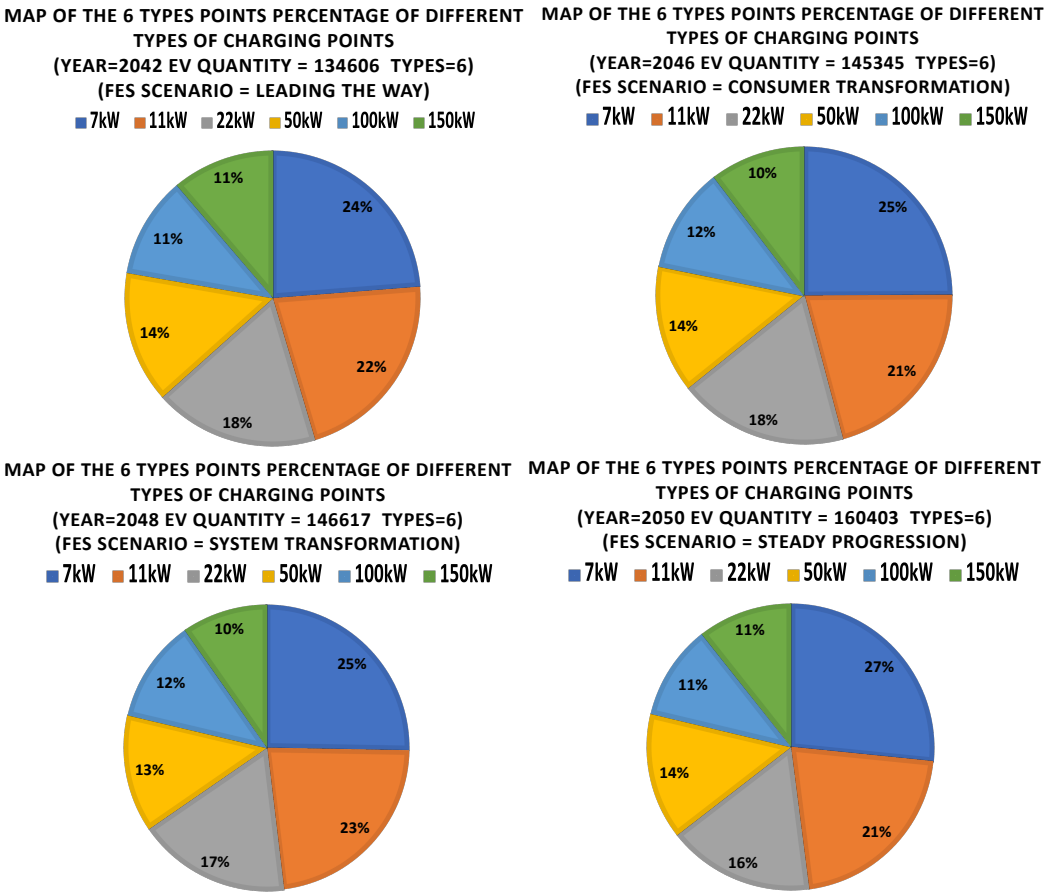

Fig. 6.15 Portion of 6 types of EV charging points for the four scenarios

6.7.1 Visualisation of the Iterative Optimisation

The optimisation process is visualised in Figure 6.16. The algorithm initially places a number of charging points with different types in the centre of the 175 areas. In the next step, the algorithm scores the available solutions by fuzzy logic and filters the better ones for the

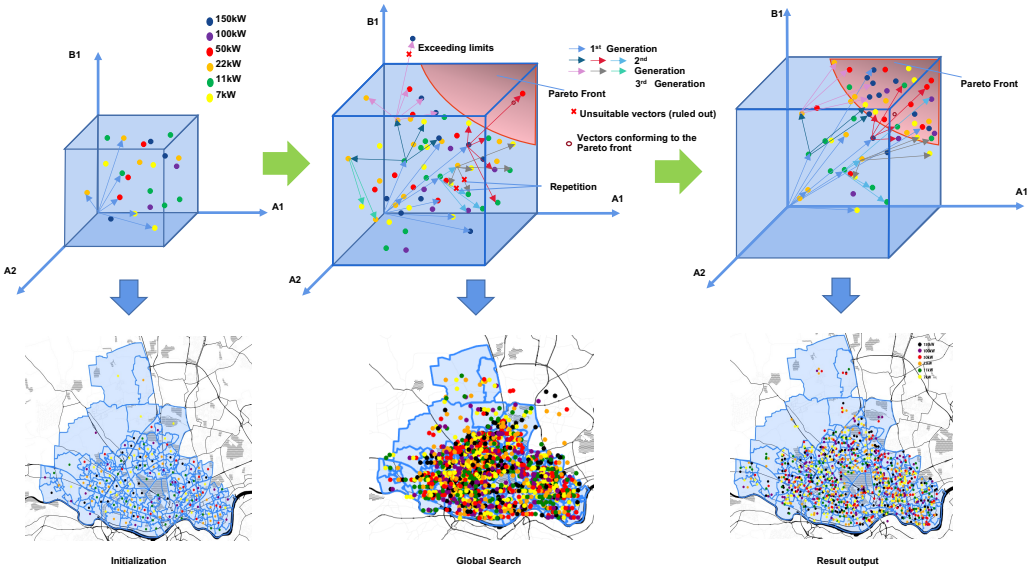

Fig. 6.16 Optimisation process of the multi-objective genetic algorithm

mutation-based generation of offspring. The generation of children in the next iteration of the algorithm expands the search direction and the search space. The first phase of the total iterations rapidly approximates the optimal envelope in the solution space. The second phase then combines different objectives through fuzzy logic to converge to an optimal solution. These two phases are also shown in Figure 6.17 that provides the total number of EV charging points in each iteration of the optimisation. The first phase is indicated by the red colour and the second phase by blue colour. The results are for four different scenarios and 100 iterations. The total number of charging points starts from a very small value (initialisation of the algorithm) and gradually increases to a peak value (first phase), then slightly decreases until converging to an optimal solution (second phase).

Figure 6.18 shows the number of 6 types of EV charging points in each iteration of the optimisation. Results are for four different scenarios and 100 iterations. Similar to the total number of charging points, the number of each type starts from a small value and gradually increases to a peak value, then slightly decreases until converging to an optimal solution. A similar trend is observed in the total expenditure presented in Figure 6.19.

6.7.2 Spatial Distribution of the Future Charging Infrastructure

Figure 6.20 shows the current and computed spatial distribution of the charging points. The left figure shows the existing charging points (source: ZAP-MAP) and the right figure shows the solution of the optimisation for future installations. The red ellipsoids represent similarities between the current charging points and the computed solution; The purple ellipsoids

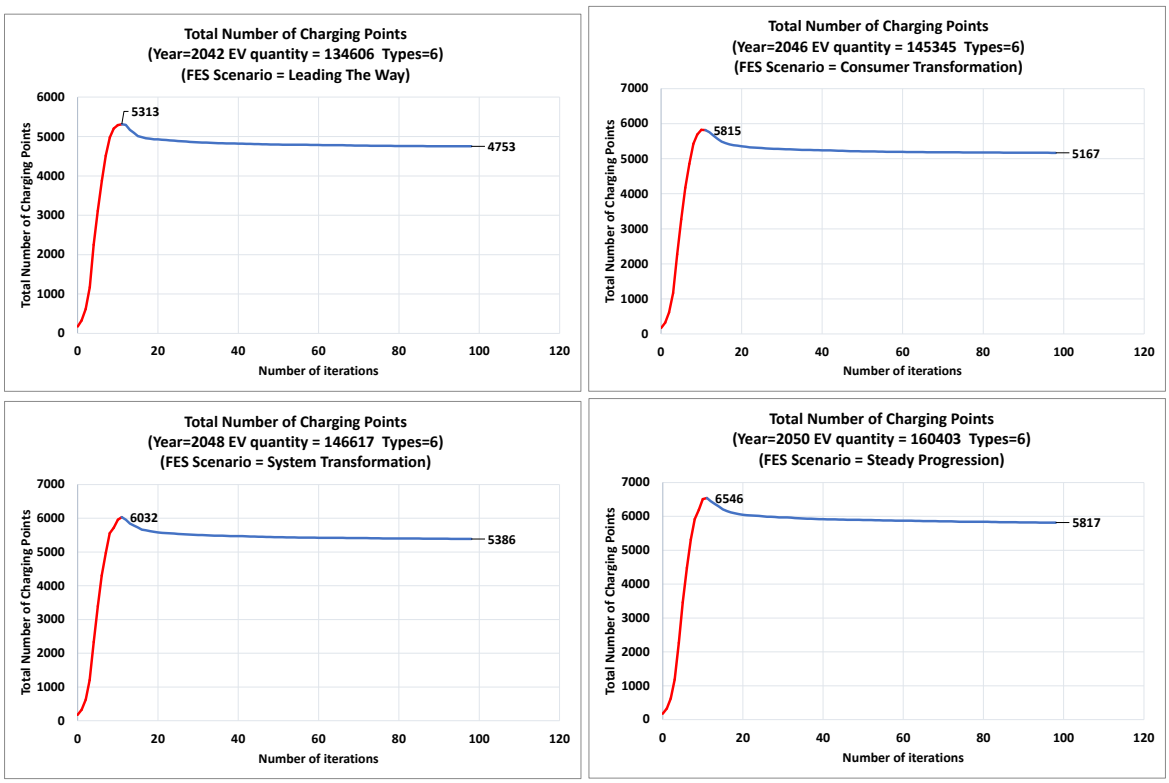

Fig. 6.17 Total number of EV charging points in each iteration of the optimisation

indicates areas with no installation but future installations is needed. This corresponds to residential areas that do not have a relevant arrangement of charging points at the moment.

6.8 Sensitivity Analysis of the Optimisation Solution

Sensitivity analysis is an important validation technique to test the robustness of an optimisation framework. In this study, sensitivity analysis has been conducted by altering input parameters, such as the number of electric vehicles and the types of charging infrastructure. By analysing the impact of these changes on the outcomes, it is demonstrated that the framework can adapt to different scenarios and produce meaningful results. This also helps to identify potential limitations and areas for improvement in the optimisation method.

6.8.1 Sensitivity to the Number of EVs

The sensitivity of the solution to the estimated number of EVs is studied by increasing and decreasing the EV numbers with 10%. The baseline for comparison is *Leading The Way* scenario for the year 2042, EV quantity 134606, and 6 types of EV charging points (cf.

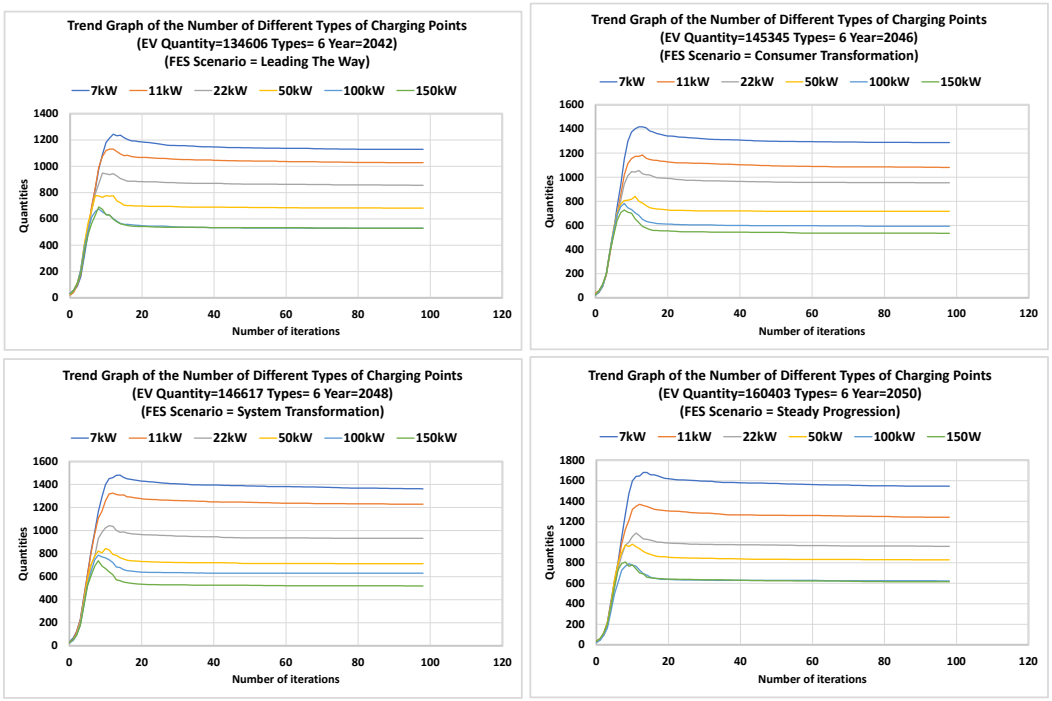

Fig. 6.18 Number of 6 types of EV charging points in each iteration of the optimisation

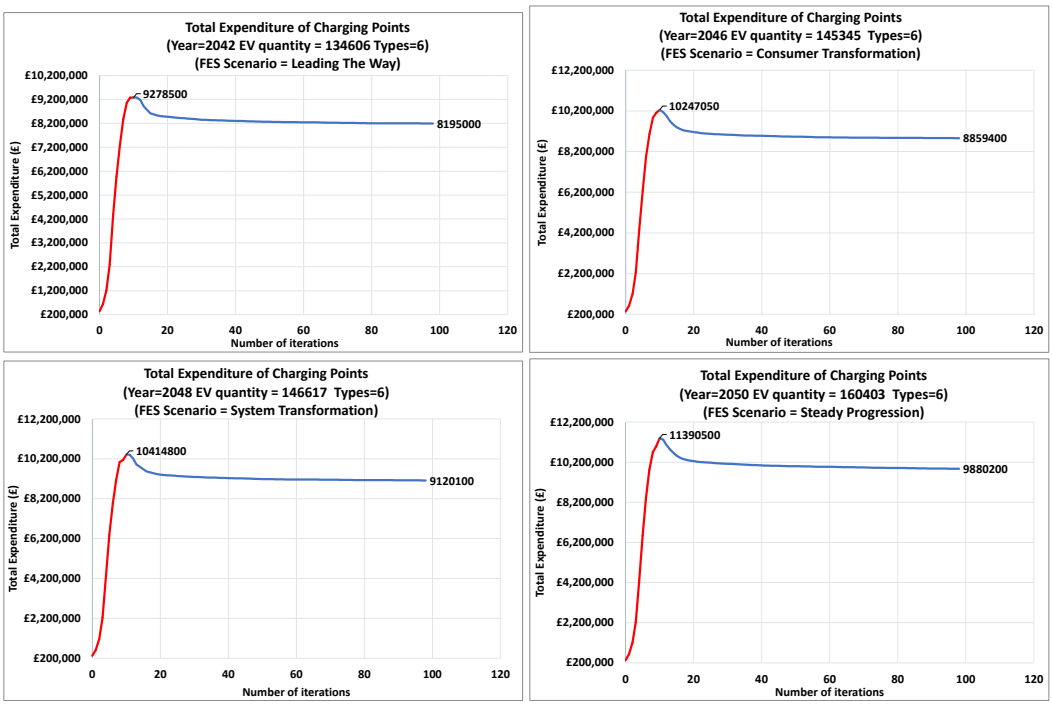

Fig. 6.19 Total Expenditure of the EV charging points

Table 6.5). The results are reported in Table 6.8. The percentages are obtained by taking average over 6 runs of the optimisation.

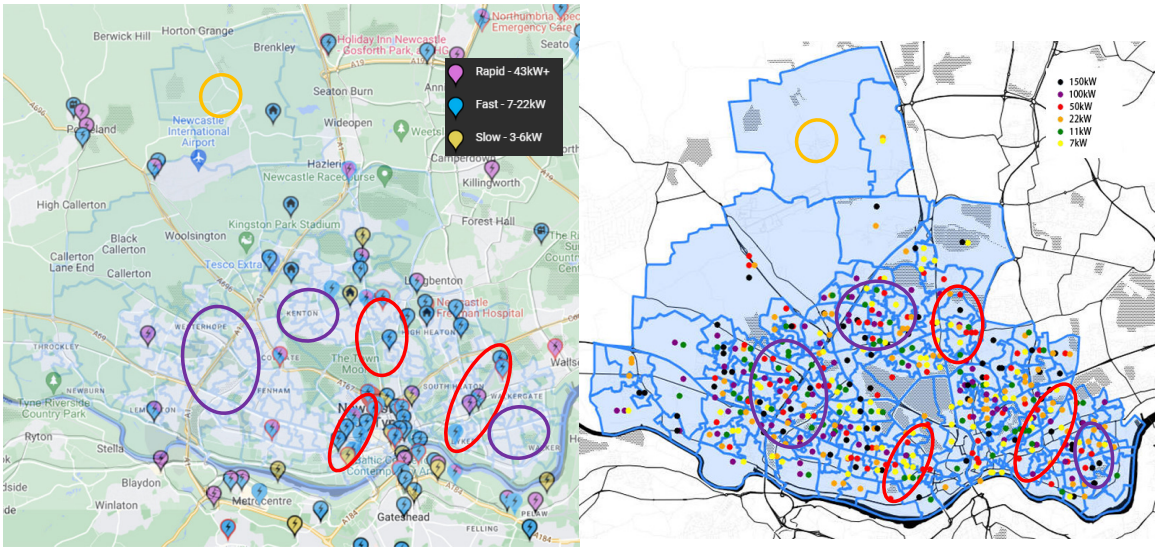

Fig. 6.20 Spatial distribution of the charging points obtained from the optimisation

A 10% increase in the number of EVs to be charged increases the total number of charging points by 27.5%, increases the total expenditure by 30.4%, decreases the number of EVs to be charged by a single charging point by 10.5%, and reduces the average operating hours of a single charging point by 17.5%.

In contrast, a 10% reduction in the number of initial EVs would reduce the total number of charging points by 5.3%, reduce the total expenditure by 7.8%, increases the number of EVs to be charged by a single charging point by 14.1%, and increases the average operating hours of a single charging point by 6.5%.

Table 6.8 Sensitivity to the number of EVs

Changes in the EV numbers	+10%	-10%
Total number of charging points	+27.5%	-5.3%
Total expenditure (£)	+30.4%	-7.8%
Average number of EVs charged by a single charging point	-10.5%	+14.1%
Average operating hours of charging points (h)	-17.5%	+6.5%

6.8.2 Sensitivity to the Number of Charging Types

Given the rapid developments in the charging point technologies, two different cases are considered to study the sensitivity of the optimisation with respect to the diversity in the types of charging points. In the first case, it is assumed that only 5 types of charging points

are available in the network by eliminating 150kW charging points. In the second case, it is assumed that an additional type of 350kW charging point is also available (in total 7 types).

The results are summarised in Table 6.9 by considering the base of comparison to be *Leading The Way* scenario for the year 2042, EV quantity 134606, and 6 types of EV charging points (cf. Table 6.5). The percentages are obtained by taking average over 6 runs of the optimisation. The table shows that the number of charging points should be increased by 10.5% when charging points with higher powers are not available. However, smaller number of charging points are needed with high powers charging points available in the network. A small variation is also observed in the total expenditure, which is 3.36% more when larger number of charging points with smaller powers need to be installed. The changes in the number of charging points in turn influences the average number of EVs charged by a single charging point: smaller (larger) number of charging points will serve higher (lower) number of EVs in average when the demand is staying roughly the same (note that the network has a fixed number of EVs). Finally, the average operating hours of charging points has stayed almost the same. This means that the EV charging infrastructure should consider installing high power charging points when they become available and increase the diversity of the charging types.

Figure 6.21 shows the portions of different types of charging points when the network has only 5 types, and compares it with the 6 types. As can be seen in the left figure, 36% of charging points are 7kW, but this is replaced in the right figure by 22% 7kW and 14% 150kW. Therefore, the optimisation algorithm suggests that a portion of increase in the number of 7kW charging points should be covered by installing high power 150kW charging points.

Table 6.9 Sensitivity to the number of charging types

Number of charging point types	7 types	5 types
Total number of charging points	-7.87%	+10.5%
Total expenditure (£)	-0.85%	+3.36%
Average number of EVs charged by a single charging point	+7.39%	-9.43%
Average operating hours of charging points (h)	-1.03%	+0.24%

6.9 Testing the Load Carrying Capacity

Load carrying capacity (LLC) is a concept used in various engineering disciplines to assess the reliability of system under extreme loads before any failure happening in the system (Abdullah et al., 2014; Randolph et al., 2004). The goal of this section is to apply the LLC concept to the EV charging infrastructure and study its reliability under extreme conditions.

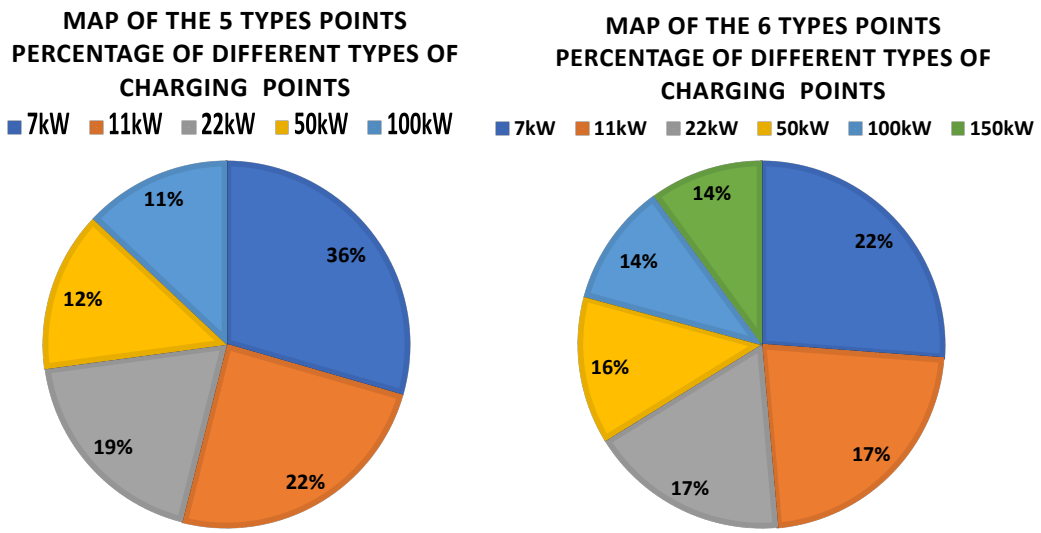

Fig. 6.21 Portions of different types of charging points

For this purpose, the optimisation of this chapter is applied to design the EV charging infrastructure for the *Leading The Way* scenario for the year 2042 with EV quantity 134,606. The number of EVs is then increased to 300,000 (more than twice) and the response of the charging infrastructure to this extremely large demand is studied.

Table 6.10 shows the results of the LCC test. The number of EVs reported in the table is increased from 134,606 (Leading The Way Scenario) to 300,000 in order to capture an extreme situation. This means 123% more EVs are visiting Newcastle upon Tyne from other areas and need charging their EV batteries. With this increase in the number of EVs, the average operating hours of charging points will increase by 41%. This is due to the fact that the distribution of additional charging requests among the charging points does not have the same proportion as the distribution of the base charging requests. The additional charging requests are served more by Rapid and Ultra-Rapid charging points than by Slow and Fast charging points. Charging the EVs by Rapid and Ultra-Rapid charging points takes less time than charging the EVs by Slow and Fast charging points. Therefore, it takes in average less time to serve the additional charging requests. This means the designed EV charging infrastructure has the capability of absorbing the increased charging demand with the increased average operating hours.

6.10 Comparison With Traditional Genetic Algorithm

When dealing with multi-objective optimisation problems, traditional genetic algorithms typically compress and integrate multiple objectives into a single function, converting the

Table 6.10 Testing the load carrying capacity for *Leading The Way* scenario

Case	Normal	Extreme	Percentage of change
Number of EVs	134,606	300,000	123%
Number of charging point types	6 types	6 types	0%
Total number of charging points	4753	4753	0%
Total expenditure (£)	8,195,000	8,195,000	0%
Average operating hours of charging points (h)	7.57	10.69	41%

problem into a single-objective optimisation. This approach demands additional knowledge of the trade-offs between different objectives and necessitates multiple iterations of tuning the weights in the objective function. This process can be costly and time-consuming.

In contrast, the optimisation method proposed in this chapter retains the multi-objective nature of the problem and provides a solution that automatically captures the trade-offs between the objectives. This is particularly advantageous when considering the distribution of total expenditures and quantities across different charging types. As Figure 6.22 (top) shows, the traditional genetic algorithm suggests higher total expenditures for satisfying the EV energy demand. The quantities of six charging types obtained from the traditional genetic algorithm and the proposed optimisation is shown in Figure 6.22 (bottom). The approach of this chapter suggests having higher number of slow and fast charging points, while traditional genetic optimisation suggest a more uniform distribution for installing charging points.

By maintaining the multi-objective structure of the problem, the new optimisation method offers a more comprehensive and efficient approach to balancing total expenditures and quantities across various charging types. This allows for more informed decision-making and resource allocation when designing and planning electric vehicle charging infrastructure, ultimately contributing to a more sustainable and accessible transportation ecosystem.

6.11 Further Considerations

This section aims to clarify the rationale behind the optimisation model's output that emphasises the slower charging types (e.g., 7kW) based on the data and insights derived from the developed baseline scenario, and addresses the evolving expectations for public EV charging infrastructure.

While the baseline scenario employed in the analysis of this chapter provided a solid foundation, it is significantly influenced by the availability of data, reflecting current trends and consumer behaviours in EV usage and charging practices. This scenario offers a realistic projection of short- to medium-term infrastructure requirements. It aligns with the gradual

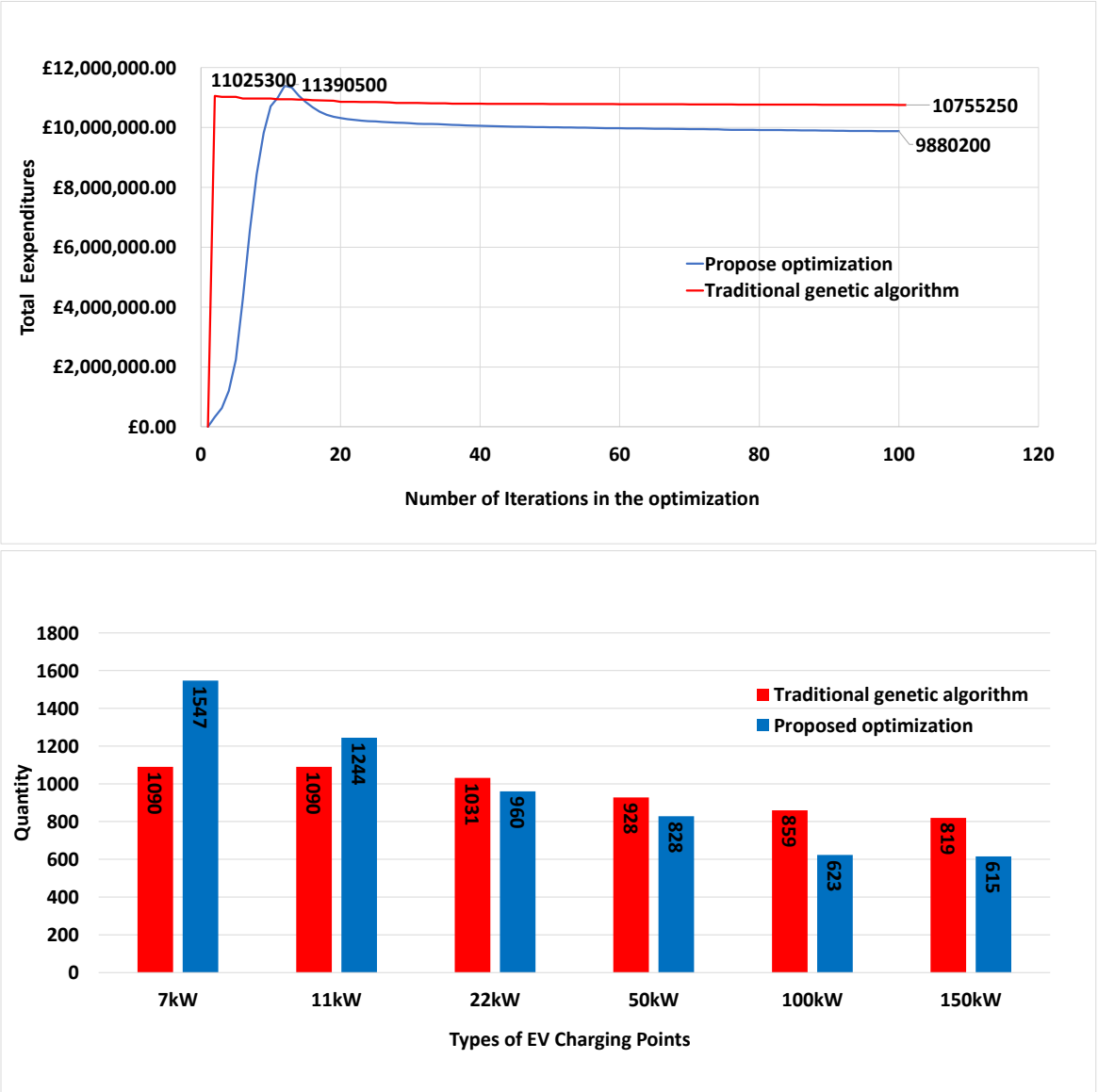

Fig. 6.22 Comparing the new optimisation method with the traditional genetic algorithm

transition towards electric mobility, ensuring that infrastructure development is both feasible and reflective of current technological and societal trends.

While the baseline scenario focuses on the current state and near-future projections, it is expected that the industry's trajectory will be shifted towards integrating faster, rapid, and ultra-rapid charging points. This shift is driven by advances in battery technology, increasing EV range, and consumer demand for shorter charging times, especially in high-traffic public areas. The optimisation framework of this thesis is designed with the flexibility to adapt to evolving charging technologies and future energy scenarios. As new data becomes available and consumer preferences shift towards rapid charging solutions, the framework

can incorporate these changes, ensuring that the EV charging infrastructure recommendations remain relevant and forward-looking. Any observed new trend can also be incorporated in the underlying assumptions used to build the data-generating baseline simulations.

6.12 Conclusions

With focus on the expansion of EV charging infrastructure and data use, this chapter mapped the generic steps of the policy cycle to the steps taken in defining and evaluating the plans for expansion of the EV charging infrastructure. This chapter applied the framework of Chapter 3 to the use case of EV charging infrastructure in Newcastle upon Tyne. A model for simulating the net zero emission policy commitment on EV charging infrastructure was built to compute the increase in the quantity of charging points with different types. The model includes two distinctive stages of simulation and optimisation. The simulation stage of the baseline model has been taken from the industrial partner, Arup Group Limited. The data of travelled distances were utilised and the model was customised to simulate the interaction between EV users and charging points within the model. The main contribution of the results presented in this chapter was on the development of a novel optimisation approach to find the optimal location of EV charging points (best implementation of the policy commitment) utilising the baseline simulation model. The relevant subsets of the output of the simulation model is used for feeding the optimisation stage of the model.

With focus on the EV charging infrastructure of Newcastle upon Tyne, the optimisation model was used to estimate and optimise the charging points types, charging points quantity, charging points locations, total expenditures, and utilisation of charging points. The optimisation was performed for four different future energy scenarios and the demand and EV quantities of the peak years for Newcastle upon Tyne were used. Quantitatively, the optimal solutions recommend installing higher number of faster charging points to reduce the percentage of slower charging points from the current 60% to around 25% in the four scenarios. Still, the optimal solutions put priority on the slower charging points (around 25%), with faster charging points having smaller portions each around 10%-13%. The optimisation shows that while 7kW charging dominates the market currently, it is more beneficial to improve charging efficiency and reduce investment costs by having a higher percentage of installations from other types of charging points in the future installations. Moreover, in the leading scenario for the year 2042 with 134,606 EVs, a total of 4,753 charging points are recommended, resulting in an average operating time of 7.57 hours per charging point. The results also illustrate the spatial distribution of charging points, with higher concentrations in urban areas and near major roads.

In order to study the factors influencing the outcome of the optimisation, a sensitivity analysis was carried out on the number of EVs and the type of charging points. The optimisation results confirmed that having more diversity in the charging point types and installing more high power points can reduce the total expenditure while having similar performance in addressing the charging demand. The analysis revealed the adaptability of the optimisation framework to different scenarios, providing valuable insights for decision-making and highlighting potential areas for improvement.

Chapter 7

Discussions and Conclusions

This chapter provides a comprehensive discussion of the research findings and their implications for validating the objectives of transport policy interventions and implementing policy commitments.

A critical review of the previous research works was covered in Chapter 2. The reviewed papers revealed the limited work in which the success of a policy intervention is assessed using data, and showed the need for a framework to validate policy interventions against their objectives and implement policy commitments using data. The review identified the following research questions in integrating machine learning and optimisation methods for validating transport policy objectives and implementing the policy commitments: **RQ1.** Given the large volume of data, what data types are relevant to the objectives of a policy intervention? **RQ2.** What machine learning techniques are suitable for combining large datasets and validating the intervention objectives? **RQ3.** Could these data-driven techniques be used for efficient optimal implementation of policy commitments?

In order to answer the above research questions, the following objectives were considered: **O1.** Identify, gather, preprocess and analyse data types relevant to a policy from different sources. **O2.** Develop suitable machine learning models based on the input processed data and the considered policy objectives and commitments. **O3.** Analyse and simulate future scenarios under the implementation of the policy commitments to gain insights on their impact in the transport network. **O4.** Study methods for validating the outcome of machine learning methods. Select and use metrics that can best describe the accuracy of the outcome and validate the outcome against domain knowledge. **O5.** Determine the potential use of optimisation methods for transport policy commitment implementation. **O6.** Apply the designed frameworks to case studies on validating the policy objectives of clean air zone and the expansion of the electric vehicle charging infrastructure, which are critical for achieving the UK's target of net-zero emissions by 2050.

By combining machine learning techniques for validating policy intervention objectives and optimisation methods for implementing policy commitments in transport, the designed research aimed to provide valuable insights and practical solutions for transport planning and policy-making. The innovative use of these methods in the chosen case studies demonstrated the potential benefits and challenges of using such technologies to improve transport systems and enhance the decision-making process.

The novelty of the research was in the selection of case studies and in proposing frameworks that combine machine learning and optimisation methods for validating the objectives of policy interventions and implementing policy commitments. For Chapters 4 and 5, the air quality case study was chosen due to its significant impact on public health and the environment. While previous studies have applied machine learning techniques to air quality monitoring and prediction, this research extended the application of machine learning methods to validating the objectives of policy intervention, specifically focusing on clean air zone. In Chapter 6, the electric vehicle charging infrastructure planning was selected as it addresses a crucial challenge in the transition towards sustainable transport systems. While optimisation techniques have been used in various infrastructure planning domains, this research contributed to the literature by developing a multi-objective optimisation framework that integrates machine learning techniques with simulation models to effectively address the challenges, real-world constraints, and multiple conflicting objectives associated with the rapid growth of electric vehicles.

7.1 Addressed Research Questions

The research questions of this study and a brief description of how they were addressed are as follows.

Research Question 1: Given the large volume of data, what data types are relevant to the objectives of a policy intervention?

After a comprehensive literature review in Chapter 2, this thesis proposed a framework to use machine learning methods for identifying the data types that are relevant to a policy objective in Chapter 3. The framework was applied to air quality and clean air zone intervention as a case study in Chapter 4 while considering the policy objective of reducing the concentration of NO_2. The datasets from the Urban Observatory in Newcastle, United Kingdom, were used.

Machine learning classifiers of decision trees (DT), K-nearest neighbours (KNN), gradient-boosted decision trees (GBDT), and light gradient-boosting machine (LGBM) were used to analyse the data from the Newcastle Urban Observatory. Correlation coefficient and feature importances were computed using these models, which were then normalised to get the relative importance of the features. A cut-off value of 1% was used as a proof of concept to identify the most important features. It was shown that the constructed models share common conclusions about the importance of features in predicting NO_2, which could be used in a voting mechanism to decide on the importance of features.

The machine learning models used in the framework stated that O_3, Month, Day, Pressure, and the Number of Cars and Taxis are important features. On the other hand, Wind Speed and Number of two-wheeled motor vehicles were less important in building a model. These findings are also confirmed by the general intuitive observations and evidences about air quality that

- There is a high correlation between NO_2 and O_3;
- The month is important as the air quality can get impacted duo to the seasons and weather conditions; and
- The day will impact the air quality as usually traffic volume might be higher during the working days and lower over the weekends. In the dataset used for building the models, the average difference in traffic volume between weekends and the rest of the week was 35%.

The implementations showed that among the selected learning models, LGBM had the highest average accuracy (88%) and KNN model has the lowest average accuracy (80%) in capturing the relations in the dataset. This shows that the LGBM model can predict the correct class in almost 9 out of 10 cases. While this is an excellent outcome, the accuracy was considered along other metrics including confusion matrix to have the full understanding of the learning and prediction performance in each class.

The findings highlighted the importance of specific data types in predicting NO_2 levels and demonstrated the potential of machine learning in identifying key features for policy objectives.

Research Question 2: What machine learning techniques are suitable for combining large datasets and validating the objectives of the intervention?

This thesis proposed a framework in Chapter 3 to use machine learning methods for validating the objectives of transport policy interventions and checking how well the objectives of the policy intervention are achieved. The framework was applied to air quality and clean air zone intervention as a case study in Chapter 5 while considering the policy objective of reducing the concentration of NO_2. The datasets from the Urban Observatory in Newcastle, United Kingdom, were used. The framework was applied using long short-term memory (LSTM) models for validating the Clean Air Zone policy objective. By utilising historical datasets and assumptions about the policy implementation, the chapter successfully built an LSTM model that predicted NO_2 concentrations with high accuracy without the implementation of the clean air zone. The historical data from the first 10 months of a year were used to build and evaluate the LSTM model, and the predictions are made for the last two months. The LSTM model predicted the NO_2 concentrations with root mean square error of 0.95.

The results demonstrated the use of machine learning methods in analysing and validating the objectives of transport policy interventions. The role of machine learning can be summarised as predicting what is going to happen in the future if the policy is not implement (using available historical data), and predicting the air quality and other related variables using transport behaviour changes in response to the implemented policy.

Table 7.1 shows the concepts from machine learning used in Chapters 4–5.

Research Question 3: Could these data-driven techniques be used for efficient optimal implementation of policy commitments?

This research question was addressed in thesis by proposing a framework in Chapter 3 that integrated data with simulation and optimisation for implementing policy commitments. The framework was applied to the use case of EV charging infrastructure expansion by creating computational models that are capable of simultaneous global search and unsupervised learning.

The study presented in Chapter 6 built a model to simulate the net-zero emission policy commitment related to EV charging infrastructure. This model calculated the increase in the quantity of charging points of different types, incorporating two distinct stages of simulation and optimisation. The simulation stage utilised a baseline model customised to simulate the interaction between EV users and charging points using data on travelled distances.

Table 7.1 Summary of the concepts from machine learning used in Chapters 4–5

Concept	**Purpose**	**Usage in the thesis**
Feature Importance	Identifies influential features on the model outcome.	Identifying data types relevant to a policy objective.
Classification Models	Used to determine feature relevance.	Classifiers can discern the most predictive features, aiding in decision-making for transport policies.
Time Series Models	Analyse data collected over time for prediction.	LSTMs are chosen for their ability to handle long-term dependencies, crucial for predicting pollution levels like NO_2 concentrations.
Performance Metrics	Evaluate the performance of the model in making predictions.	Metrics including accuracy, precision, recall, F1 score, and confusion matrix determine the effectiveness of models in making accurate predictions.

The primary contribution of the presented results lay in the development of a novel optimisation approach for determining the optimal location of EV charging points, effectively implementing the policy commitment. Relevant subsets of the simulation model output fed into the optimisation stage.

Focusing on the EV charging infrastructure in Newcastle upon Tyne, United Kingdom, the optimisation model estimated and optimised charging point types, quantities, locations, total expenditures, and utilisation. The optimisation was conducted for four different future energy scenarios, using peak-year demand and EV quantities for Newcastle upon Tyne. Quantitatively, variations in charging point types were within a 3% range of the total, prioritising slower charging points in optimal solutions.

The study revealed that while 7kW slow charging dominated the market at the time, future installations should prioritise improving efficiency and reducing costs with other charging point types. In the leading scenario for 2042, with 134,606 EVs, the optimal recommendation was 4,753 charging points. The results also depicted the spatial distribution of charging points, with higher concentrations in urban areas and near major roads.

To understand the factors influencing the optimisation outcomes, a sensitivity analysis was conducted on the number of EVs and charging point types. Results confirmed that greater diversity in charging point types and increased installation of high-power points could reduce

total expenditure while maintaining performance in addressing charging demand. The analysis underscored the adaptability of the optimisation framework to different scenarios, offering valuable insights for decision-making and identifying potential areas for improvement.

The proposed optimisation approach proved to be generalisable and applicable to any baseline model providing estimates of future EV charging demands in specific areas. The multi-objective optimisation framework is poised to have offered more scientific and comprehensive support for EV charging infrastructure within the context of a net-zero emission strategy.

The term 'framework' was employed in reference to the presented approaches, as they are characterised by a high level of flexibility, making them applicable to various policy objectives. The specifics and selection of machine learning models can be tailored based on the unique requirements of the intended policy objective.

7.2 Key Findings and Implications for Transport Policy Evaluation

The research presented in this thesis extends our understanding of how advanced techniques such as machine learning, multi-objective optimisation, and simulation can be utilised in validating the objectives and implementing transport policies. It offers valuable insights for local authorities involved in designing and implementing policies including the clean air zone and expansion of EV charging infrastructure in Newcastle upon Tyne.

The Newcastle clean air zone that levies charges for non-compliant taxis, buses, coaches, and heavy goods vehicles, and is set to extend to vans and light goods vehicles, necessitates an evaluation of its effectiveness in enhancing air quality within legal limits and mitigating traffic-related pollution in the city centre. The proposed frameworks play a crucial role in modelling and comprehending the relationship between collected data, implemented charges, and the reduction of air pollution in Newcastle. This research contributes to fostering a more sustainable urban environment by offering insights into the efficacy of clean air zone interventions. These interventions, focused on improving air quality, reducing NO_2 concentrations, and promoting sustainable transportation solutions, aid Newcastle in advancing toward its climate and air quality objectives.

The main takeaways from the research of this thesis are:

- **Identifying Relevant Data for Transport Policy Objective Validation:** This research underlined the role of accurate data selection and gathering in the context of transport policy objectives. Using machine learning classification methods, various data types

related to the policy objectives were identified and analysed. This approach reinforced the significance of finding and incorporating relevant datasets into the objective validation processes of policy interventions.

- **Machine Learning for Validating Policy Objectives:** The application of machine learning techniques for validating policy objectives indicated considerable potential, performing strongly in both classification and time-series forecasting tasks. The proposed use of machine learning with simulations of actual transport conditions offers a potential advancement towards more pragmatic policy objective validation.
- **Multi-objective Optimisation and Simulation for Policy Commitment Implementation:** The research underscored the effectiveness of employing multi-objective optimisation and simulation for policy planning and execution. This method provides a mechanism to strike a balance between disparate objectives, resulting in more sustainable and efficient policy outcomes. It demonstrated significant efficacy, especially in planning an optimal EV charging infrastructure, where factors like cost, location, and user demand were considered.
- **Flexible Data-Driven Frameworks:** The frameworks proposed in this thesis are designed to be high-level and flexible, able to cater to different policy objectives. The specifics can be adjusted based on the unique requirements of each policy objective. This flexibility makes them a valuable tool for policymakers, providing them with an evidence-based methodology to guide their decisions and design effective interventions. Utilising machine learning and data analysis, policymakers can understand which aspects of their transport system have the most significant impact.
- **Value of Data-Driven Approaches:** The research emphasised the role of data-driven methodologies in policy-making. Utilising real-world data, the models produced more accurate and reliable results, highlighting the importance of quality data in evaluating policy objectives. This approach can facilitate better policy decisions and enhancing the efficiency of the implementation.
- **Bridging Theory and Practice:** The findings of this research contribute towards narrowing the gap between academic research and practical application in transport policy-making. It exhibits the real-world application of machine learning, multi-objective optimisation, and simulation in tangible policy contexts, providing a solid direction for both practitioners and researchers.

In summary, the findings of this research provide practical insights that can guide more effective decision-making and enrich the application of data-driven techniques in transport

policy. The study underscores the importance of employing machine learning techniques and optimisation methods to develop frameworks that can assist in the analysis, objective validation of policy interventions, and implementation of transport policy commitments. The three frameworks discussed in this study represent innovative and effective approaches to addressing various challenges and uncertainties in the development and execution of transport policies.

7.3 Limitations of the Proposed Data-Driven Frameworks

7.3.1 Technical and Data Limitations

The proposed frameworks utilise machine learning models based on data. The following aspects should be considered that may limit the extend to which the framework can be applied.

1. **Data Quality**: Machine learning models are highly dependent on the quality of data that is used to train them. Poor quality data can lead to inaccurate or unreliable results, biased models, and reduced genralisation of the model to new data points. This point was observed when analysing the datasets from Newcastle Urban Observatory, which included many invalid or inaccurate data points. These invalid data points and other data points outside of the normal range of variables were discarded from the dataset, otherwise they would have reduced the quality of the trained machine learning models.

2. **Data Quantity**: Machine learning models require datasets that are large enough to achieve high levels of accuracy and performance. This can be a challenge when data is scarce or difficult to obtain. This point was observed when analysing the datasets from Newcastle Urban Observatory. By analysing and preprocessing the datasets and eliminating invalid data points, it became clear that only one year worth of data can be used regarding training and prediction of NO_2 concentrations. For making accurate and long horizon predictions, the time duration in which accurate data is stored must be larger.

3. **Overfitting**: Machine learning models can sometimes learn the training data too well, capturing noise and random fluctuations as if they were genuine patterns. This would result in poor performance on new or unseen data. To address overfitting, techniques such as regularisation or using simpler models should be employed. Additionally, increasing the size of the training dataset can help the model generalise better to new, unseen instances. These points were considered in the first use case of this thesis on

air quality by having machine learning models with different levels of complexity, changing the proportion of training and test dataset, and using appropriate loss functions with regularisation in training the models.

4. **Computing Power**: Machine learning algorithms can be computationally expensive and require significant computing power, making them difficult to implement on smaller or less powerful devices.

5. **Improving baseline simulation models:** The accuracy and reliability of the optimisation solutions obtained by the framework used in Chapter 6 depend on the underlying baseline model that generates the EV demands and also on the quality of the datasets used. Since more data points are becoming available over time, the baseline model must be continuously updated.

To improve the outcomes of the data-driven frameworks of this thesis, it is crucial to continue enhancing the quality of the data, reduce missing data points, and minimise observation errors through precise sensor calibration. In doing so, the effectiveness of interventions can be accurately measured and continually improved, further assisting policymakers in their ongoing efforts to implement better policy interventions.

7.3.2 Limitations of the Frameworks in the Policy Design Cycle

Although the proposed data-driven frameworks contribute to improving and enhancing various stages of the policy cycle, the application of data-driven methods to the transport policy design and implementation has the following limitations:

- **Focus on Measurable and Quantifiable Outcomes:** Data-driven evaluations often focus on easily quantifiable metrics like reduction in travel times or emission levels, potentially neglecting equally important but less quantifiable aspects such as public sentiment, long-term environmental impacts, improved quality of life, and social cohesion. This may lead to missing broader lessons from qualitative analysis and stakeholder feedback, such as community engagement and social equity impacts.

- **Inadequate Qualitative Insights:** Data-driven methods might not fully capture qualitative insights, such as user satisfaction or public acceptance, which are crucial for evaluating the success and acceptance of transport policies.

- **Bias in Data Availability:** Data may be more readily available for urban areas with advanced monitoring systems, leading to a focus on these regions while rural or underserved areas with less data may be overlooked.

- **Lack of Contextual Understanding:** Relying solely on data might miss the underlying causes of transport issues, such as social inequality or economic disparities, that are not easily captured in quantitative datasets.
- **Overreliance on Simulation Models:** Simulation models in transport can oversimplify complex human behaviours and interactions, potentially leading to policies that do not fully address real-world complexities. These models focus on average behaviours and may ignore the diverse mobility needs of different population groups, such as the elderly, disabled, or low-income individuals.
- **Inflexibility in Policy Design:** Data-driven policies might not fully account for future changes in technology, urban development, or social trends, leading to rigid solutions that may become outdated. Therefore, the decisions made in the policy cycle needs to be updated more frequently with respect to the availability of new data and technology. Moreover, data-driven implementation plans may be too rigid, not allowing for adjustments based on local conditions or unforeseen challenges, such as community resistance or unexpected environmental impacts.
- **Risk of Misinterpretation:** Complex data analyses and models can be misunderstood by policymakers, leading to decisions that do not accurately reflect the underlying issues and resulting in inefficient or counterproductive infrastructure investments.
- **Overconfidence in Predictive Models:** Relying heavily on predictive models can create a false sense of certainty, potentially underestimating risks and uncertainties inherent in transport systems.
- **Political Manipulation:** Data can be selectively used to support specific political agendas, such as prioritising certain infrastructure projects for electoral gains rather than based on actual need or effectiveness.
- **Implementation Gaps:** Data-driven methods may not fully capture practical implementation challenges, like funding constraints, stakeholder coordination, or logistical hurdles, leading to gaps between policy design and on-the-ground realities.
- **Insufficient Consideration of Human Factors:** Overemphasis on data might neglect the social and cultural aspects of transport policy, such as how people adapt to new transport modes or infrastructure changes.
- **Attribution Challenges:** In complex transport systems, it can be difficult to attribute specific outcomes directly to particular policies, as many factors can influence re-

sults. For instance, an observed decrease in air pollution might also be influenced by economic downturns or changes in fuel prices that reduce the use of vehicles.

- **Preventing Radical Policy Innovation:** Overreliance on existing data can stifle innovation by discouraging the exploration of new and untested ideas that lack historical data support. For instance, innovative transport solutions like autonomous vehicles might not have enough historical data to support their immediate implementation.

- **Inertia:** Data-driven methods can reinforce existing biases and practices, making it difficult to implement novel solutions such as new public transport systems or alternative mobility schemes.

- **Robustness:** A key challenge in data-driven policy design is dealing with uncertainty, which can arise from various sources such as limited data availability, noisy data, or structural misspecification. In the context of machine learning, policies derived from data are particularly vulnerable to these uncertainties. If these policies are not robust to errors in the underlying simulations or discrepancies between the simulated environment and real-world conditions, their effectiveness in real-world applications can be significantly compromised. Ensuring that learned policies can withstand these uncertainties requires further research for their successful implementation and adaptation to complex and dynamic real-world scenarios.

- **Explainability:** Another aspect of data-driven policy design is the explainability of the learned policies. In many cases, policies generated through machine learning models can be complex, making it difficult for policymakers, stakeholders, and the public to understand the rationale behind specific decisions. This lack of explainability can lead to mistrust and resistance to policy implementation. To ensure that data-driven policies are both effective and widely accepted, further research is needed to make the learned policies interpretable by providing clear explanations of how they are derived, the data and assumptions they rely on, and the underlying mechanisms driving their recommendations. Enhanced explainability not only facilitates informed decision-making but also builds trust and fosters greater accountability in the policy-making process.

- **Simplicity:** For data-driven policies to be widely accepted, simplicity is a highly desirable feature. Policies that are overly complex or intricate can be difficult for stakeholders and the general public to comprehend and support. Clear and straightforward policies are more likely to gain public trust, facilitate smoother implementation, and

encourage broader compliance. Machine learning models can produce highly sophisticated and nuanced outputs. Therefore, there is a risk that the resulting policies may be seen as too complicated or inaccessible. Further research is needed for simplifying these data-driven policies without compromising their effectiveness.

- **Rate of change:** A policy that changes too frequently can create confusion, reduce compliance, and undermine trust among stakeholders, as constant adjustments may be perceived as indecisiveness or lack of reliability. Conversely, a policy that remains static for too long risks becoming outdated and unresponsive to evolving conditions and emerging data. To strike the right balance, further research is needed to establish a framework that allows policies to adapt at a pace that reflects the dynamics of the environment they operate in. This can be achieved by continuously monitoring relevant data, setting predefined thresholds or triggers for policy evaluation, and incorporating feedback loops that inform when and how adjustments should be made.

Incorporating expert knowledge alongside data-driven methods is essential to mitigate the limitations inherent in relying solely on quantitative data for transport policy design and implementation. Experts bring contextual understanding, interpretative skills, and practical experience that complement the empirical rigour of data-driven approaches. By integrating qualitative insights, expert judgement, and nuanced perspectives, policymakers can ensure a more comprehensive understanding of transport issues, addressing not only the quantifiable aspects but also the social, cultural, and behavioural dimensions. This combination enhances the flexibility, relevance, and responsiveness of policies, leading to more effective and sustainable transport solutions that are better aligned with the diverse needs of communities and the dynamic nature of urban environments.

7.4 Ethical Considerations

This research complies with ethical guidelines for data collection and analysis. All datasets used are publicly available, and no personal or sensitive information has been accessed during the research process.

7.5 Validity, Reliability, Scalability, and Transferability

In this thesis, various methods and techniques have been applied to address validity, reliability, scalability, and transferability of research findings throughout the research process.

7.5.1 Validity

The following approaches have been employed to ensure validity:

Expert validation: A key component of the validation process in this research involved the collaboration with consultants from Arup. This partnership provided a deep pool of domain knowledge that was invaluable in refining and validating the models developed in Chapters 4, 5, and 6.

In Chapter 5, the machine learning techniques used for policy objective validation in the air quality case study were thoroughly vetted. The experts from Arup provided critical feedback on the practical implications of the results obtained from these models. This ensured that the machine learning results were grounded in practical considerations that are likely to be effective in the real world.

In Chapter 6, the optimisation framework developed for electric vehicle charging infrastructure planning was subjected to the same rigorous review process. The experts provided insight into realistic constraints, appropriate objectives, and relevant performance indicators in this domain. Their expertise was instrumental in refining the framework, ensuring that it aligns with industry best practices and real-world requirements.

The collaboration process across both chapters included the following:

- The findings and results were shared with the consultants.
- Meetings and discussions were conducted to obtain the consultant's advice and insights.
- This advice was utilised to enhance the practicality and feasibility of the research.
- The improved research was presented back to the consultants for their agreement and further validation.

By engaging with these experts, the research became more reliable and applicable. Their insights ensured that the research outcomes were not just theoretically sound but also practical, thereby increasing the likelihood of acceptance by stakeholders involved in air quality improvement, electric vehicle charging station deployment, and policymakers.

Validating the Results of Chapters 4–5. In order to validate the results in Chapters 4 and 5, where machine learning techniques have been applied for policy objective validation on the air quality case study, the following approaches have been employed:

- Performance metrics: The performance of the classification and LSTM models has been evaluated using relevant metrics such as precision, recall, F1-score, confusion matrix and mean squared error. This helps to quantify the accuracy and effectiveness of the models.

- Model interpretability: To further validate the results, visualisations and explanations of the feature importance and LSTM model's results have been done. This will help to understand the rationale behind the models' predictions and the relationships between the input features and the target variable (NO_2 concentrations).

Validating the Results of Chapter 6. By employing the following diverse validation methods, the robustness and reliability of the optimisation framework for electric vehicle charging infrastructure planning are comprehensively demonstrated. These validation approaches provide strong evidence of the effectiveness and generalisability of the method, increasing confidence in the research findings and their potential applicability in real-world situations.

- Comparison with alternative methods: In this validation approach, the performance of the proposed optimisation solution, which incorporates LSTM and fuzzy logic, has been compared with a simpler genetic algorithm that does not use these techniques. By comparing the outcomes, the added value of the approach is demonstrated. The results showed that the solution provided lower costs and more accurate quantities for the electric vehicle charging infrastructure, indicating that the optimisation framework is more effective than the alternative method.

- Sensitivity analysis: Sensitivity analysis is an important validation technique to test the robustness of an optimisation framework. In this study, sensitivity analysis has been conducted by altering input parameters, such as the number of electric vehicles and the types of charging infrastructure. By analysing the impact of these changes on the outcomes, it is demonstrated that the framework can adapt to different scenarios and produce meaningful results. This also helps to identify potential limitations and areas for improvement in the optimisation method.

7.5.2 Reliability

Reliability refers to the consistency and stability of research findings across different contexts and conditions. In this study, reliability has been addressed by using rigorous and systematic methodologies, thorough data pre-processing, and employing a diverse set of validation techniques for Chapters 4, 5 and 6, as detailed in the previous sections. By ensuring the validity of the models and techniques employed, the reliability of the findings has been improved.

7.5.3 Transferability and Scalability

Transferability refers to the extent to which the research findings can be applied to other contexts, settings, or populations. While the specific datasets and timeframes used in this study may limit the transferability of the findings, the presented framework and the employed methodologies can be applied to other datasets and timeframes, allowing for broader applicability in future research. The diverse validation methods used in both Chapters 4 and 5, including performance metrics, sensitivity analysis, and expert validation, demonstrate the robustness and transferability of the proposed approaches.

In the Machine Learning literature, recent advances have highlighted the effectiveness of *transfer learning*, which is a technique for re-using knowledge learned during data-driven modelling of a problem to boost performance on the data-driven solution of a related problem (Weiss et al., 2016). Transfer learning has been used recently for solving various problems in intelligent transport systems (Baumann et al., 2018; Huang et al., 2021; Ünlü, 2021). The frameworks developed in this thesis can be extended in the context of transfer learning to answer the research questions in other geographical areas. This will be highly effective when the data is expensive or difficult to collect in the considered geographical area.

The frameworks and methodologies developed in this thesis are scalable with respect to the following two aspects:

- The size of the available datasets: The models developed in this thesis using the available datasets can be computed also for larger datasets with larger number of physical variables. The required computational resources will increase with respect to the size of the dataset, but the proposed frameworks will remain applicable.

- The size of the geographical area: The proposed methodologies can be applied to other larger geographical areas. Note that the EV infrastructure case study utilises an aggregate model of the energy consumption, which makes the simulations scalable with respect to the number of EVs the in the network.

Based on the insights gained through the research of this thesis, future research directions will be provided in the next chapter with the goal of setting an outlook for future use of data-driven methods in transport systems.

Chapter 8

Outlook on Future Research

This thesis has presented a comprehensive investigation into transport policy objective validation and the implementation of policy commitments, with a primary focus on quantitative modelling. The research findings have shed light on the effectiveness of machine learning techniques for policy objective validation and the advantages of data-driven methodologies for policy implementation. Through a combination of advanced data analytics, simulation models, and optimisation algorithms, this research has provided valuable insights and actionable recommendations for decision-makers in the transport sector.

One of the key findings of this research is the effectiveness of quantitative modelling in validating the objectives of policy interventions. The application of machine learning models, such as classification and time series models, has demonstrated their ability to predict and evaluate the impacts of clean air zone interventions and the expansion of electric vehicle charging infrastructure. By leveraging large and complex datasets, these models have offered quantitative measures of policy outcomes, enabling a more evidence-based approach to evaluation of policy interventions and implementation of policy commitments.

Quantitative modelling has proven advantageous in several ways. It enables the analysis of vast amounts of data, capturing intricate relationships between variables and providing quantitative indicators of policy effectiveness. The integration of simulation models and multi-objective optimisation has allowed for the efficient implementation of policy commitments, balancing multiple objectives and providing optimal solutions.

However, it is important to acknowledge that quantitative modelling should be complemented by qualitative considerations. Qualitative factors, such as stakeholder engagement and expert opinions, are vital in understanding the social, economic, and political dimensions of transport policies. The integration of both quantitative and qualitative approaches offers a comprehensive and holistic understanding of policy challenges and their potential solutions.

As the field of transport policy continues to evolve, so does the need for advanced, data-driven techniques for policy evaluation and implementation. The research conducted in this thesis has opened up several avenues for future research in the field of transport policy evaluation and commitment implementation. Moving forward, the following research directions are considered to be influential in integrating data-driven methods with transport systems.

- The first use case of this thesis focused on the aggregate effect of the policy intervention on the vehicles entering the zone. It would be interesting to refine this aggregate effect and study how different modes of transport will be affected by the intervention. This will allow for a better understanding of the response of the transport users to the implemented clean air zone, and will provide better insights into applying the framework of this thesis for validating the objectives of policy interventions.

- The proposed frameworks have focused on utilising machine learning models that do not include any information from the physical or chemical models of the variables. The integration of these data-driven frameworks with physics-based models related to the policy interventions could improve the accuracy and performance of the approach.

- To reduce uncertainty in the outcome of the learned models, these models could be updated continuously whenever new additional data points become available. For enabling this continuous model update with real-time data, machine learning models need to be integrated with digital twin models of the transport systems under study. By leveraging digital twin models, policymakers can create virtual replicas of transport systems and simulate policy impacts in real-time. This approach would provide more accurate and up-to-date insights, supporting adaptive policy decision-making.

- Additional work is needed to further enhance the simulation models of transport systems by incorporating factors such as demographic and socioeconomic variables.

- Reducing computational effort of analysing and simulating large-scale transport systems remains a constant challenge in the future. Agent-based modelling is a potential avenue in building simulation models for large-scale transport systems.

- Further exploration of advanced data analytics techniques can enhance the performance and predictive capabilities of quantitative models. Deep learning algorithms, ensemble modelling, reinforcement learning, and other emerging techniques offer opportunities to improve the accuracy of predictions, capture complex relationships in transport data, and handle large-scale and high-dimensional datasets. Future research should focus

on assessing whether the current data quality and data quantity available in relation to transport policies are appropriate for such learning methods that may struggle when the available dataset is limited, and that require a large amount of training data to perform effectively and generalise well to new unseen examples.

- Building on the work of Chapter 6, further research could explore the incorporation of various data types in transport policy formulation and evaluation. This could involve investigating the potential benefits and challenges of integrating unconventional data sources, such as social media data into the transport policy-making process and policy implementations.

- As part of the approach of this thesis, reasonable assumptions were considered on the aggregate response of the transport system to the policy interventions. Future research could explore refining these assumptions and considering more granular aspects that affect the transport policy objectives.

- This research underscored the benefits of bridging the gap between theory and practice in transport policy implementation and their objective validation. Future work should continue to promote collaboration between researchers and practitioners, aiming to facilitate the translation of research findings into practical solutions specifically focused on different stages of data life cycle.

- The results of this thesis, while developed with a focus on specific decarbonisation strategies within the transport sector, have the potential to be applied to other areas, such as active travel infrastructure investment, through an approach analogous to *transfer learning* in machine learning (Huang et al., 2021; Torrey and Shavlik, 2010; Weiss et al., 2016). Just as transfer learning involves taking a model trained on one task and adapting it to perform well on a related but distinct task, the frameworks and methodologies developed in this thesis can be adapted to other decarbonisation policies.

  The principles underlying the identification of relevant datasets, validation of policy objectives, and optimisation of policy implementation are broadly applicable. However, similar to how transfer learning requires fine-tuning a model for a new task, applying these frameworks to different policy areas, such as active travel, would necessitate adjustments. These adaptations would address the unique datasets, stakeholder dynamics, and implementation challenges specific to active travel infrastructure.

- In situations where there are no existing equivalent implementations to learn from, such as when dealing with a radical new policy approach, validating the objectives of

the policy would rely heavily on simulations, pilot studies, and iterative feedback loops. The framework would need to incorporate scenario analysis and modelling to predict potential outcomes, using expert judgement to interpret these results. This approach allows for the validation of policy objectives in a simulated environment, ensuring that the policy is robust and aligned with desired outcomes even in uncharted territory. In the absence of existing models to guide implementation, the optimisation framework would focus on a flexible, adaptive strategy. This could involve phased rollouts, where the policy is implemented incrementally, with continuous monitoring and adjustments based on real-time data and feedback. The framework would employ adaptive management principles, allowing policymakers to refine their approach dynamically as new data and insights become available.

- With the current advances in the development of large language models (LLMs) such as ChatGPT (Du, 2024), LLMs can parse and analyse vast amounts of existing policy documents, research papers, and related literature to identify gaps, trends, and innovative ideas that might not be immediately apparent through traditional methods. LLMs can assist in generating new policy ideas by creatively combining concepts from different domains, leading to the development of radical policies that break new ground. This approach would require develOping an LLM that is trained specifically on existing policy and technical documents.

The outlook for data-driven methods in validation of transport policy objectives and their implementation is highly promising, with ongoing advancements and innovations contributing to transformative changes. These data-driven methods will play a central role in shaping effective and evidence-based transportation policies. They are expected to contribute to evidence-based decision making, dynamic policy adjustments, optimising resource allocation, and achieving environmental and sustainability goals.

References

[1] Md A Abdullah, Kashem M Muttaqi, Ashish P Agalgaonkar, and Danny Sutanto. A noniterative method to estimate load carrying capability of generating units in a renewable energy rich power grid. *IEEE Transactions on sustainable Energy*, 5(3): 854–865, 2014.

[2] Alexander Alexander and Haris Sriwindono. The comparison of genetic algorithm and ant colony optimization in completing travelling salesman problem. In *Proceedings of the 2nd International Conference of Science and Technology for the Internet of Things, ICSTI 2019, September 3rd 2019, Yogyakarta, Indonesia*, 2020.

[3] Taghreed Alghamdi, Sifatul Mostafi, Ghadeer Abdelkader, and Khalid Elgazzar. A comparative study on traffic modeling techniques for predicting and simulating traffic behavior. *Future Internet*, 14(10):294, 2022.

[4] Osyczka Andrzej and Krenich Stanislaw. Evolutionary algorithms for global optimization. *Global Optimization: Scientific and Engineering Case Studies*, pages 267–300, 2006.

[5] Mauro Annunziato and Stefano Pizzuti. Adaptive parameterization of evolutionary algorithms driven by reproduction and competition. In *Proceedings of the European Symposium on Intelligent Techniques (ESIT 2000), Aachen, Germany*, pages 31–35, 2000.

[6] Itamar Arel, Cong Liu, Tom Urbanik, and Airton G Kohls. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems*, 4(2):128–135, 2010.

[7] P Aruna and V Vasan Prabhu. Evolution and recent advancements in electric vehicle (EV) technology. In *Emerging Solutions for e-Mobility and Smart Grids: Select Proceedings of ICRES 2020*, pages 91–109. Springer, 2021.

[8] Ove Arup and Partners Limited. Enabling the roll out of rapid electric vehicle charging stations, 2022. URL https://www.arup.com/projects/assessment-of-electric-vehicle-charging-on-uk-motorways.

[9] PG Balaji, X German, and Dipti Srinivasan. Urban traffic signal control using reinforcement learning agents. *IET Intelligent Transport Systems*, 4(3):177–188, 2010.

[10] John Bates. History of demand modelling. In *Handbook of Transport Modelling: 2nd Edition*, pages 11–34. Emerald Group Publishing Limited, 2007.

[11] Ulrich Baumann, Yuan-Yao Huang, Claudius Gläser, Michael Herman, Holger Banzhaf, and J Marius Zöllner. Classifying road intersections using transfer-learning on a deep neural network. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 683–690. IEEE, 2018.

[12] Islam Safak Bayram, Michael Devetsikiotis, and Raka Jovanovic. Optimal design of electric vehicle charging stations for commercial premises. *International Journal of Energy Research*, 46(8):10040–10051, 2022.

[13] BEIS. Net zero strategy: Build back greener, 2021. URL https://www.gov.uk/government/publications/transitioning-to-zero-emission-cars-and-vans-2035-delivery-plan.

[14] Pauline Bennet, Carola Doerr, Antoine Moreau, Jeremy Rapin, Fabien Teytaud, and Olivier Teytaud. Nevergrad: black-box optimization platform. *ACM SIGEVOlution*, 14(1):8–15, 2021.

[15] Dimitri Bertsekas. *Nonlinear programming*, volume 4. Athena scientific, 2016.

[16] Michel Bierlaire. Discrete choice models. *Operations research and decision aid methodologies in traffic and transportation management*, pages 203–227, 1998.

[17] E Chandra Blessie and E Karthikeyan. Sigmis: a feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, 6(3): 385–394, 2012.

[18] Fergus Bolger and George Wright. Use of expert knowledge to anticipate the future: Issues, analysis and directions, 2017.

[19] Peter Bonsall, Ronghui Liu, and William Young. Modelling safety-related driving behaviourâĂŤimpact of parameter values. *Transportation Research Part A: Policy and Practice*, 39(5):425–444, 2005.

[20] Gianluca Bontempi, Souhaib Ben Taieb, and Yann-Aël Le Borgne. Machine learning strategies for time series forecasting. *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2*, pages 62–77, 2013.

[21] Simon Box and Ben Waterson. An automated signalized junction controller that learns strategies by temporal difference reinforcement learning. *Engineering applications of artificial intelligence*, 26(1):652–659, 2013.

[22] Breathe. Everyone has a right to clean air, 2020. URL https://www.breathe-cleanair.com/. Accessed: 2022-02-20.

[23] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[24] Martin Breunig, Patrick Erik Bradley, Markus Jahn, Paul Kuper, Nima Mazroob, Norbert Rösch, Mulhim Al-Doori, Emmanuel Stefanakis, and Mojgan Jadidi. Geospatial data management research: Progress and future directions. *ISPRS International Journal of Geo-Information*, 9(2):95, 2020.

[25] David Browne and Lisa Ryan. Comparative analysis of evaluation techniques for transport policies. *Environmental Impact Assessment Review*, 31(3):226–233, 2011.

[26] David Brownstone. Discrete choice modeling for transportation. 2001.

[27] Sabine Brunswicker, Laia Pujol Priego, and Esteve Almirall. Transparency in policy making: A complexity view. *Government Information Quarterly*, 36(3):571–591, 2019.

[28] Martin Bulmer, Elizabeth Coates, and Louise Dominian. Evidence-based policy making. In *Making policy in theory and practice*, pages 87–104. Policy Press, 2007.

[29] Hua Cai, Xiaoping Jia, Anthony SF Chiu, Xiaojun Hu, and Ming Xu. Siting public electric vehicle charging stations in beijing using big-data informed travel patterns of the taxi fleet. *Transportation Research Part D: Transport and Environment*, 33:39–46, 2014.

[30] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[31] Jin Cao, Monica Menendez, and Rashid Waraich. Impacts of the urban parking system on cruising traffic and policy development: the case of zurich downtown area, switzerland. *Transportation*, 46:883–908, 2019.

[32] Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. A machine learning approach to predict air quality in california. *Complexity*, 2020, 2020.

[33] Mario Catalano and Fabio Galatioto. Enhanced transport-related air pollution prediction through a novel metamodel approach. *Transportation Research Part D: Transport and Environment*, 55:262–276, 2017.

[34] Briefing Paper Catapult Transport Systems. The case for government involvement to incentivise data sharing in the UK intelligent mobility sector, 2017. URL https://s3-eu-west-1.amazonaws.com/media.ts.catapult/wp-content/uploads/2017/04/12092544/15460-TSC-Q1-Report-Document-Suite-single-pages.pdf.

[35] Marco Cavazzuti. Deterministic optimization. *Optimization Methods: From Theory to Design Scientific and Technological Aspects in Mechanics*, pages 77–102, 2013.

[36] Census. Census 2016 small area population statistics, 2016. URL https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/.

[37] Qianwen Chao, Huikun Bi, Weizi Li, Tianlu Mao, Zhaoqi Wang, Ming C Lin, and Zhigang Deng. A survey on visual traffic simulation: Models, evaluations, and applications in autonomous driving. In *Computer Graphics Forum*, volume 39, pages 287–308. Wiley Online Library, 2020.

[38] Lee Chapman. Transport and climate change: a review. *Journal of transport geography*, 15(5):354–367, 2007.

[39] Vijay Chauhan, Arunima Jaiswal, and Junaid Khan. Web page ranking using machine learning approach. In *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pages 575–580. IEEE, 2015.

[40] Shuyan Chen and Wei Wang. Decision tree learning for freeway automatic incident detection. *Expert systems with applications*, 36(2):4101–4105, 2009.

[41] Tao Chen, Bowen Zhang, Hajir Pourbabak, Abdollah Kavousi-Fard, and Wencong Su. Optimal routing and charging of an electric vehicle fleet for high-efficiency dynamic transit systems. *IEEE Transactions on Smart Grid*, 9(4):3563–3572, 2016.

[42] Tze-Ming Chen, Ware G Kuschner, Janaki Gokhale, and Scott Shofer. Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *The American journal of the medical sciences*, 333(4):249–256, 2007.

[43] Zong-Gan Chen, Zhi-Hui Zhan, Sam Kwong, and Jun Zhang. Evolutionary computation for intelligent transportation in smart cities: a survey. *IEEE Computational Intelligence Magazine*, 17(2):83–102, 2022.

[44] Qiuming Cheng. Quantitative simulation and prediction of extreme geological events. *Science China Earth Sciences*, 65(6):1012–1029, 2022.

[45] Johan Christensen. Expert knowledge and policymaking: a multi-disciplinary research agenda. *Policy & Politics*, 49(3):455–471, 2021.

[46] Clean Air Greater Manchester. Greater manchester clean air plan, 2020. URL https://cleanairgm.com/clean-air-zone/. Accessed: 2022-02-20.

[47] European Commission. Communication from the commission to the european parliament, the council, the european economic and social committee, the committee of the regions and the european investment bank - a policy framework for climate and energy in the period from 2020 to 2030, 2014. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52014DC0015.

[48] Caitlin D Cottrill. Data and digital systems for UK transport: change and its implications. 2018.

[49] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[50] Flávio Cunto and Frank F Saccomanno. Calibration and validation of simulated vehicle safety performance at signalized intersections. *Accident analysis & prevention*, 40(3):1171–1179, 2008.

[51] Guowen Dai, Changxi Ma, and Xuecai Xu. Short-term traffic flow prediction method for urban road sections based on space–time analysis and GRU. *IEEE Access*, 7: 143025–143035, 2019.

[52] James Damsere-Derry, Beth E Ebel, Charles N Mock, Francis Afukaar, Peter Donkor, and Thomas Ojo Kalowole. Evaluation of the effectiveness of traffic calming measures on vehicle speeds and pedestrian injury severity in ghana. *Traffic injury prevention*, 20(3):336–342, 2019.

[53] Huw TO Davies and Sandra M Nutley. *What works?: Evidence-based policy and practice in public services*. Policy Press, 2000.

[54] Giada De Marchi, Giulia Lucertini, and Alexis Tsoukiàs. From evidence-based policy making to policy analytics. *Annals of operations research*, 236(1):15–38, 2016.

[55] Carolina de Mattos Affonso and Mladen Kezunovic. Technical and economic impact of pv-bess charging station on transformer life: A case study. *IEEE Transactions on Smart Grid*, 10(4):4683–4692, 2018.

[56] Kalyanmoy Deb and Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies: Introductory tutorials in optimization and decision support techniques*, pages 403–449. Springer, 2013.

[57] DEFRA. Clean air strategy. 2019. URL https://www.gov.uk/government/publications/clean-air-strategy-2019.

[58] Defra. Emissions Factors Toolkit, 2022. URL https://laqm.defra.gov.uk/air-quality/air-quality-assessment/. Accessed: 2022-02-20.

[59] Defra Joint Air Quality Unit. Clean air zone framework: Principles for setting up clean air zones in England. February 2020.

[60] Pietro Dell'Acqua, Francesco Bellotti, Riccardo Berta, and Alessandro De Gloria. Time-aware multivariate nearest neighbor regression methods for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3393–3402, 2015.

[61] Department for Transport. Road use statistics great britain 2016, 2016. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/514912/road-use-statistics.pdf.

[62] Department for Transport. Road traffic statistics, 2022. URL https://roadtraffic.dft.gov.uk/regions/11. Accessed: 2022-02-20.

[63] Chauhan Dev, Naman Biyani, Nirmal P Suthar, Prashant Kumar, and Priyanshu Agarwal. Structured prediction in NLP: A survey. *arXiv preprint arXiv:2110.02057*, 2021.

[64] Alan Díaz-Manríquez, Gregorio Toscano-Pulido, and Wilfrido Gómez-Flores. On the selection of surrogate models in evolutionary optimization algorithms. In *2011 IEEE congress of evolutionary computation (CEC)*, pages 2155–2162. IEEE, 2011.

[65] Lucian Dimitrievici, Daniel-Eduard Constantin, and Luminita Moraru. The analysis of the correlations between NO2 column, O3 column and uv radiation at global level using space observations. In *AIP Conference Proceedings*, volume 1796, page 030009. AIP Publishing LLC, 2017.

[66] Arnaud Doniec, René Mandiau, Sylvain Piechowiak, and Stéphane Espié. A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, 21(8):1443–1454, 2008.

[67] Driving and Road Transport. Vehicle mileage and occupancy, 2013a. URL https://www.gov.uk/government/statistical-data-sets/nts09-vehicle-mileage-and-occupancy.

[68] Driving and Road Transport. Region and rural-urban classification, 2013b. URL https://www.gov.uk/government/statistical-data-sets/nts99-travel-by-region-and-area-type-of-residence.

[69] Yanliang Du. Large models in transportation infrastructure: a perspective. *Intelligent Transportation Infrastructure*, 3:liae007, 2024.

[70] Luca DâĂŹAcierno, Mariano Gallo, and Bruno Montella. Ant colony optimisation approaches for the transportation assignment problem. *WIT Transactions on the Built Environment*, 111:37–48, 2010.

[71] Jose A Egea, Rafael Martí, and Julio R Banga. An evolutionary method for complex-process optimization. *Computers & Operations Research*, 37(2):315–324, 2010.

[72] Agoston Endre Eiben and Selmar K Smit. Evolutionary algorithm parameters and methods to tune them. *Autonomous search*, pages 15–36, 2012.

[73] Amr Elfar, Alireza Talebpour, and Hani S Mahmassani. Machine learning approach to short-term traffic congestion prediction in a connected environment. *Transportation Research Record*, 2672(45):185–195, 2018.

[74] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of animal ecology*, 77(4):802–813, 2008.

[75] Richard B Ellison, Stephen P Greaves, and David A Hensher. Five years of London's low emission zone: Effects on vehicle fleet composition and air quality. *Transportation Research Part D: Transport and Environment*, 23:25–33, 2013.

[76] European Environment Agency. Air quality in Europe – 2020 report. 2020.

[77] Jonny Evans, Ben Waterson, and Andrew Hamilton. Forecasting road traffic conditions using a context-based random forest algorithm. *Transportation planning and technology*, 42(6):554–572, 2019.

[78] Jonathan E Fieldsend and Richard M Everson. Multi-objective optimisation in the presence of uncertainty. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 243–250. IEEE, 2005.

[79] Thomas B Fischer. Transport policy making and sea in liverpool, amsterdam and berlinâĂŤ1997 and 2002. *Environmental Impact Assessment Review*, 24(3):319–336, 2004.

[80] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[81] Evelyn Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.

[82] Stephanie Forrest. Genetic algorithms. *ACM computing surveys (CSUR)*, 28(1):77–80, 1996.

[83] Ellie Fossey, Carol Harvey, Fiona McDermott, and Larry Davidson. Understanding and evaluating qualitative research. *Australian & New Zealand journal of psychiatry*, 36(6):717–732, 2002.

[84] Federico Gallo, Francesco Corman, and Nicola Sacco. Real-time occupancy predictions of public transport vehicles. In *22nd Swiss Transport Research Conference (STRC 2022)*. STRC, 2022.

[85] Ying Gao, Lei Shi, and Pingjing Yao. Study on multi-objective genetic algorithm. In *Proceedings of the 3rd World Congress on Intelligent Control and Automation (Cat. No. 00EX393)*, volume 1, pages 646–650. IEEE, 2000.

[86] Rodolfo Garcia-Flores, Soumya Banerjee, and George Mathews. Using optimisation and machine learning to validate the value of infrastructure investments. 2017.

[87] Geotab and Met Office. Temperature tool for ev range, 2022. URL https://www.geotab.com/uk/fleet-management-solutions/ev-temperature-tool/.

[88] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*, 3(Aug): 115–143, 2002.

[89] Jonathan L Gifford. Flexible transport systems. In *Handbook of Transport Strategy, Policy and Institutions*, pages 597–611. Emerald Group Publishing Limited, 2005.

[90] Moshe Givoni. Re-assessing the results of the London congestion charging scheme. *Urban Studies*, 49(5):1089–1105, 2012.

[91] Moshe Givoni. Addressing transport policy challenges through policy-packaging, 2014.

[92] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1(2). MIT press Cambridge, 2016.

[93] Stefan Gössling. Urban transport transitions: Copenhagen, city of cyclists. *Journal of Transport Geography*, 33:196–206, 2013.

[94] HM Government. Net zero strategy: Build back greener, 2021a. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1033990/net-zero-strategy-beis.pdf.

[95] UK Government. Workplace charging scheme: guidance for applicants, 2021b. URL https://www.gov.uk/guidance/workplace-charging-scheme-guidance-for-applicants#about-the-scheme.

[96] gov.uk. Congestion charge, 2021. URL https://tfl.gov.uk/modes/driving/congestion-charge.

[97] Gov.UK. Electric vehicle homecharge scheme: guidance for customers, 2021. URL https://www.gov.uk/government/publications/customer-guidance-electric-vehicle-homecharge-scheme/electric-vehicle-homecharge-scheme-guidance-for-customers.

[98] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[99] Yosra Hajjaji, Wadii Boulila, Imed Riadh Farah, Imed Romdhani, and Amir Hussain. Big data and iot-based applications in smart environments: A systematic review. *Computer Science Review*, 39:100318, 2021.

[100] Ronald A Halim, Lucie Kirstein, Olaf Merk, and Luis M Martinez. Decarbonization pathways for international maritime transport: A model-based policy impact assessment. *Sustainability*, 10(7):2243, 2018.

[101] Suqin Han, Hai Bian, Yinchang Feng, Aixia Liu, Xiangjin Li, Fang Zeng, Xiaoling Zhang, et al. Analysis of the relationship between o3, no and no2 in tianjin, china. *Aerosol and Air Quality Research*, 11(2):128–139, 2011.

[102] KS Harishkumar, KM Yogesh, Ibrahim Gad, et al. Forecasting air pollution particulate matter (pm2.5) using machine learning regression models. *Procedia Computer Science*, 171:2057–2066, 2020.

[103] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 485–585, 2009.

[104] M Hatzopoulou and EJ Miller. Transport policy evaluation in metropolitan areas: The role of modelling in decision-making. *Transportation Research Part A: policy and practice*, 43(4):323–338, 2009.

[105] David A Hensher and Kenneth J Button. *Handbook of transport modelling*. Emerald Group Publishing Limited, 2007.

[106] HM Government. The road to zero, 2018. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/739460/road-to-zero.pdf.

[107] HM Government. Transitioning to zero emission cars and vans: 2035 delivery plan, 2021. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1005301/transitioning-to-zero-emission-cars-vans-2035-delivery-plan.pdf.

[108] Kristel MR Hoen, Tarkan Tan, Jan C Fransoo, and Geert-Jan van Houtum. Switching transport modes to meet voluntary carbon emission targets. *Transportation Science*, 48(4):592–608, 2014.

[109] Steven CH Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 6(3):1–26, 2010.

[110] Claire Holman, Roy Harrison, and Xavier Querol. Review of the efficacy of low emission zones to improve urban air quality in european cities. *Atmospheric Environment*, 111:161–169, 2015.

[111] Ye Hong, Yanan Xin, Henry Martin, Dominik Bucher, and Martin Raubal. A clustering-based framework for individual travel behaviour change detection. In *11th International Conference on Geographic Information Science (GIScience 2021)-Part II*, volume 208, page 4. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[112] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2, 2018.

[113] Cheng-Lung Huang and Chieh-Jen Wang. A ga-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with applications*, 31(2): 231–240, 2006.

[114] Yunjie Huang, Xiaozhuang Song, Shiyao Zhang, and JQ James. Transfer learning in traffic prediction with graph neural networks. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3732–3737. IEEE, 2021.

[115] Luke Hutchinson, Ben Waterson, Bani Anvari, and Denis Naberezhnykh. Potential of wireless power transfer for dynamic charging of electric vehicles. *IET intelligent transport systems*, 13(1):3–12, 2019.

[116] Michael Iacono, David Levinson, and Ahmed El-Geneidy. Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature*, 22(4): 323–340, 2008.

[117] Ulrike Illmann and Jan Kluge. Public charging infrastructure and the market diffusion of electric vehicles. *Transportation Research Part D: Transport and Environment*, 86: 102413, 2020.

[118] Ditsuhi Iskandaryan, Francisco Ramos, and Sergio Trilles. Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. *Applied Sciences*, 10(7):2401, 2020.

[119] Daniel J Jacob. *Introduction to atmospheric chemistry*. Princeton university press, 1999.

[120] Marianna Jacyna, Mariusz Wasiak, Konrad Lewczuk, and Michał Kłodawski. Simulation model of transport system of poland as a tool for developing sustainable transport. *Archives of Transport*, 31(3):23–35, 2014.

[121] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3(1):1–25, 2016.

[122] Philip James, Ronnie Das, Agata Jalosinska, and Luke Smith. Smart cities and a data-driven response to COVID-19. *Dialogues in Human Geography*, 10(2):255–259, 2020.

[123] Majid Emami Javanmard and SF Ghaderi. A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data. *Sustainable Cities and Society*, 82:103886, 2022.

[124] Long Jia, Zechun Hu, Yonghua Song, and Zhuowei Luo. Optimal siting and sizing of electric vehicle charging stations. In *2012 IEEE International Electric Vehicle Conference*, pages 1–6. IEEE, 2012.

[125] Zhuojun Jiang, Lei Dong, Lun Wu, and Yu Liu. Quantifying navigation complexity in transportation networks. *PNAS Nexus*, 1(3):pgac126, 2022.

[126] Yaochu Jin, Markus Olhofer, Bernhard Sendhoff, et al. On evolutionary optimization with approximate fitness functions. In *Gecco*, pages 786–793, 2000.

[127] R Daniel Jonsson. Policy application and validation of the transport and land-use model landscapes. 2008.

[128] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[129] Ferhat Karaca. Determination of air quality zones in turkey. *Journal of the Air & Waste Management Association*, 62(4):408–419, 2012.

[130] Vincent Kaufmann, Christophe Jemelin, Géraldine Pflieger, and Luca Pattaroni. Socio-political analysis of french transport policies: The state of the practices. *Transport Policy*, 15(1):12–22, 2008.

[131] Milad Kazemi and Sadegh Soudjani. Formal policy synthesis for continuous-state systems via reinforcement learning. In *Integrated Formal Methods: 16th International Conference, IFM 2020, Lugano, Switzerland, November 16–20, 2020, Proceedings 16*, pages 3–21. Springer, 2020.

[132] Milad Kazemi, Mateo Perez, Fabio Somenzi, Sadegh Soudjani, Ashutosh Trivedi, and Alvaro Velasquez. Translating omega-regular specifications to average objectives for model-free reinforcement learning. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022),*, 2022.

[133] Milad Kazemi, Mateo Perez, Fabio Somenzi, Sadegh Soudjani, Ashutosh Trivedi, and Alvaro Velasquez. Assume-guarantee reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21223–21231, 2024.

[134] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[135] Laura Keast, Lindsay Bramwell, Kamal Jyoti Maji, Judith Rankin, and Anil Namdeo. Air quality outside schools in Newcastle upon Tyne, UK: An investigation into NO2 and PM concentrations and PM respiratory deposition. *Atmosphere*, 13(2):172, 2022.

[136] Katherine A Keith, Christoph Teichmann, Brendan O'Connor, and Edgar Meij. Uncertainty over uncertainty: Investigating the assumptions, annotations, and text measurements of economic policy uncertainty. *arXiv preprint arXiv:2010.04706*, 2020.

[137] Gopalan Kesavaraj and Sreekumar Sukumaran. A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–7. IEEE, 2013.

[138] Masanobu Kii, Hitomi Nakanishi, Kazuki Nakamura, and Kenji Doi. Transportation and spatial development: An overview and a future direction. *Transport Policy*, 49: 148–158, 2016.

[139] Daehyon Kim and Jeong Lee. Application of neural network model to vehicle emissions. *International Journal of Urban Sciences*, 14(3):264–275, 2010.

[140] Kyoungok Kim. Identifying the structure of cities by clustering using a new similarity measure based on smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2002–2011, 2019.

[141] Donald Knuth. Knuth: Computers and typesetting, 2016. URL http://www-cs-faculty.stanford.edu/~uno/abcde.html.

[142] Mrigank Krishan, Srinidhi Jha, Jew Das, Avantika Singh, Manish Kumar Goyal, and Chandrra Sekar. Air quality modelling using long short-term memory (lstm) over nct-delhi, india. *Air Quality, Atmosphere & Health*, 12(8):899–908, 2019.

[143] Manoj Kumar, Dr Husain, Naveen Upreti, Deepti Gupta, et al. Genetic algorithm: Review and application. *Available at SSRN 3529843*, 2010.

[144] Sarah LaMonaca and Lisa Ryan. The state of play in electric vehicle charging services–a review of infrastructure provision, players, and policies. *Renewable and Sustainable Energy Reviews*, 154:111733, 2022.

[145] Michela Le Pira, Edoardo Marcucci, and Valerio Gatta. Role-playing games as a mean to validate agent-based models: An application to stakeholder-driven urban freight transport policy-making. *Transportation Research Procedia*, 27:404–411, 2017.

[146] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[147] Constança Martins Leite de Almeida, Semida Silveira, Erik Jeneulis, and Francesco Fuso-Nerini. Using the sustainable development goals to evaluate possible transport policies for the city of curitiba. *Sustainability*, 13(21):12222, 2021.

[148] Carola Leone, Michela Longo, Luis M Fernández-Ramírez, and Pablo García-Triviño. Multi-objective optimization of pv and energy storage systems for ultra-fast charging stations. *IEEE Access*, 10:14208–14224, 2022.

[149] Huan-Yan Li, Wei Xu, Yanjuan Cui, Zhou Wang, Meng Xiao, and Zhi-Xiao Sun. Preventive maintenance decision model of urban transportation system equipment based on multi-control units. *IEEE Access*, 8:15851–15869, 2019.

[150] Huanhuan Li, Jingxian Liu, Ryan Wen Liu, Naixue Xiong, Kefeng Wu, and Tai-hoon Kim. A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis. *Sensors*, 17(8):1792, 2017.

[151] Linchao Li, Yi Lin, Bowen Du, Fan Yang, and Bin Ran. Real-time traffic incident detection based on a hybrid deep learning model. *Transportmetrica A: transport science*, 18(1):78–98, 2022.

[152] Ming-Hua Lin, Jung-Fa Tsai, and Chian-Son Yu. A review of deterministic optimization methods in engineering and management. *Mathematical Problems in Engineering*, 2012, 2012.

[153] Todd Litman. *Introduction to multi-modal transportation planning*. Victoria Transport Policy Institute Victoria, 2017.

[154] Bowen Liu, John R Bryson, Deniz Sevinc, Matthew A Cole, Robert JR Elliott, Suzanne E Bartington, William J Bloss, and Zongbo Shi. Assessing the impacts of BirminghamâĂŹs clean air zone on air quality: Estimates from a machine learning and synthetic control approach. *Environmental and Resource Economics*, 86(1): 203–231, 2023.

[155] Karl Löfgren and C William R Webster. The value of big data in government: The case of âĂŸsmart citiesâĂŹ. *Big Data & Society*, 7(1):2053951720912775, 2020.

[156] Kathryn G Logan, John D Nelson, and Astley Hastings. Low emission vehicle integration: Will national grid electricity generation mix meet UK net zero? *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 236 (1):159–175, 2022.

[157] Xiangyong Lu, Kaoru Ota, Mianxiong Dong, Chen Yu, and Hai Jin. Predicting transportation carbon emission with urban big data. *IEEE Transactions on Sustainable Computing*, 2(4):333–344, 2017.

[158] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

[159] Yixuan Ma, Zhenji Zhang, and Alexander Ihler. Multi-lane short-term traffic forecasting with convolutional LSTM network. *IEEE Access*, 8:34629–34643, 2020.

[160] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[161] Moataz Mahmoud, Ryan Garnett, Mark Ferguson, and Pavlos Kanaroglou. Electric buses: A review of alternative powertrains. *Renewable and Sustainable Energy Reviews*, 62:673–684, 2016.

[162] Khadija Ait Mamoun, Lamia Hammadi, Abdessamad El Ballouti, and Eduardo Souza De Cursi. Optimisation models for transportation network design under uncertainty: a literature review. *arXiv preprint arXiv:2201.07161*, 2021.

[163] Greg Marsden and Louise Reardon. Questions of governance: Rethinking the study of transportation policy. *Transportation Research Part A: Policy and Practice*, 101: 238–251, 2017.

[164] Sourabh Mehta. How to use genetic algorithm for hyperparameter tuning of ml models, 2022. URL https://analyticsindiamag.com/how-to-use-genetic-algorithm-for-hyperparameter-tuning-of-ml-models/.

[165] Robert Messenger and Lewis Mandell. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, 67(340):768–772, 1972.

[166] Dave Milne and David Watling. Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*, 76:235–244, 2019.

[167] RK Mishra, Manoranjan Parida, and S Rangnekar. Evaluation and analysis of traffic noise along bus rapid transit system corridor. *International Journal of Environmental Science & Technology*, 7:737–750, 2010.

[168] Dietmar PF Möller. Introduction to transportation analysis, modeling and simulation. *Simulation Foundations, Methods and Applications. London: Springer London*, 2014.

[169] Seyed Mohammad Hossein Moosavi, Amiruddin Ismail, and Choon Wah Yuen. Using simulation model as a tool for analyzing bus service reliability and implementing improvement strategies. *PLoS One*, 15(5):e0232799, 2020.

[170] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *Gi/itg workshop mmbnet*, volume 7, 2007.

[171] Daisik Nam, Hyunmyung Kim, Jaewoo Cho, and R Jayakrishnan. A model based on deep learning for predicting travel mode choice. In *Proceedings of the transportation research board 96th annual meeting transportation research board, Washington, DC, USA*, pages 8–12, 2017.

[172] Mahima Nama, Ankita Nath, Nancy Bechra, Jitendra Bhatia, Sudeep Tanwar, Manish Chaturvedi, and Balqies Sadoun. Machine learning-based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. *International Journal of Communication Systems*, 34(9):e4814, 2021.

[173] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62, 2017.

[174] National Institute of Environmental. Air pollution and your health, 2019. URL https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm.

[175] NationalGrid. Future energy scenarios, 2022a. URL https://www.nationalgrideso.com/document/263951/download.

[176] NationalGrid. A net zero future, 2022b. URL https://www.nationalgrideso.com/future-energy/future-energy-scenarios.

[177] United Nations. Climate action, 2015. URL https://www.un.org/en/climatechange/net-zero-coalition.

[178] Ricardo Navares and José L Aznarte. Predicting air quality with deep learning LSTM: Towards comprehensive models. *Ecological Informatics*, 55:101019, 2020.

[179] John L Nazareth. Conjugate gradient method. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):348–353, 2009.

[180] Myriam Neaimeh, Robin Wardle, Andrew M Jenkins, Jialiang Yi, Graeme Hill, Padraig F Lyons, Yvonne Hübner, Phil T Blythe, and Phil C Taylor. A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts. *Applied Energy*, 157:688–698, 2015.

[181] MH Nehrir, Caisheng Wang, Kai Strunz, Hirohisa Aki, Rama Ramakumar, James Bing, Zhixhin Miao, and Ziyad Salameh. A review of hybrid renewable/alternative energy systems for electric power generation: Configurations, control, and applications. *IEEE transactions on sustainable energy*, 2(4):392–403, 2011.

[182] John Nellthorp and Peter Mackie. *Transport Projects, Programmes and Policies: Evaluation Needs and Capabilities*. Routledge, 2017.

[183] Hoang Nguyen, Le-Minh Kieu, Tao Wen, and Chen Cai. Deep learning methods in transportation domain: a review. *IET Intelligent Transport Systems*, 12(9):998–1004, 2018.

[184] Michael Nicholas. Estimating electric vehicle charging infrastructure costs across major us metropolitan areas. *The International Council on Clean Transportation*, 2019.

[185] Saeid Nikbakht, Cosmin Anitescu, and Timon Rabczuk. Optimizing the neural network hyperparameters utilizing genetic algorithm. *Journal of Zhejiang University-Science A*, 22(6):407–426, 2021.

[186] Jovial Niyogisubizo, Lyuchao Liao, Fumin Zou, Guangjie Han, Eric Nziyumva, Ben Li, and Yuyuan Lin. Predicting traffic crash severity using hybrid of balanced bagging classification and light gradient boosting machine. *Intelligent Data Analysis*, 27(1): 79–101, 2023.

[187] Nomis. Car or van availability, 2013a. URL https://www.nomisweb.co.uk/census/2011/ks404ew.

[188] Nomis. Dwellings, household spaces and accommodation type, 2013b. URL https://www.nomisweb.co.uk/census/2011/data_finder.

[189] Nomis. Movement of people and distance between las, 2013c. URL https://www.nomisweb.co.uk/census/2011/ks404ew.

[190] Northhighland worldwide consulting. Local transport data discovery, 2018. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730787/local-transport-data-summary.pdf.

[191] nyc.gov. Vision zero in new york city, 2020. URL https://www.nyc.gov/content/visionzero/pages/.

[192] Office of Air and Radiation (6301A). Air quality guide for nitrogen dioxide, February 2011. URL https://www.airnow.gov/sites/default/files/2018-06/no2.pdf.

[193] OFGEM. Implications of the transition to electric vehicles, 2018. URL https://www.ofgem.gov.uk/sites/default/files/docs/2018/07/ofg1086_future_insights_series_5_document_master_v5.pdf.

[194] Hichem Omrani. Predicting travel mode of individuals by machine learning. *Transportation research procedia*, 10:840–849, 2015.

[195] OpenChargeMap. Open charge map api, 2021. URL https://openchargemap.org/site/develop/api#/.

[196] A Orun, David Elizondo, E Goodyer, and D Paluszczyszyn. Use of bayesian inference method to model vehicular air pollution in local urban areas. *Transportation Research Part D: Transport and Environment*, 63:236–243, 2018.

[197] Colin Palfrey, Paul Thomas, and Ceri Phillips. *Evaluation for the real world: the impact of evidence in policy making*. Policy Press, 2012.

[198] Surbhi Pareek, A Sujil, Saurabh Ratra, and Rajesh Kumar. Electric vehicle charging station challenges and opportunities: A future perspective. In *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, pages 1–6. IEEE, 2020.

[199] Lumír Pečenỳ, Pavol Meško, Rudolf Kampf, and Jozef Gašparík. Optimisation in transport and logistic processes. *Transportation Research Procedia*, 44:15–22, 2020.

[200] Jan A Persson, Paul Davidsson, Stefan J Johansson, and Fredrik Wernstedt. Combining agent-based approaches and classical optimization techniques. In *Third European Workshop on Multi-Agent Systems*, 2005.

[201] John K. Philips. Importance of transportation. 2023.

[202] Roger A Pielke and M Uliasz. Use of meteorological models as input to regional and mesoscale air quality models – limitations and strengths. *Atmospheric environment*, 32(8):1455–1466, 1998.

[203] Transport Planning. Average number of trips made and distance travelled, 2013. URL https://www.gov.uk/government/statistical-data-sets/nts01-average-number-of-trips-made-and-distance-travelled.

[204] Dorina Pojani and Dominic Stead. Policy design for sustainable urban transport in the global south. *Policy Design and Practice*, 1(2):90–102, 2018.

[205] European Union Policy. The deployment of alternative fuels infrastructure, 2017. URL https://eur-lex.europa.eu/eli/dir/2014/94/oj.

[206] Robi Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

[207] Nicholas G Polson and Vadim O Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017.

[208] Boris T Polyak. NewtonâĂŹs method and its use in optimization. *European Journal of Operational Research*, 181(3):1086–1096, 2007.

[209] Pascal Poudenx. The effect of transportation policies on energy consumption and greenhouse gas emission from urban passenger transportation. *Transportation Research Part A: Policy and Practice*, 42(6):901–909, 2008.

[210] André Queirós, Daniel Faria, and Fernando Almeida. Strengths and limitations of qualitative and quantitative research methods. *European journal of education studies*, 2017.

[211] David J Rader. *Deterministic operations research: models and methods in linear optimization*. John Wiley & Sons, 2010.

[212] Nurshahrily Idura Ramli and MI Mohamed Rawi. An overview of traffic congestion detection and classification techniques in vanet. *Indonesian Journal of Electrical Engineering and Computer Science*, 20(1):437–444, 2020.

[213] MF Randolph, MB Jamiolkowski, and L Zdravković. Load carrying capacity of foundations. In *Advances in geotechnical engineering: The Skempton conference: Proceedings of a three day conference on advances in geotechnical engineering, organised by the Institution of Civil Engineers and held at the Royal Geographical Society, London, UK, on 29–31 March 2004*, pages 207–240. Thomas Telford Publishing, 2004.

[214] Khaled Rasheed, Xiao Ni, and Swaroop Vattam. Methods for using surrogate models to speed up genetic algorithm optimization: Informed operators and genetic engineering. *Knowledge Incorporation in Evolutionary Computation*, pages 103–122, 2005.

[215] Hossein Rastgoftar and Jean-Baptiste Jeannin. A physics-based finite-state abstraction for traffic congestion control. *arXiv preprint arXiv:2101.07865*, 2021.

[216] Rajat Rastogi. Promotion of non-motorized modes as a sustainable transportation option: policy and planning issues. *Current Science*, pages 1340–1348, 2011.

[217] Research Oxford University. Policy Engagement, 2024. URL https://www.ox.ac.uk/research/using-research-engage/policy-engagement/guidance-policy-engagement-internationally. Accessed: 2024-08-07.

[218] Marta Rinaldi, Teresa Murino, Elisa Gebennini, Donato Morea, and Eleonora Bottani. A literature review on quantitative models for supply chain risk management: can they be applied to pandemic disruptions? *Computers & Industrial Engineering*, page 108329, 2022.

[219] Jean-Paul Rodrigue, Theo Notteboom, and Brian Slack. 9.1–the nature of transport policy. *The Geography of Transport Systems*, 2023.

[220] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[221] Yves Rybarczyk and Rasa Zalakeviciute. Assessing the covid-19 impact on air quality: A machine learning approach. *Geophysical Research Letters*, 48(4):e2020GL091202, 2021.

[222] S Muhammad Bagher Sadati, Jamal Moshtagh, Miadreza Shafie-Khah, Abdollah Rastgou, and João PS Catalão. Optimal charge scheduling of electric vehicles in solar energy integrated power systems considering the uncertainties. *Electric Vehicles in Energy Systems: Modelling, Integration, Analysis, and Optimization*, pages 73–128, 2020.

[223] Ian Sanderson. Evaluation, policy learning and evidence-based policy making. *Public administration*, 80(1):1–22, 2002.

[224] Georgina Santos, Hannah Behrendt, and Alexander Teytelboym. Part II: Policy instruments for sustainable road transport. *Research in transportation economics*, 28 (1):46–91, 2010.

[225] David Sarabia-Jácome, Carlos E. Palau, Manuel Esteve, and Fernando Boronat. Seaport data space for improving logistic maritime operations. *IEEE Access*, 8:4372–4382, 2020.

[226] Robert G Sargent. Verification and validation of simulation models. In *Proceedings of the 2010 winter simulation conference*, pages 166–183. IEEE, 2010.

[227] Ch Ravi Sekhar, E Madhu, et al. Mode choice analysis using random forrest decision trees. *Transportation Research Procedia*, 17:644–652, 2016.

[228] Bloomberg Professional Services. Enterprise datasets content and data, 2022. URL https://www.bloomberg.com/professional/datasets/?bbgsum-page=DG-WS-PROF-SOLU-DATACONT&mpam-page=21140&tactic-page=429341.

[229] Neha Sharma and Prithwis Kumar De. Application of machine learning to predict CO2 emission from transport sector to mitigate climate change. In *Towards Net-Zero Targets: Usage of Data Science for Long-Term Sustainability Pathways*, pages 197–218. Springer, 2022.

[230] Huan-Jyh Shyur. A quantitative model for aviation safety risk assessment. *Computers & Industrial Engineering*, 54(1):34–44, 2008.

[231] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[232] Office For National Statist. Lower layer super output area population estimates, 2021. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates.

[233] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[234] Gary Stein, Bing Chen, Annie S Wu, and Kien A Hua. Decision tree classifier for network intrusion detection with ga-based feature selection. In *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, pages 136–141, 2005.

[235] Wilma F Strydom, Nikki Funke, Shanna Nienaber, Karen Nortje, and Maronel Steyn. Evidence-based policymaking: a review. *South African Journal of Science*, 106(5): 1–8, 2010.

[236] Andrew Sturman and Peyman Zawar-Reza. Application of back-trajectory techniques to the delimitation of urban clean air zones. *Atmospheric Environment*, 36(20):3339–3350, 2002.

[237] Vignesh Subramanian, Farzaneh Farhadi, and Sadegh Soudjani. Reinforcement learning for stochastic max-plus linear systems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 5631–5638. IEEE, 2023.

[238] A Suleiman, MR Tight, and AD Quinn. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*, 10(1):134–144, 2019.

[239] Bin Sun, Wei Cheng, Prashant Goswami, and Guohua Bai. Short-term traffic forecasting using self-adjusting k-nearest neighbours. *IET Intelligent Transport Systems*, 12 (1):41–48, 2018.

[240] Olle Sundström and Carl Binding. Optimization methods to plan the charging of electric vehicle fleets. In *Proceedings of the international conference on control, communication and power engineering*, pages 28–29, 2010.

[241] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[242] Pasi Tanskanen. The evolutionary structural optimization method: theoretical aspects. *Computer methods in applied mechanics and engineering*, 191(47-48):5485–5498, 2002.

[243] Tharsis Teoh, Bastiaan van Berne, Ivo Hindriks, Rut Waldenfels, Timo Eichhorn, Todor Ivanov, Minsung Hong, Rajendra Akerkar, Maruša Benkic, Jayant Sangwan, et al. Leveraging big data for managing transport operations. *European Commission*, 2019.

[244] Yan Tian, Kaili Zhang, Jianyuan Li, Xianxuan Lin, and Bailin Yang. LSTM-based traffic flow prediction with missing data. *Neurocomputing*, 318:297–305, 2018.

[245] Piyush Tiwari, Hidekazu Itoh, and Masayuki Doi. Shippers' port and carrier selection behaviour in china: a discrete choice analysis. *Maritime Economics & Logistics*, 5: 23–39, 2003.

[246] Ali Tizghadam, Hamzeh Khazaei, Mohammad HY Moghaddam, and Yasser Hassan. Machine learning in transportation, 2019.

[247] Ana Isabel Torre-Bastida, Javier Del Ser, Ibai Laña, Maitena Ilardia, Miren Nekane Bilbao, and Sergio Campos-Cordobés. Big data for transportation and mobility: recent advances, trends and challenges. *IET Intelligent Transport Systems*, 12(8):742–755, 2018.

[248] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[249] Department For Transport. Electric charge point analysis: Domestics, 2018a. URL https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-domestics.

[250] Department For Transport. Electric charge point analysis: Local authority rapids, 2018b. URL https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-local-authority-rapids.

[251] Department For Transport. Electric charge point analysis: Public sector fasts, 2018c. URL https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-public-sector-fasts.

[252] Department For Transport. National trip end model (ntem), 2022. URL https://data.gov.uk/dataset/11bc7aaf-ddf6-4133-a91d-84e6f20a663e/national-trip-end-model-ntem.

[253] Transport for London. London low emission zone: Impacts monitoring baseline report. London, 2008.

[254] Transport for London. Ultra low emission zone, 2021. URL https://tfl.gov.uk/modes/driving/ultra-low-emission-zone.

[255] Alexander Trott, Sunil Srinivasa, Douwe van der Wal, Sebastien Haneuse, and Stephan Zheng. Building a foundation for data-driven, interpretable, and robust policy design using the ai economist. *arXiv preprint arXiv:2108.02904*, 2021.

[256] Energy Saver Trust. What is net zero and how can we get there, 2021. URL https://energysavingtrust.org.uk/what-is-net-zero-and-how-can-we-get-there/.

[257] Alexander J Turner, Jinsol Kim, Helen Fitzmaurice, Catherine Newman, Kevin Worthington, Katherine Chan, Paul J Wooldridge, Philipp Köehler, Christian Frankenberg, and Ronald C Cohen. Observed impacts of COVID-19 on urban co2 emissions. *Geophysical Research Letters*, 47(22):e2020GL090037, 2020.

[258] UK Government. Clean air zones, 2021. URL https://www.gov.uk/guidance/driving-in-a-clean-air-zone#types-of-clean-air-zones.

[259] UK Government. Low-emission vehicles eligible for a plug-in grant, 2021. URL https://www.gov.uk/plug-in-vehicle-grants.

[260] Ramazan Ünlü. Safer-driving: Application of deep transfer learning to build intelligent transportation systems. *Deep Learning and Big Data for Intelligent Transportation: Enabling Technologies and Future Trends*, pages 135–150, 2021.

[261] Anna Urbanek. Data-driven transport policy in cities: A literature review and implications for future developments. In *Scientific And Technical Conference Transport Systems Theory And Practice*, pages 61–74. Springer, 2018.

[262] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53:5929–5955, 2020.

[263] Anne Fleur van Veenstra and Bas Kotterink. Data-driven policy making: The policy lab approach. In *Electronic Participation: 9th IFIP WG 8.5 International Conference, ePart 2017, St. Petersburg, Russia, September 4-7, 2017, Proceedings 9*, pages 100–111. Springer, 2017.

[264] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions*. Springer Science & Business Media, 2014.

[265] Matthew Veres and Medhat Moussa. Deep learning for intelligent transportation systems: A survey of emerging trends. *IEEE Transactions on Intelligent transportation systems*, 21(8):3152–3168, 2019.

[266] Pradnya A Vikhar. Evolutionary algorithms: A critical review and its future prospects. In *2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC)*, pages 261–265. IEEE, 2016.

[267] Shaghayegh Vosough, Hossain Poorzahedy, and Robin Lindsey. Predictive cordon pricing to reduce air pollution. *Transportation Research Part D: Transport and Environment*, 88:102564, 2020.

[268] Peter Vovsha, Eric Petersen, and Robert Donnelly. Microsimulation in travel demand modeling: Lessons learned from the new york best practice model. *Transportation Research Record*, 1805(1):68–77, 2002.

[269] Xiumin Wang, Chau Yuen, Naveed Ul Hassan, Ning An, and Weiwei Wu. Electric vehicle charging station placement for urban public bus systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(1):128–139, 2016.

[270] Yang Wang, Yu Xiao, Jianhui Lai, and Yanyan Chen. An adaptive k nearest neighbour method for imputation of missing traffic data based on two similarity metrics. *Archives of Transport*, 54, 2020.

[271] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[272] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

[273] Terri Wills. Briefing document the ukâĂŹs transition to electric vehicles, 2021. URL https://www.theccc.org.uk/wp-content/uploads/2020/12/The-UKs-transition-to-electric-vehicles.pdf.

[274] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[275] Michalis Xyntarakis and Constantinos Antoniou. Data science and data visualization. In *Mobility Patterns, Big Data and Transport Analytics*, pages 107–144. Elsevier, 2019.

[276] Jie Yang, Jing Dong, and Liang Hu. A data-driven optimization-based approach for siting and sizing of electric taxi charging stations. *Transportation Research Part C: Emerging Technologies*, 77:462–477, 2017.

[277] Yandong Yang. Development of the regional freight transportation demand prediction models based on the regression analysis methods. *Neurocomputing*, 158:42–47, 2015.

[278] Steve Yearley, Steve Cinderby, John Forrester, Peter Bailey, and Paul Rosen. Participatory modelling and the local governance of the politics of UK air pollution: a three-city case study. *Environmental Values*, 12(2):247–262, 2003.

[279] AHR Zaied. Quantitative models for planning and scheduling of flexible manufacturing system. *Emirates Journal for Engineering Research*, 13(2):11–19, 2008.

[280] ZapMap. Ev charging connector types, 2022a. URL https://www.zap-map.com/charge-points/connectors-speeds/.

[281] ZapMap. Ev charging statistics 2022, 2022b. URL https://www.zap-map.com/statistics/#points.

[282] Cong Zhang, Yuanan Liu, Fan Wu, Bihua Tang, and Wenhao Fan. Effective charging planning based on deep reinforcement learning for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 22(1):542–554, 2020.

[283] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639, 2011.

[284] Weibin Zhang, Yinghao Yu, Yong Qi, Feng Shu, and Yinhai Wang. Short-term traffic flow prediction based on spatio-temporal analysis and cnn deep learning. *Transportmetrica A: Transport Science*, 15(2):1688–1711, 2019a.

[285] Xinyu Zhang, Yafei Zheng, and Shouyang Wang. A demand forecasting method based on stochastic frontier analysis and model average: An application in air travel demand forecasting. *Journal of Systems Science and Complexity*, 32(2):615–633, 2019b.

[286] Yanbin Zhao, Kun Zhang, Xiaotian Xu, Huizhong Shen, Xi Zhu, Yanxu Zhang, Yongtao Hu, and Guofeng Shen. Substantial changes in nitrogen dioxide and ozone after excluding meteorological impacts during the COVID-19 outbreak in mainland china. *Environmental Science & Technology Letters*, 7(6):402–408, 2020.

[287] Yuzhe Zhao, Jingmiao Zhou, Yujun Fan, and Haibo Kuang. An expected utility-based optimization of slow steaming in sulphur emission control areas by applying big data analytics. *IEEE Access*, 8:3646–3655, 2019.

[288] Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75, 2017.

[289] Aimin Zhou, Bo-Yang Qu, Hui Li, Shi-Zheng Zhao, Ponnuthurai Nagaratnam Suganthan, and Qingfu Zhang. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and evolutionary computation*, 1(1):32–49, 2011.

[290] Changjiu Zhou. Robot learning with ga-based fuzzy reinforcement learning agents. *Information Sciences*, 145(1-2):45–68, 2002.

[291] Guangyou Zhou, Zhiwei Zhu, and Sumei Luo. Location optimization of electric vehicle charging stations: Based on cost model and genetic algorithm. *Energy*, 247: 123437, 2022.

[292] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. *arXiv preprint arXiv:0906.4258*, 2009.

Appendix A

Contact List for the Datasets

The following persons were contacted related to the datasets relevant to the research of this thesis. The list is in alphabetic order.

1. Nick Bec, City Modelling Lab Business Lead at Arup
2. Alastair Boswell, Director Future Mobility at Arup
3. Grace Carol, Head of Private EV Charging Infrastructure Policy at DFT
4. Gerard Casey, City Modelling Lab at Arup
5. Alex Finkel, Business Intelligence at Nexus
6. Rachelle Forsyth-Ward, Strategic Transport Advisor at Transport Northeast
7. Patrizia Franco, Technical Lead in Transport and Demand Modelling at Connected Places Catapult
8. Martin Gilmour, Deputy director, Planning, Transport and Housing at DFT
9. Louise Guidi, Head of Office and Private Secretary to Chief Scientific Adviser at DFT
10. Neil Harris, Research Associate at Newcastle University (Urban Observatory)
11. John Hodgson, Air Quality and Environmental Permitting at Arup
12. Philip James, Professor at Newcastle University, Director National Urban Observatory Facility Newcastle
13. Jennine Jonczyk, Program Manager at Newcastle University (Urban Observatory)

14. Tristan Joubert, Senior Consultant at Arup
15. Ben Kidd, Research Lead at Arup
16. Philip Meikle, Transport Strategy Director at Transport Northeast
17. Munyaradzki Mutyora, Senior Consultant at Arup
18. Lewis Reed, Intelligent Mobility Engineer at Arup
19. Dominic-AG Taylor, Transport Associate Director at Arup
20. Roger Witte, Principal Transport Modeller at DFT.

Appendix B

Python Code for the Results Reported in Chapter 4

The results reported in Chapter 4 are computed by developing an extensive code in the Python programming environment. The code is reported below.

```
import pandas as pd
import numpy as np
import yellowbrick
from sklearn.model_selection import train_test_split as tts
from sklearn.metrics import accuracy_score , precision_score ,
                                recall_score , roc_auc_score ,
                                confusion_matrix , f1_score
from sklearn.svm import SVC ,LinearSVC
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure

figure(figsize=(27, 23), dpi=80)
plt.rc('font', size=24)
plt.rc('axes', titlesize=28)
plt.rc('axes', labelsize=28)
plt.rc('xtick', labelsize=28)
plt.rc('ytick', labelsize=28)
plt.rc('legend', fontsize=24)
plt.rc('figure', titlesize=24)
plt.rcParams["figure.figsize"] = (25,19)
plt.rcParams["font.family"] = "Times New Roman"


plt.rcParams['axes.facecolor'] = 'white'
```

```
plt.rcParams['figure.facecolor'] = 'white'
plt.rcParams["font.weight"] = "bold"
plt.rcParams["axes.labelweight"] = "bold"
plt.rcParams['axes.titleweight'] = "bold"
plt.rcParams['axes.titlepad'] = 50
plt.rcParams['axes.labelpad'] = 50


import datetime
from datetime import datetime
from sklearn.preprocessing import MinMaxScaler
from keras.models import Sequential
from keras.layers import Dense, Dropout, LSTM
import tensorflow as tf
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import GradientBoostingClassifier
import seaborn as sns
from yellowbrick.target import FeatureCorrelation
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.linear_model import LogisticRegression


def set_tags_no2(labels):
  templ=np.array(labels,dtype='object')
  for i in range(len(labels)):
    if labels[i]>200:
      templ[i]='Very unhealthy'
    elif labels[i]<=200 and labels[i]>150:
      templ[i]='Unhealthy'
    elif labels[i]<=150 and labels[i]>100:
      templ[i]='Unhealthy for sensitive group'
    elif labels[i]<=100 and labels[i]>50:
      templ[i]='Moderate'
    else:
      templ[i]='Good'
  return templ

def parse(x):
  return datetime.strptime(x, '%Y %m %d %H')
def to_integer(dt_time):
    return 10000*dt_time.year + 100*dt_time.month + dt_time.day

def series_to_supervised(data, n_in=1, n_out=1, dropnan=True):
```

```
  n_vars = 1 if type(data) is list else data.shape[1]
  df = pd.DataFrame(data)
  cols, names = list(), list()
  # input sequence (t-n, ... t-1)
  for i in range(n_in, 0, -1):
    cols.append(df.shift(i))
    names += [('var%d(t-%d)' % (j+1, i)) for j in range(n_vars)]
  # forecast sequence (t, t+1, ... t+n)
  for i in range(0, n_out):
    cols.append(df.shift(-i))
    if i == 0:
      names += [('var%d(t)' % (j+1)) for j in range(n_vars)]
    else:
      names += [('var%d(t+%d)' % (j+1, i)) for j in range(n_vars)]
  # put it all together
  agg = pd.concat(cols, axis=1)
  agg.columns = names
  # drop rows with NaN values
  if dropnan:
    agg.dropna(inplace=True)
  return agg

def get_score_after_permutation(model, X, y, curr_feat):
    """ return the score of model when curr_feat is permuted """

    X_permuted = X.copy()
    col_idx = list(X.columns).index(curr_feat)
    # permute one column
    X_permuted.iloc[:, col_idx] = np.random.permutation(
        X_permuted[curr_feat].values)

    permuted_score = model.score(X_permuted, y)
    return permuted_score

def get_feature_importance(model, X, y, curr_feat):
    """ compare the score when curr_feat is permuted """

    baseline_score_train = model.score(X, y)
    permuted_score_train = get_score_after_permutation(model, X, y,
                                        curr_feat)

    # feature importance is the difference between the two scores
    feature_importance = baseline_score_train - permuted_score_train
    return feature_importance
```

```
def permutation_importance(model, X, y, n_repeats=10):
    """Calculate importance score for each feature."""

    importances = []
    for curr_feat in X.columns:
        list_feature_importance = []
        for n_round in range(n_repeats):
            list_feature_importance.append(
                get_feature_importance(model, X, y, curr_feat))

        importances.append(list_feature_importance)

    return {'importances_mean': np.mean(importances, axis=1),
            'importances_std': np.std(importances, axis=1),
            'importances': importances}

def plot_importantes_features(perm_importance_result, feat_name):
    """ bar plot the feature importance """

    fig, ax = plt.subplots(figsize=(10,20))

    indices = perm_importance_result['importances_mean'].argsort()
    y=perm_importance_result['importances_mean'][indices]
    plt.barh(range(len(indices)),
             y,
             xerr=perm_importance_result['importances_std'][indices],
                                                color='blue')

    ax.set_yticks(range(len(indices)))
    _ = ax.set_yticklabels(feat_name[indices])
    for i, v in enumerate(y):
      ax.text(0, i, str(round(v, 2)), color='black', fontweight='bold
                                              ')
    plt.legend()

def correlation(ino,oto):
  X, y = ino, oto

  # Create a list of the feature names
  features = np.array(X.columns.tolist())

  # Instantiate the visualizer
  visualizer = FeatureCorrelation(labels=features)
```

```
  visualizer.fit(X, y)        # Fit the data to the visualizer
  visualizer.show()

def confplt(classifier,X_test,y_test,class_names):
  titles_options = [
    ("Confusion matrix, without normalization", None),
    ("Normalised confusion matrix", "true"),
  ]
  for title, normalize in titles_options:
    disp = ConfusionMatrixDisplay.from_estimator(
        classifier,
        X_test,
        y_test,
        display_labels=class_names,
        cmap=plt.cm.Blues,
        normalize=normalize,
    )
    olaf=""
    if str(type(classifier).__name__)=="LGBMClassifier":
      olaf="LGBM Classifier"
    elif str(type(classifier).__name__)=="DecisionTreeClassifier":
      olaf="DT Classifier"
    elif str(type(classifier).__name__)=="KNeighborsClassifier":
      olaf="KNN Classifier"
    else:
      olaf="GBDT Classifier"
    disp.ax_.set_title(title+" for "+olaf)
    plt.tick_params(axis=u'both', which=u'both',length=0)
    plt.grid(b=None)

"""#data processing"""

DF2018=pd.read_csv('mix2018.csv')
DF2018.reset_index(drop=True, inplace=True)

tf=DF2018['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF2018['two wheeled motor vehicles']=twmv
DF2018['cars and taxis']=cat
DF2018['buses and coaches']=bac
```

```
DF2018['lgvs']=lgvs

labels2018=DF2018['NO2']
templ2018=set_tags_no2(labels2018)
DF2018.NO2=templ2018
trainx18,testx18,trainy18,testy18=tts(DF2018.drop('NO2', axis=1),
                                      DF2018['NO2'],test_size=0.30,
                                      random_state=300)

DF2018.info()

DF2018.describe(include = [object])

"""#classifiers
LGBMClassifier
"""

Xs_of_train=trainx18.values
Ys=trainy18.values
Ys_of_train=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4


log18 = LGBMClassifier(random_state=1234, num_leaves=31,
                                      learning_rate=0.05,
                                      feature_fraction=0.8,
                                      bagging_fraction=0.8, max_depth=-1
                                      )
log18.fit(Xs_of_train, Ys_of_train)
log18_train_res = log18.predict(Xs_of_train)

com_acc_train = accuracy_score(log18_train_res, Ys_of_train)
print(f'Training accuracy : {com_acc_train*100}%')
```

```
"""#classifiers
DecisionTree
"""

Xs_of_train=trainx18.values
Ys=trainy18.values
Ys_of_train=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

DT18 = DecisionTreeClassifier(max_depth=10, min_samples_split=2,
                        min_samples_leaf=1, random_state=1234)
DT18.fit(Xs_of_train, Ys_of_train)

DT18_train_res = DT18.predict(Xs_of_train)
dtacc_trainc = accuracy_score(DT18_train_res, Ys_of_train)

print(f'Training accuracy : {dtacc_trainc*100}%')


"""#classifiers
KNeighborsClassifier
"""

Xs_of_train=trainx18.values
Ys=trainy18.values
Ys_of_train=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
```

```
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

KN18=KNeighborsClassifier(n_neighbors=5)
KN18.fit(Xs_of_train, Ys_of_train)

KN18_train_res = KN18.predict(Xs_of_train)
knacc_trainc = accuracy_score(KN18_train_res, Ys_of_train)
print(f'Training accuracy : {knacc_trainc*100}%')


"""#classifiers
GradientBoostingClassifier
"""

Xs_of_train=trainx18.values
Ys=trainy18.values
Ys_of_train=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

BRT18 = GradientBoostingClassifier(max_depth=10, learning_rate=0.1,
                                   n_estimators=100, random_state=1234)
BRT18.fit(Xs_of_train, Ys_of_train)

BRT18_train_res = BRT18.predict(Xs_of_train)
brtacc_trainc = accuracy_score(BRT18_train_res, Ys_of_train)
print(f'Training accuracy : {brtacc_trainc*100}%')

"""correlation"""

Ys=DF2018["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
```

```
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

correlation(DF2018.drop('NO2', axis=1),Ys_of_train)

"""#results"""

Xs_of_train=trainx18.values
Ys=trainy18.values
Ys_of_train=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4


Xs_of_test=testx18.values
Ys=testy18.values
y_test=np.zeros(Ys.shape)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    y_test[i]=0
  elif(Ys[i]=='Moderate'):
    y_test[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    y_test[i]=2
  elif(Ys[i]=='Unhealthy'):
    y_test[i]=3
  elif(Ys[i]=='Very unhealthy'):
```

```
    y_test[i]=4

#LGBm
y_pred = log18.predict(Xs_of_test)
print("accuracy :", accuracy_score(y_pred, y_test))
print("precision :",precision_score(y_test, y_pred, average=None))
print("recall:",recall_score(y_test, y_pred, average=None))

lgbm18_res=[]
lgbm18_res.append(precision_score(y_test, y_pred, average=None))
lgbm18_res.append(recall_score(y_test, y_pred, average=None))
lgbm18_res.append(f1_score(y_test, y_pred, average=None))
lgbm18_res.append(accuracy_score(y_pred, y_test))
import warnings
warnings.filterwarnings('ignore')
perm_importance_result_train = permutation_importance(
    log18, trainx18, Ys_of_train, n_repeats=15)
class_names=["0","1","2","3","4"]
confplt(log18,Xs_of_test,y_test,class_names)

plot_importantes_features(perm_importance_result_train, trainx18.
                                    columns)

y_pred = DT18.predict(Xs_of_test)

print("accuracy :", accuracy_score(y_pred, y_test))
print("precision :",precision_score(y_test, y_pred, average=None))
print("recall:",recall_score(y_test, y_pred, average=None))

DT18_res=[]
DT18_res.append(precision_score(y_test, y_pred, average=None))
DT18_res.append(recall_score(y_test, y_pred, average=None))
DT18_res.append(f1_score(y_test, y_pred, average=None))
DT18_res.append(accuracy_score(y_pred, y_test))
import warnings
warnings.filterwarnings('ignore')
perm_importance_result_train = permutation_importance(
    DT18, trainx18, Ys_of_train, n_repeats=15)

class_names=["0","1","2","3","4"]
confplt(DT18,Xs_of_test,y_test,class_names)
plot_importantes_features(perm_importance_result_train, trainx18.
                                    columns)
```

```
#KN
y_pred = KN18.predict(Xs_of_test)

print("accuracy :", accuracy_score(y_pred, y_test))
print("precision :",precision_score(y_test, y_pred, average=None))
print("recall:",recall_score(y_test, y_pred, average=None))

KN18_res=[]
KN18_res.append(precision_score(y_test, y_pred, average=None))
KN18_res.append(recall_score(y_test, y_pred, average=None))
KN18_res.append(f1_score(y_test, y_pred, average=None))
KN18_res.append(accuracy_score(y_pred, y_test))
import warnings
warnings.filterwarnings('ignore')
perm_importance_result_train = permutation_importance(
    KN18, trainx18, Ys_of_train, n_repeats=15)

class_names=["0","1","2","3","4"]
confplt(KN18,Xs_of_test,y_test,class_names)
plot_importantes_features(perm_importance_result_train, trainx18.
                                    columns)

#BRT18
y_pred = BRT18.predict(Xs_of_test)

print("accuracy :", accuracy_score(y_pred, y_test))
print("precision :",precision_score(y_test, y_pred, average=None))
print("recall:",recall_score(y_test, y_pred, average=None))

BRT18_res=[]
BRT18_res.append(precision_score(y_test, y_pred, average=None))
BRT18_res.append(recall_score(y_test, y_pred, average=None))
BRT18_res.append(f1_score(y_test, y_pred, average=None))
BRT18_res.append(accuracy_score(y_pred, y_test))
import warnings
warnings.filterwarnings('ignore')
perm_importance_result_train = permutation_importance(
    BRT18, trainx18, Ys_of_train, n_repeats=15)

class_names=["0","1","2","3","4"]
confplt(BRT18,Xs_of_test,y_test,class_names)
plot_importantes_features(perm_importance_result_train, trainx18.
                                    columns)
```

```
xaxis = [2,4,6,8]
xaxis_dec2 = [x-0.60 for x in xaxis]
xaxis_dec = [x-0.30 for x in xaxis]
xaxis_zer = [x for x in xaxis]
xaxis_inc = [x+0.30 for x in xaxis]
xaxis_inc2 = [x+0.60 for x in xaxis]
LABELS = ["LGBM", "DT", "GBDT", 'KNN']

recall_lgbm=lgbm18_res[1]
recall_DT=DT18_res[1]
recall_BRT=BRT18_res[1]
recall_KNN=KN18_res[1]
rec0 = [recall_lgbm[0],recall_DT[0],recall_BRT[0],recall_KNN[0]]
rec1 = [recall_lgbm[1],recall_DT[1],recall_BRT[1],recall_KNN[1]]
rec2 = [recall_lgbm[2],recall_DT[2],recall_BRT[2],recall_KNN[2]]
rec3 = [recall_lgbm[3],recall_DT[3],recall_BRT[3],recall_KNN[3]]
rec4 = [recall_lgbm[4],recall_DT[4],recall_BRT[4],recall_KNN[4]]


ax = plt.subplot(111)
ax.bar(xaxis_dec2, rec0, width=0.15,color='green',align='center',
                                    label = 'Good')
ax.bar(xaxis_dec, rec1, width=0.15,color='orange',align='center',
                                    label = 'Moderate')
ax.bar(xaxis_zer, rec2, width=0.15,color='blue',align='center', label
                                     = 'Unhealthy for sensitive group'
                                    )
ax.bar(xaxis_inc, rec3, width=0.15,color='purple',align='center',
                                    label = 'Unhealthy')
ax.bar(xaxis_inc2, rec4, width=0.15,color='red',align='center', label
                                     = 'Very unhealthy')

ax.set_ylim((0, 1))
plt.xticks(xaxis, LABELS, fontsize=35)
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5)
plt.ylabel('Recall',fontsize=50,fontweight='bold')
plt.title("Recall score",fontsize=50,fontweight='bold')
plt.show()

xaxis = [2,4,6,8]
xaxis_dec2 = [x-0.60 for x in xaxis]
xaxis_dec = [x-0.30 for x in xaxis]
xaxis_zer = [x for x in xaxis]
```

```
xaxis_inc = [x+0.30 for x in xaxis]
xaxis_inc2 = [x+0.60 for x in xaxis]
LABELS = ["LGBM", "DT", "GBDT", 'KNN']

f1_lgbm=lgbm18_res[2]
f1_DT=DT18_res[2]
f1_BRT=BRT18_res[2]
f1_KNN=KN18_res[2]
rec0 = [f1_lgbm[0],f1_DT[0],f1_BRT[0],f1_KNN[0]]
rec1 = [f1_lgbm[1],f1_DT[1],f1_BRT[1],f1_KNN[1]]
rec2 = [f1_lgbm[2],f1_DT[2],f1_BRT[2],f1_KNN[2]]
rec3 = [f1_lgbm[3],f1_DT[3],f1_BRT[3],f1_KNN[3]]
rec4 = [f1_lgbm[4],f1_DT[4],f1_BRT[4],f1_KNN[4]]

ax = plt.subplot(111)
ax.bar(xaxis_dec2, rec0, width=0.15,color='green',align='center',
                                    label = 'Good')
ax.bar(xaxis_dec, rec1, width=0.15,color='orange',align='center',
                                    label = 'Moderate')
ax.bar(xaxis_zer, rec2, width=0.15,color='blue',align='center', label
                                     = 'Unhealthy for sensitive group'
                                    )
ax.bar(xaxis_inc, rec3, width=0.15,color='purple',align='center',
                                    label = 'Unhealthy')
ax.bar(xaxis_inc2, rec4, width=0.15,color='red',align='center', label
                                     = 'Very unhealthy')

ax.set_ylim((0, 1))
plt.xticks(xaxis, LABELS,fontsize=40)
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5,fontsize=30)
plt.ylabel('F1',fontsize=50,fontweight='bold')
plt.title("F1 score",fontsize=50,fontweight='bold')
plt.show()

xaxis = [2,4,6,8]
xaxis_dec2 = [x-0.60 for x in xaxis]
xaxis_dec = [x-0.30 for x in xaxis]
xaxis_zer = [x for x in xaxis]
xaxis_inc = [x+0.30 for x in xaxis]
xaxis_inc2 = [x+0.60 for x in xaxis]
LABELS = ["LGBM", "DT", "GBDT", 'KNN']

precision_lgbm=lgbm18_res[0]
```

```
precision_DT=DT18_res[0]
precision_BRT=BRT18_res[0]
precision_KNN=KN18_res[0]
rec0 = [precision_lgbm[0],precision_DT[0],precision_BRT[0],
                                    precision_KNN[0]]
rec1 = [precision_lgbm[1],precision_DT[1],precision_BRT[1],
                                    precision_KNN[1]]
rec2 = [precision_lgbm[2],precision_DT[2],precision_BRT[2],
                                    precision_KNN[2]]
rec3 = [precision_lgbm[3],precision_DT[3],precision_BRT[3],
                                    precision_KNN[3]]
rec4 = [precision_lgbm[4],precision_DT[4],precision_BRT[4],
                                    precision_KNN[4]]



ax = plt.subplot(111)
ax.bar(xaxis_dec2, rec0, width=0.15,color='green',align='center',
                                    label = 'Good')
ax.bar(xaxis_dec, rec1, width=0.15,color='orange',align='center',
                                    label = 'Moderate')
ax.bar(xaxis_zer, rec2, width=0.15,color='blue',align='center', label
                                     = 'Unhealthy for sensitive group'
                                    )
ax.bar(xaxis_inc, rec3, width=0.15,color='purple',align='center',
                                    label = 'Unhealthy')
ax.bar(xaxis_inc2, rec4, width=0.15,color='red',align='center', label
                                     = 'Very unhealthy')

ax.set_ylim((0, 1))
plt.xticks(xaxis, LABELS, fontsize=30)
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5)
plt.ylabel('Precision', fontsize=50,fontweight='bold')
plt.title("Precision score", fontsize=80,fontweight='bold')
plt.show()

xaxis = [1,2,3,4]
xaxis_dec = [x for x in xaxis]
xaxis_inc = [x for x in xaxis]
LABELS = ["LGBM", "DT", "GBDT", 'KNN']

acc_lgbm=lgbm18_res[3]
acc_DT=DT18_res[3]
```

```
acc_BRT=BRT18_res[3]
acc_KNN=KN18_res[3]
rec0 = [acc_lgbm,acc_DT,acc_BRT,acc_KNN]


ax = plt.subplot(111)
ax.bar(xaxis_dec, rec0, width=0.2,color=['green','orange','blue','
                                    pink'],align='center')

ax.set_ylim((0, 1))
plt.xticks(xaxis, LABELS, fontsize=30)
plt.xlabel('Methods', fontsize=50,fontweight='bold')
plt.ylabel('Accuracy', fontsize=50,fontweight='bold')
plt.title("Accuracy score", fontsize=50,fontweight='bold')
plt.show()
```

Appendix C

Python Code for the Results Reported in Chapter 5

The results reported in Chapter 5 are computed by developing an extensive code in the Python programming environment. The code is reported below.

```
DF2018=pd.read_csv('mix2018.csv')
DF2018.reset_index(drop=True, inplace=True)

tf=DF2018['Traffic Flow'].values
twmv=tf* 0.0063
cat=tf*0.8
bac=tf*0.0124
lgvs=tf*0.1813
DF2018['two wheeled motor vehicles']=twmv
DF2018['cars and taxis']=cat
DF2018['buses and coaches']=bac
DF2018['lgvs']=lgvs


DF2018=DF2018.drop('Traffic Flow',axis=1)
DF2018=DF2018.drop('Congestion',axis=1)

DF2018.drop('hour',axis=1,inplace=True)
DF2018.reset_index(drop=True, inplace=True)

labels=DF2018['NO2'].values
templ=set_tags_no2(labels)
DF2018.NO2=templ
DF2018.fillna(0,inplace=True)
```

```
uncoded_Y=DF2018.NO2.values
Coded_Y=np.zeros(uncoded_Y.shape)
for i in range(len(uncoded_Y)):
  if(uncoded_Y[i]=='Good'):
    Coded_Y[i]=0
  elif(uncoded_Y[i]=='Moderate'):
    Coded_Y[i]=1
  elif(uncoded_Y[i]=='Unhealthy for sensitive group'):
    Coded_Y[i]=2
  elif(uncoded_Y[i]=='Unhealthy'):
    Coded_Y[i]=3
  elif(uncoded_Y[i]=='Very unhealthy'):
    Coded_Y[i]=4
DF2018.NO2=Coded_Y


DF18=DF2018
date18=[]
for i in range(len(DF18)):
  date18.append(i+1)
DF18.drop(['year','month','day'],axis=1,inplace=True)
DF18['date_by_day']=date18

DF18.index = DF18['date_by_day']
DF18=DF18.sort_index(ascending=True, axis=0)
plt.figure(figsize=(16,8))
plt.ylim(ymin=0,ymax=4)
plt.plot(DF18['NO2'], label='NO2 level')


trainx18,testx18,trainy18,testy18=tts(DF18.drop('NO2', axis=1),DF18['
                                  NO2'],test_size=0.30,random_state=
                                  300)



#process 2018 data for lstm
scaler = MinMaxScaler(feature_range=(0, 1))
values18 = scaler.fit_transform(trainx18)
trainx18 = series_to_supervised(values18, 1, 1)
tsvalues18 = scaler.fit_transform(testx18)
testx18 = series_to_supervised(tsvalues18, 1, 1)
testlab18=testy18
```

```
trainlab18=trainy18

trainx18=trainx18.values
testx18=testx18.values
trainlab18=trainlab18[1:]
testlab18=testlab18[1:]
trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1])
                                    )
testx18 = testx18.reshape((testx18.shape[0], 1, testx18.shape[1]))
print(trainx18.shape, trainlab18.shape, testx18.shape, testlab18.
                                    shape)

"""#model"""

#2018
model18 = Sequential()
model18.add(LSTM(200, input_shape=(trainx18.shape[1], trainx18.shape[
                                    2])))
model18.add(Dense(1))
model18.compile(loss="mean_squared_error", optimizer='sgd')
# fit network
#,
history18pre = model18.fit(trainx18, trainlab18, epochs=500,
                                    validation_data=(testx18,
                                    testlab18), batch_size=1, verbose=
                                    2, shuffle=True)
# plot history
plt.plot(history18pre.history['loss'], label='Train')
plt.plot(history18pre.history['val_loss'], label='Test')
plt.xlabel('Epoch')
plt.xlim(xmin=0)
plt.ylabel('MSE')
plt.legend()
plt.show()

from sklearn.metrics import mean_squared_error
yhat18 = model18.predict(testx18)
print(yhat18.tolist())

testlab18_new=testlab18.values.tolist()
print(testlab18_new)

rmse18 = np.sqrt(mean_squared_error(testlab18, yhat18))
```

```
total_rms = np.sqrt(mean_squared_error(testlab18, np.zeros(len(
                                    testlab18_new))))
total_rms2 = np.sqrt(mean_squared_error(yhat18, np.zeros(len(
                                    testlab18_new))))
rmse18per = 100*rmse18/total_rms

print('Test RMSE for 2018: %.3f' % rmse18)
print('Test root mean square for 2018: %.3f' % total_rms)
print('Prediction root mean square for 2018: %.3f' % total_rms2)
print('Test RMSE percent for 2018: %.3f' % rmse18per)

predict_these=[2,8,10,20,30,40,50,60,70]
ys=model18.predict(testx18[predict_these,:,:])

for i in range(len(predict_these)):
  print('predicted ',np.round(ys[i]),'for day',predict_these[i],'of
                                    2018. the true value is',
                                    testlab18_new[predict_these[i]])

DF18=pd.read_csv('mix2018.csv')
monthvals=DF18.month.unique()


tf=DF18['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF18['two wheeled motor vehicles']=twmv
DF18['cars and taxis']=cat
DF18['buses and coaches']=bac
DF18['lgvs']=lgvs
DF18.drop('Traffic Flow',axis=1)
labels2018=DF18['NO2']
templ2018=set_tags_no2(labels2018)
DF18.NO2=templ2018
Ys=DF18["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
```

```
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

DF18.NO2=Ys_of_train


print(monthvals)
for j in monthvals:

  tempdf=DF18.loc[DF18['month'] == j]
  trainx18=tempdf.drop('NO2', axis=1)

  scaler = MinMaxScaler(feature_range=(0, 1))
  values18 = scaler.fit_transform(trainx18)
  trainx18 = series_to_supervised(values18, 1, 1)
  trainlab18=tempdf['NO2'].values

  trainx18=trainx18.values
  trainlab18=trainlab18[:48]
  trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1
                                        ]))
  trainx18 =trainx18[:,:,:48]
  yhats = model18.predict(trainx18)
  scaler = MinMaxScaler(feature_range=(0, 4))
  yhats = scaler.fit_transform(yhats)
  plt.plot(yhats)
  plt.yticks([0,1,2,3,4])
  plt.xlabel('Time and values serialized')
  plt.ylabel('NO2 Labels prediction')
  plt.title('Prediction by month '+str(j)+'th before intervention')
  # 1st 2nd 3rd 4th 5th 6th 7th
  plt.xlim(xmin=0)
  plt.ylim(ymin=0,ymax=4)
  plt.legend()
  plt.show()

"""#Policy"""

DF18=pd.read_csv('mix2018.csv')

DF18.reset_index(drop=True, inplace=True)
```

```
tf=DF18['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF18['two wheeled motor vehicles']=twmv
DF18['cars and taxis']=cat
DF18['buses and coaches']=bac
DF18['lgvs']=lgvs

DF18=DF18.drop('Traffic Flow',axis=1)
DF18=DF18.drop('Congestion',axis=1)


DF18.drop('hour',axis=1,inplace=True)
DF18.reset_index(drop=True, inplace=True)

# 18
twmv18=DF18["two wheeled motor vehicles"].values
twmv18=np.round(twmv18*0.8)
cat18=DF18["cars and taxis"].values
cat18=np.round(cat18*0.8)
bac18=DF18["buses and coaches"].values
bac18=np.round(bac18*0.9)
lgvs18=DF18["lgvs"].values
lgvs18=np.round(lgvs18*0.8)
co18=DF18["CO"].values
co18=co18-18
pm2518=DF18["PM2_5"].values
pm2518=pm2518 -16
par18=DF18["Particle Count"].values
par18=par18 -10
pm118=DF18["PM1"].values
pm118=pm118 -10
pm1018=DF18["PM10"].values
pm1018=pm1018 -21
pm418=DF18["PM 4"].values
pm418=pm418 -23
o318=DF18["O3"].values
o318=o318 -29
no18=DF18["NO"].values
no18=no18 -18
```

```
nox18=DF18["NOx"].values
nox18=nox18 -24
no218=DF18["NO2"].values
no218=no218 -25


DF18["two wheeled motor vehicles"]=twmv18
DF18["cars and taxis"]=cat18
DF18["buses and coaches"]=bac18
DF18["lgvs"]=lgvs18
DF18["CO"]=co18
DF18["PM2_5"]=pm2518
DF18["Particle Count"]=par18
DF18["PM1"]=pm118
DF18["PM10"]=pm1018
DF18["PM 4"]=pm418
DF18["O3"]=o318
DF18["NO"]=no18
DF18["NOx"]=nox18
DF18["NO2"]=no218

labels18=DF18['NO2'].values
templ18=set_tags_no2(labels18)
DF18.NO2=templ18
DF18.fillna(0,inplace=True)


uncoded_Y=DF18.NO2.values
Coded_Y=np.zeros(uncoded_Y.shape)
for i in range(len(uncoded_Y)):
  if(uncoded_Y[i]=='Good'):
    Coded_Y[i]=0
  elif(uncoded_Y[i]=='Moderate'):
    Coded_Y[i]=1
  elif(uncoded_Y[i]=='Unhealthy for sensitive group'):
    Coded_Y[i]=2
  elif(uncoded_Y[i]=='Unhealthy'):
    Coded_Y[i]=3
  elif(uncoded_Y[i]=='Very unhealthy'):
    Coded_Y[i]=4
DF18.NO2=Coded_Y


Af18=DF18.NO2.values
```

```
trainx18,testx18,trainy18,testy18=tts(DF18.drop('NO2', axis=1),DF18['
                                    NO2'],test_size=0.30,random_state=
                                    300)


scaler = MinMaxScaler(feature_range=(0, 1))
values18 = scaler.fit_transform(trainx18)
trainx18 = series_to_supervised(values18, 1, 1)
tsvalues18 = scaler.fit_transform(testx18)
testx18 = series_to_supervised(tsvalues18, 1, 1)
testlab18=testy18
trainlab18=trainy18

trainx18=trainx18.values
testx18=testx18.values
trainlab18=trainlab18[1:]
testlab18=testlab18[1:]
trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1])
                                    )
testx18 = testx18.reshape((testx18.shape[0], 1, testx18.shape[1]))
print(trainx18.shape, trainlab18.shape, testx18.shape, testlab18.
                                    shape)

DF18.columns

DF18

DF18.columns

DF18p=DF18
DF18p.drop(['year','month','day'],inplace=True,axis=1)

exclude = ['NOx', 'Brood']
DF18["two wheeled motor vehicles"] = DF18["two wheeled motor vehicles
                                    "] + 30
DF18["buses and coaches"] = DF18["buses and coaches"] + 0.89
DF18p = DF18p.loc[:, DF18p.columns.difference(exclude)]

DF18p.hist(figsize=(26,26), layout=(7,4), color = 'green')
plt.xticks(fontsize=40)
font = {'family' : 'normal',
        'weight' : 'bold',
```

```
        'size'   : 40}
plt.rc('font', **font)
plt.subplots_adjust(left=4, bottom=4, right=5, top=5)
plt.xlim(xmin=0)
plt.show()



# #print quartile plot for outlier detection
DF18p.plot(kind='box', subplots=True, figsize=(26,26), layout=(8,4),
            sharex=False, sharey=False, fontsize=(18), color='red')
plt.subplots_adjust(left=4, bottom=4, right=5, top=5)
plt.show()

#2018
model18a = Sequential()
model18a.add(LSTM(200, input_shape=(trainx18.shape[1], trainx18.shape
                                    [2])))
model18a.add(Dense(1))
model18a.compile(loss='mean_squared_error', optimizer='sgd')
# fit network
#,
history18pos = model18a.fit(trainx18, trainlab18, epochs=500,
                                    validation_data=(testx18,
                                    testlab18), batch_size=1, verbose=
                                    2, shuffle=True)
# plot history
plt.plot(history18pos.history['loss'], label='Train')
plt.plot(history18pos.history['val_loss'], label='Test')
plt.xlabel('Epoch')
plt.ylabel('MSE')
plt.xlim(xmin=0)
plt.legend()
plt.show()

from sklearn.metrics import mean_squared_error
yhat18af = model18a.predict(testx18)

testlab18_new=testlab18.values.tolist()
print(testlab18_new)

rmse18 = np.sqrt(mean_squared_error(testlab18, yhat18af))
rmse18per = 100*rmse18/np.sqrt(np.sum(testlab18**2))
```

```
print('Test RMSE for 2018: %.3f' % rmse18)
print('Test RMSE percent for 2018: %.3f' % rmse18per)

predict_these=[2,8,10,20,30,40,50,60,70]
ys=model18a.predict(testx18[predict_these,:,:])

print("\n")
for i in range(len(predict_these)):
  print('predicted ',np.round(ys[i]),'for day',predict_these[i])

plt.plot(history18pre.history['val_loss'], label='Loss before policy'
                                    ,color= 'firebrick',linewidth=5)
plt.plot(history18pos.history['val_loss'], label='Loss after policy',
                                    color= 'darkcyan',linewidth=5)
plt.xlabel('Epoch',fontsize=40, fontweight='bold')
plt.ylabel('MSE',fontsize=40, fontweight='bold')
plt.ylim(ymin=0,ymax=2.5)
plt.xlim(xmin=0,xmax=280)
plt.legend(fontsize=40)
plt.show()

DF18=pd.read_csv('mix2018.csv')
monthvals=DF18.month.unique()


tf=DF18['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF18['two wheeled motor vehicles']=twmv
DF18['cars and taxis']=cat
DF18['buses and coaches']=bac
DF18['lgvs']=lgvs

# 18
twmv18=DF18["two wheeled motor vehicles"].values
twmv18=np.round(twmv18*0.8)
cat18=DF18["cars and taxis"].values
cat18=np.round(cat18*0.8)
bac18=DF18["buses and coaches"].values
bac18=np.round(bac18*0.9)
lgvs18=DF18["lgvs"].values
lgvs18=np.round(lgvs18*0.8)
```

```
co18=DF18["CO"].values
co18=co18-18
pm2518=DF18["PM2_5"].values
pm2518=pm2518 -16
par18=DF18["Particle Count"].values
par18=par18 -10
pm118=DF18["PM1"].values
pm118=pm118 -10
pm1018=DF18["PM10"].values
pm1018=pm1018 -21
pm418=DF18["PM 4"].values
pm418=pm418 -23
o318=DF18["O3"].values
o318=o318 -29
no18=DF18["NO"].values
no18=no18 -18
nox18=DF18["NOx"].values
nox18=nox18 -24
no218=DF18["NO2"].values
no218=no218 -25


DF18["two wheeled motor vehicles"]=twmv18
DF18["cars and taxis"]=cat18
DF18["buses and coaches"]=bac18
DF18["lgvs"]=lgvs18
DF18["CO"]=co18
DF18["PM2_5"]=pm2518
DF18["Particle Count"]=par18
DF18["PM1"]=pm118
DF18["PM10"]=pm1018
DF18["PM 4"]=pm418
DF18["O3"]=o318
DF18["NO"]=no18
DF18["NOx"]=nox18
DF18["NO2"]=no218

labels2018=DF18['NO2']
templ2018=set_tags_no2(labels2018)
DF18.NO2=templ2018
Ys=DF18["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
```

```
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

DF18.NO2=Ys_of_train


print(monthvals)
for j in monthvals:

  tempdf=DF18.loc[DF18['month'] == j]
  trainx18=tempdf.drop('NO2', axis=1)

  scaler = MinMaxScaler(feature_range=(0, 1))
  values18 = scaler.fit_transform(trainx18)
  trainx18 = series_to_supervised(values18, 1, 1)
  trainlab18=tempdf['NO2'].values

  trainx18=trainx18.values
  trainlab18=trainlab18[:52]
  trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1
                                ]))
  trainx18 =trainx18[:,:,:52]
  yhats = model18a.predict(trainx18)
  scaler = MinMaxScaler(feature_range=(0, 4))
  yhats = scaler.fit_transform(yhats)
  plt.plot(yhats)
  plt.xlabel('Time and values serialized')
  plt.ylabel('NO2 labels prediction')
  plt.title('Prediction by month '+str(j)+'th after intervention')
  plt.yticks([0,1,2,3,4])
  plt.ylim(ymin=0,ymax=4)
  plt.xlim(xmin=0)
  plt.legend()
  plt.show()

DF2018=pd.read_csv('mix2018.csv')
```

```
# print(DF2019.head())
DF2018.reset_index(drop=True, inplace=True)

labels2018=DF2018['NO2']
templ2018=set_tags_no2(labels2018)
DF2018.NO2=templ2018


Ys=DF2018["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

B418=Ys_of_train

color=['green','orange','blue','purple','red']
labels=['Good','Moderate','Unhealthy for sensitive group','Unhealthy'
                                    ,'Very unhealthy']
fig, ax = plt.subplots(1,1, figsize=(18,10))

hist, bins = np.histogram(B418,bins=5,density=True)
hist=hist/hist.sum()
for w,x,y,z in zip(np.ceil(bins[:-1]), hist.astype(np.float32)*100,
                                    color, labels):
    ax.bar(w,x, color = y, width=(bins[1]-bins[0]), label = z)
ax.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5,fontsize=18)
ax.set_ylim(ymin=0,ymax=40)
ax.set_title('Before Intervention',fontsize=28)
ax.set_ylabel('NO2 Percentage',fontsize=28)

fig, ax = plt.subplots(1,1, figsize=(18,10))

hist, bins = np.histogram(Af18,bins=5,density=True)
hist=hist/hist.sum()
```

```
for w,x,y,z in zip(np.ceil(bins[:-1]), hist.astype(np.float32)*100,
                                    color, labels):
    ax.bar(w,x, color = y, width=(bins[1]-bins[0]), label = z)
ax.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5,fontsize=10)
ax.set_ylim(ymin=0,ymax=40)
ax.set_title('After Intervention',fontsize=28)
ax.set_ylabel('NO2 Percentage',fontsize=28)


xaxis = [3,9,16,23,30]
xaxis_dec = [x-0.5 for x in xaxis]
xaxis_inc = [x+0.5 for x in xaxis]
color=['green','orange','blue','purple','red']
labels=['Good','Moderate','Unhealthy for sensitive group','Unhealthy'
                                    ,'Very unhealthy']



hist, bins = np.histogram(B418,bins=5,density=True)
hist=hist/hist.sum()
hist2, bins2 = np.histogram(Af18,bins=5,density=True)
hist2=hist2/hist2.sum()



rec0 = [hist[0].astype(np.float32)*100,hist[1].astype(np.float32)*100
                                    ,hist[2].astype(np.float32)*100,
                                    hist[3].astype(np.float32)*100,
                                    hist[4].astype(np.float32)*100]
rec1 = [hist2[0].astype(np.float32)*100,hist2[1].astype(np.float32)*
                                    100,hist2[2].astype(np.float32)*
                                    100,hist2[3].astype(np.float32)*
                                    100,hist2[4].astype(np.float32)*
                                    100]

ax = plt.subplot(111)
ax.bar(xaxis_dec, rec0, width=0.5,align='center', label = 'Before
                                    policy', color= 'firebrick')
ax.bar(xaxis_inc, rec1, width=0.5,align='center', label = 'After
                                    policy', color= 'darkcyan')

ax.set_ylim((0, 40))
plt.xticks(xaxis, labels, fontsize=18)
```

```
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
          fancybox=True, shadow=True, ncol=5,fontsize=30)
plt.ylabel('NO2 Percentage', fontsize=40,fontweight='bold')
plt.title("Before and After Policy", fontsize=40, fontweight='bold')
plt.show()

count, bins_count = np.histogram(B418, bins=10)
pdf = count / sum(count)
cdf = np.cumsum(pdf)
plt.plot(bins_count[1:], cdf, label="CDF Before policy",color= '
                                    firebrick',linewidth=10)
plt.legend(fontsize=30)
plt.xlabel('Bins count',fontsize=40, fontweight='bold')
plt.ylabel('CDF',fontsize=40, fontweight='bold')
plt.title('CDF by Policy',fontsize=40, fontweight='bold')

count, bins_count = np.histogram(Af18, bins=10)
pdf = count / sum(count)
cdf = np.cumsum(pdf)
plt.plot(bins_count[1:], cdf, label="CDF After policy",color= '
                                    darkcyan',linewidth=10)
plt.ylim(ymax=1.0)
plt.xlim(xmin=0)
plt.legend(fontsize=40)

plt.show()

colors = ['purple']

data=B418

bp=plt.boxplot(data, patch_artist = True,positions=[2],
                notch ='True', vert = 0,widths=0.4)

for patch, color in zip(bp['boxes'], colors):
    patch.set_facecolor(color)

for whisker in bp['whiskers']:
    whisker.set(color ='#8B008B',
                linewidth = 1.5,
                linestyle =":")

for cap in bp['caps']:
    cap.set(color ='#8B008B',
```

```
            linewidth = 2)

for median in bp['medians']:
    median.set(color ='red',
               linewidth = 3)

for flier in bp['fliers']:
    flier.set(marker ='D',
              color ='#e7298a',
              alpha = 0.5)

plt.legend()

colors = ['red']
data=Af18

bp=plt.boxplot(data, patch_artist = True,positions=[1],
                notch ='True', vert = 0,widths=0.4)

for patch, color in zip(bp['boxes'], colors):
    patch.set_facecolor(color)

for whisker in bp['whiskers']:
    whisker.set(color ='#8B008B',
                linewidth = 1.5,
                linestyle =":")

for cap in bp['caps']:
    cap.set(color ='#8B008B',
            linewidth = 2)

for median in bp['medians']:
    median.set(color ='red',
               linewidth = 3)

for flier in bp['fliers']:
    flier.set(marker ='D',
              color ='#e7298a',
              alpha = 0.5)

plt.legend()
plt.ylabel('boxes')
plt.title('box distrobution of data')
plt.show()
```

```
scaler = MinMaxScaler(feature_range=(0, 4))
yhat18 = scaler.fit_transform(yhat18)
yhat18af = scaler.fit_transform(yhat18af)
plt.plot(yhat18, label='Prediction by days before policy', color= '
                                    firebrick',linewidth=3)
plt.plot(yhat18af, label='Prediction by days after policy', color= '
                                    darkcyan',linewidth=3)
plt.ylabel('NO2 labels predictions')
plt.yticks([0,1,2,3,4])
plt.xlabel('Serialised time and data', fontweight='bold')
plt.title('Prediction by days', fontweight='bold')
plt.ylim(ymin=0, ymax=4)
plt.xlim(xmin=0)
plt.legend()
plt.show()

from IPython.core.pylabtools import figsize
DF18b=pd.read_csv('mix2018.csv')
monthvals=DF18b.month.unique()


tf=DF18b['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF18b['two wheeled motor vehicles']=twmv
DF18b['cars and taxis']=cat
DF18b['buses and coaches']=bac
DF18b['lgvs']=lgvs
DF18b.drop('Traffic Flow',axis=1)
labels2018=DF18b['NO2']
templ2018=set_tags_no2(labels2018)
DF18b.NO2=templ2018
Ys=DF18b["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
```

```
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

DF18b.NO2=Ys_of_train

#after
DF18a=pd.read_csv('mix2018.csv')
monthvals=DF18a.month.unique()

tf=DF18a['Traffic Flow'].values
twmv=np.round(tf* 0.0063)
cat=np.round(tf*0.8)
bac=np.round(tf*0.0124)
lgvs=np.round(tf*0.1813)
DF18a['two wheeled motor vehicles']=twmv
DF18a['cars and taxis']=cat
DF18a['buses and coaches']=bac
DF18a['lgvs']=lgvs

# 18
twmv18=DF18a["two wheeled motor vehicles"].values
twmv18=np.round(twmv18*0.8)
cat18=DF18a["cars and taxis"].values
cat18=np.round(cat18*0.8)
bac18=DF18a["buses and coaches"].values
bac18=np.round(bac18*0.9)
lgvs18=DF18a["lgvs"].values
lgvs18=np.round(lgvs18*0.8)
co18=DF18a["CO"].values
co18=co18-18
pm2518=DF18a["PM2_5"].values
pm2518=pm2518 -16
par18=DF18a["Particle Count"].values
par18=par18 -10
pm118=DF18a["PM1"].values
pm118=pm118 -10
pm1018=DF18a["PM10"].values
pm1018=pm1018 -21
pm418=DF18a["PM 4"].values
pm418=pm418 -23
o318=DF18a["O3"].values
o318=o318 -29
```

```
no18=DF18a["NO"].values
no18=no18 -18
nox18=DF18a["NOx"].values
nox18=nox18 -24
no218=DF18a["NO2"].values
no218=no218 -45


DF18a["two wheeled motor vehicles"]=twmv18
DF18a["cars and taxis"]=cat18
DF18a["buses and coaches"]=bac18
DF18a["lgvs"]=lgvs18
DF18a["CO"]=co18
DF18a["PM2_5"]=pm2518
DF18a["Particle Count"]=par18
DF18a["PM1"]=pm118
DF18a["PM10"]=pm1018
DF18a["PM 4"]=pm418
DF18a["O3"]=o318
DF18a["NO"]=no18
DF18a["NOx"]=nox18
DF18a["NO2"]=no218

labels2018=DF18a['NO2']
templ2018=set_tags_no2(labels2018)
DF18a.NO2=templ2018
Ys=DF18a["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

DF18a.NO2=Ys_of_train

print(monthvals)
for j in monthvals:
```

```
#before
tempdf=DF18b.loc[DF2018['month'] == j]
trainx18=tempdf.drop('NO2', axis=1)

scaler = MinMaxScaler(feature_range=(0, 1))
values18 = scaler.fit_transform(trainx18)
trainx18 = series_to_supervised(values18, 1, 1)
trainlab18=tempdf['NO2'].values

trainx18=trainx18.values
trainlab18=trainlab18[:48]
trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1
                                    ]))
trainx18 =trainx18[:,:,:48]
yhats = model18.predict(trainx18)
scaler = MinMaxScaler(feature_range=(0, 4))
yhats = scaler.fit_transform(yhats)

#after
tempdf=DF18a.loc[DF2018['month'] == j]
trainx18=tempdf.drop('NO2', axis=1)

scaler = MinMaxScaler(feature_range=(0, 1))
values18 = scaler.fit_transform(trainx18)
trainx18 = series_to_supervised(values18, 1, 1)
trainlab18=tempdf['NO2'].values

trainx18=trainx18.values
trainlab18=trainlab18[:52]
trainx18 = trainx18.reshape((trainx18.shape[0], 1, trainx18.shape[1
                                    ]))
trainx18 =trainx18[:,:,:52]
yhats2 = model18a.predict(trainx18)
scaler = MinMaxScaler(feature_range=(0, 2))
yhats2 = scaler.fit_transform(yhats2)
plt.plot(yhats, label='Prediction before policy',color= 'firebrick'
                                    ,linewidth=5)
plt.plot(yhats2, label='Prediction after policy',color= 'darkcyan',
                                    linewidth=5)
plt.ylabel('NO2 Concentration Class',fontsize=60,fontweight='bold')
plt.yticks([0,1,2,3,4])
plt.xlabel('Time Samples',fontsize=60,fontweight='bold')
plt.title('Predictions for the ' + str(j)+ 'th'+ ' Month ',fontsize
                                    =60,fontweight='bold')
```

```
  plt.ylim(ymin=0,ymax=4)
  plt.xlim(xmin=0)
  plt.legend(fontsize=60)
  plt.show()

DF2018=pd.read_csv('mix2018.csv')
DF2018.reset_index(drop=True, inplace=True)
labels2018=DF2018['NO2']
templ2018=set_tags_no2(labels2018)
DF2018.NO2=templ2018
Ys=DF2018["NO2"].values
Ys_of_train=np.zeros_like(Ys)
for i in range(len(Ys)):
  if(Ys[i]=='Good'):
    Ys_of_train[i]=0
  elif(Ys[i]=='Moderate'):
    Ys_of_train[i]=1
  elif(Ys[i]=='Unhealthy for sensitive group'):
    Ys_of_train[i]=2
  elif(Ys[i]=='Unhealthy'):
    Ys_of_train[i]=3
  elif(Ys[i]=='Very unhealthy'):
    Ys_of_train[i]=4

B418=Ys_of_train
DF18=DF2018
DF18.NO2=B418

fig, ax = plt.subplots()
DF2018['month'].value_counts().plot(ax=ax, kind='bar', color=['brown'
                                    , 'blue', 'purple', 'green', '
                                    orange', 'yellow', 'red'], figsize
                                    =(25,25), fontsize=(25))

plt.ylabel('Value Counts',fontsize=(30))
plt.xlabel('Months',fontsize=(30))
plt.xlim(xmin=0)
plt.show()

DF2018['NO2'].value_counts()

DF18.columns

labels, counts = np.unique(DF18.month.values, return_counts=True)
```

```
plt.bar(labels, counts, align='center',color="green")
plt.xticks(labels)
plt.ylabel("count")
plt.xlabel("month")
plt.xlim(xmin=0)
plt.show()

labels, counts = np.unique(DF18.day.values, return_counts=True)
plt.bar(labels, counts, align='center',color="green")
plt.xticks(labels)
plt.ylabel("count")
plt.xlabel("day")
plt.xlim(xmin=0)
plt.show()

labels, counts = np.unique(DF18.hour.values, return_counts=True)
plt.bar(labels, counts, align='center',color="green")
plt.xticks(labels)
plt.ylabel("count")
plt.xlabel("hour")
plt.xlim(xmin=0)
plt.show()

DF2018.hist(figsize=(26,26), layout=(6,5), color = 'green')
plt.xticks(fontsize=10)
plt.xlim(xmin=0)
plt.subplots_adjust(left=4, bottom=4, right=5, top=5)
plt.show()
```

Appendix D

Python Code for the Results Reported in Chapter 6

The results reported in Chapter 6 are computed by developing an extensive code in the Python programming environment. The code is reported below.

```
import os
import random
import time
import pandas as pd
import numpy as np
import math as ma

from tqdm import *
from scipy import spatial
from operator import itemgetter, attrgetter
from collections import Counter
from copy import deepcopy


'''
Initialize LSTM-GA
'''
class LSTM_GA():
    def __init__(self, type= 6,
                 url = str(),
                 ev_quantity = int(),
                 generation_num = int(),
                 constant_ev_quantity = True,
                 output_dir = str(),
```

```
            worst_case_eval = False
            ):

        self.type  = type
        self.type_list = [[7, 2], [11, 2], [22, 3], [50, 4], [100, 5]
                                        , [150, 5], [200, 6]]    #
                                        available charger types [
                                        Charging station capacity,
                                         Cost share]
        self.baseMoney = 550  # Basement of  Cost share
        self.dataset = self.load_dataset(url)
        self.boundary = [[54.9641, -1.76835] , [55.059, -1.53996]]  #
                                         Using the coordinates of
                                        the top-left and bottom-
                                        right points to define
                                        basic rectangular regions
        self.ev_quantity = ev_quantity
        self.generation_num = generation_num
        self.population = []  #  Population pool of each generation
        self.hunt_count = []  #  Prey captured in each generation:
                                        Fully charged electric
                                        cars
        self.satisfy_demand = []  # Satisfaction rate in each
                                        generation
        self.elits = []*6
        self.worst_case_eval = worst_case_eval
        self.not_satisfied = []  #  Uncaught prey in each generation:
                                         Uncharged electric cars
        self.ev_food_list = []   #  Food list in each generation:
                                        List of electric cars
        if self.type == 6:
            self.type_use = self.type_list[:-1]
        else:
            self.type_use = self.type_list[:]
        self.sum_charger_num = 0  #  Counting the number of charging
                                        stations
        self.avg_work_time = 0    #  Calculating the working time of
                                        charging stations
        self.sum_cost = []        #  Total cost in each generation  =
                                         Basement of  Cost share x
                                         Cost share x  Number
        self.type_count = []      #  Total type of chargers for each
                                        generation
        self.constant_ev_quantity = constant_ev_quantity
```

```
        self.utilization_rate = []

    def load_dataset(self,url,sheet_name=0):
        if url.find(".xlsx"):
            return pd.read_excel(url, sheet_name=sheet_name)
        elif url.find(".csv"):
            return pd.read_csv(url)

    def select_items(self):
        try:
            self.latitude = self.dataset['LSOA/DZ centre point
                                          latitude']
            self.longitude = self.dataset['LSOA/DZ centre point
                                          longitude']
            self.EV_power_demand = self.dataset['Total EV power
                                          demand']
            self.vehicles_exit_prob = self.dataset['vehicles
                                          percentage']
        except Exception:
            raise("The data columns are not suitable, please check
                                          columns")

    def cos_sim(self, vec1, vec2):
        return 1 - spatial.distance.cosine(vec1, vec2)

    def crossover(self, geneset3, geneset4, base_prob = 0.2):
        geneset1 = deepcopy(geneset3)
        geneset2 = deepcopy(geneset4)
        Crossover_prob = random.random()
        if Crossover_prob > base_prob:
            l = random.randrange(0, 3, 1)
            temp = geneset1[l]
            geneset1[l] = geneset2[l]
            geneset2[l] = temp
        mutation_prob_1 = random.random()
        if mutation_prob_1 > 0.5 :
            geneset1 = self.mutation(geneset1)
        mutation_prob_2 = random.random()
        if mutation_prob_2 > 0.5 :
            geneset2 = self.mutation(geneset2)
        return geneset1, geneset2

    def mutation(self,geneset):
        prob_lati = random.random()
```

```
        if prob_lati > 0.5:
            mut_lati = geneset[0] * (1 + 0.045 * (np.random.normal(
                                                  loc=0.0, scale=1.0,
                                                  size=None)))
        else:
            mut_lati = geneset[0]
        prob_long = random.random()
        if prob_long > 0.5:
            mut_long = geneset[1] * (1 + 0.058 * (np.random.normal(
                                                  loc=0.0, scale=1.0,
                                                  size=None)))
        else:
            mut_long = geneset[1]
        prob_type = random.random()
        if prob_type > 0.5:
            mut_type = random.sample(self.type_use, 1)
            gene_type = mut_type[0][0]
            gene_cost = mut_type[0][1]
        else:
            gene_type = geneset[2]
            gene_cost = geneset[3]
        return [mut_lati, mut_long, gene_type, gene_cost, 0, 0, 0]

    def initial_population(self):
        start_time = time.time()
        get_food_score = 0
        worktime_score = 0
        generation = 0
        pdbr = tqdm(range(0, self.latitude.size),leave = False, desc
                                            = "initializing")
        for i in pdbr:
            type_sample = random.sample(self.type_use, 1)

            self.population.append([self.latitude[i],self.longitude[i
                                                  ],type_sample[0][0],
                                                  type_sample[0][1],
                                                  get_food_score,
                                                  worktime_score,
                                                  generation])

        print(f'finish initial using time {(time.time()-start_time)}'
                                                )
        return self.population
```

```
    def generate_food(self,last_ev_food_count):
        # Clearing and regenerating
        self.ev_food_list = []
        for areas in range(0, len(self.vehicles_exit_prob)):
            if random.random() > 0.5 :
                areas_ev_quantity = ma.ceil(self.vehicles_exit_prob[
                                                areas] *
                                                last_ev_food_count
                                                )
            else:
                areas_ev_quantity = ma.floor(self.vehicles_exit_prob[
                                                areas] *
                                                last_ev_food_count
                                                )
            assert areas_ev_quantity != 0
            for ev in range(areas_ev_quantity):

                ev_lati = self.latitude[areas] * (1 + 0.090 * np.
                                                random.normal(loc=
                                                0.0, scale=1.0,
                                                size=None))
                ev_long = self.longitude[areas] * (1 + 0.116 * np.
                                                random.normal(loc=
                                                0.0, scale=1.0,
                                                size=None))

                if self.worst_case_eval:
                    ev_demand=self.EV_power_demand[areas]/
                                                        areas_ev_quantity
                                                        *(1 + np.
                                                        random.normal(
                                                        loc=0.0, scale
                                                        =1.0, size=
                                                        None))
                else:
                    ev_demand=self.EV_power_demand[areas]/
                                                        areas_ev_quantity
                                                         * (2 - random
                                                        .random())

                while(ev_demand > 240 or ev_demand <= 0):
                    ev_demand = self.EV_power_demand[areas] * (1 + np
                                                        .random.normal
                                                        (loc=0.0,
```

```
                                                    scale=1.0,
                                                    size=None))

                self.ev_food_list.append([ev_lati, ev_long, ev_demand
                                                    ])
        return self.ev_food_list

    def hunt(self):
        hunt_food_num = 0
        for i in range(0, len(self.ev_food_list)):
            distance = []
            for j in range(0, len(self.population)):
                distance.append([np.power((np.abs(self.population[j][
                                                    0] - self.
                                                    ev_food_list[i][0]
                                                    ) + np.abs(self.
                                                    population[j][1] -
                                                     self.ev_food_list
                                                    [i][1])),2),j])
            distance.sort()
            for k in range(0, len(distance)):
                work_hour_score = self.population[distance[k][1]][5]
                                                    + \ self.
                                                    ev_food_list[i][2]
                                                    /(self.population[
                                                    distance[k][1]][2]
                                                    )
                if((work_hour_score) < 20 * 0.8):
                    self.population[distance[k][1]][5] += self.
                                                        ev_food_list[i
                                                        ][2] / self.
                                                        population[
                                                        distance[k][1]
                                                        ][2]
                    self.population[distance[k][1]][4] += 1
                    hunt_food_num += 1
                    break
        self.hunt_count.append(hunt_food_num)
        self.satisfy_demand.append(hunt_food_num/self.ev_quantity)

    def check_list(self, generation_now:int, is_offspring=False):
        data_set = []
        for charger in range(len(self.population)):
            for charger_other in range(len(self.population)):
```

```
                if charger_other == charger:
                    break
                if self.cos_sim(self.population[charger], self.
                                                    population[
                                                    charger_other]) ==
                                                     1.:
                    self.population[charger_other][0] = 0
            is_in_boundry_x = self.boundary[0][0] < self.population[
                                                charger][0] < self.
                                                boundary[1][0]
            is_in_boundry_y = self.boundary[0][1] < self.population[
                                                charger][1] < self.
                                                boundary[1][1]
            if is_in_boundry_x and is_in_boundry_y:
                if (generation_now == self.generation_num):
                    if self.population[charger][4] > 0 and self.
                                                    population[
                                                    charger][5] >
                                                    0:
                        data_set.append(self.population[charger])
                elif is_offspring:
                    self.population[charger][4] = 0
                    self.population[charger][5] = 0
                    data_set.append(self.population[charger])
                else:
                    if self.population[charger][4] > 0 and self.
                                                    population[
                                                    charger][5] >
                                                    0:
                        self.population[charger][4] = 0
                        self.population[charger][5] = 0
                        data_set.append(self.population[charger])
        self.population = data_set

    def get_utilization(self):
        w1 = 0.6
        w2 = 0.4
        Tm = 10
        self.utilization_rate.append(self.satisfy_demand[-1] * w1 + (
                                            1-((self.avg_work_time -
                                            Tm) / Tm) ** 2) * w2)

    def get_offsprings(self, generation_now:int):
        offspring = []
```

```
        for single in self.population:
            other = random.sample(self.population, 1)
            other = other[0]
            if single[6] < generation_now - 4:
                    break

            elif single[4] >= self.hunt_food_score[2] and single[5] >
                                                = self.work_hour_rank[
                                                2]:
                        self.elits.append(single)
                        geneset1, geneset2 = self.crossover(single,
                                                            other)
                        geneset1[6] = generation_now
                        geneset2[6] = generation_now
                        offspring.append(geneset1)
                        offspring.append(geneset2)

            elif single[4] >= self.hunt_food_score[1] or single[5] >=
                                                 self.work_hour_rank[1
                                                ]:
                        geneset1, geneset2 = self.crossover(single,
                                                            other)
                        geneset1[6] = generation_now
                        geneset2[6] = generation_now
                        survivor_prob = random.random()
                        if survivor_prob > 0.5:
                            offspring.append(geneset1)
                        offspring.append(geneset2)

            elif single[4] >= self.hunt_food_score[0] or single[5] >=
                                                 self.work_hour_rank[0
                                                ]:
                        geneset1, geneset2 = self.crossover(single,
                                                            other)
                        geneset1[6] = generation_now
                        geneset2[6] = generation_now
                        survivor_prob = random.random()
                        if survivor_prob > 0.4:
                            survivor_prob = random.random()
                            if survivor_prob > 0.5:
                                offspring.append(geneset1)
                            else:
                                offspring.append(geneset2)
            else:
```

```
                single[0] = 0
        self.population += offspring
        self.check_list(generation_now=generation_now, is_offspring=
                                            True)

    def get_rank(self,generation_now):

        self.population.sort(key=itemgetter(4), reverse=True)
        hunt_food_score_limit_top = self.population[0][4]
        hunt_food_score_limit_1 = hunt_food_score_limit_top * 0.2
        hunt_food_score_limit_2 = hunt_food_score_limit_top * 0.5
        hunt_food_score_limit_3 = hunt_food_score_limit_top * 0.9

        self.population.sort(key=itemgetter(5),reverse=True)
        work_hour_score_limit_top = self.population[0][5]
        work_hour_score_limit_1 = work_hour_score_limit_top * 0.2
        work_hour_score_limit_2 = work_hour_score_limit_top * 0.5
        work_hour_score_limit_3 = work_hour_score_limit_top * 0.9

        if self.elits != []:
            temp = []

            for elit in self.elits:
                if elit[6]>= generation_now - 5:
                    temp.append(elit)
                elif random.random() > 0.5:
                    temp.append(elit)
            self.elits = temp

            elits_mean = np.array(self.elits).mean(axis=0)
            avg_time = elits_mean[5]
            self.avg_work_time = avg_time
            avg_hunt = elits_mean[4]

            hunt_food_score_limit_1 = 0.8 * hunt_food_score_limit_1 +
                                            0.2 * avg_hunt
            hunt_food_score_limit_2 = 0.9 * hunt_food_score_limit_2 +
                                            0.1 * avg_hunt
            hunt_food_score_limit_3 = 0.99 * hunt_food_score_limit_3
                                          + 0.01 * avg_hunt

            work_hour_score_limit_1 = 0.8 * work_hour_score_limit_1 +
                                            0.2 * avg_time
```

```
            work_hour_score_limit_2 = 0.9 * work_hour_score_limit_2 +
                                                0.1 * avg_time
            work_hour_score_limit_3 = 0.99 * work_hour_score_limit_3
                                              + 0.01 * avg_time

        self.work_hour_rank = [self.satisfy_demand[-1] * item for
                                              item in [
                                              work_hour_score_limit_1,
                                              work_hour_score_limit_2,
                                              work_hour_score_limit_3]]
        self.hunt_food_score = [self.satisfy_demand[-1] * item for
                                              item in[
                                              hunt_food_score_limit_1,
                                              hunt_food_score_limit_2,
                                              hunt_food_score_limit_3]]


    def write_data(self):
        try:
            self.sum_type = pd.DataFrame(data=self.
                                                train_total_population
                                                ,
                                            columns=['Chargers Number'])
            self.sum_cost = pd.DataFrame(data=self.sum_cost,
                                            columns=['Toral Cost'])
            self.ev_quantity = pd.DataFrame(data=self.
                                                train_total_evfood,
                                                columns=['Total EV
                                                Quantity'])
            self.hunt_count = pd.DataFrame(data=self.hunt_count,
                                                columns=['Satisfied EV
                                                 Quantity'])
            self.satisfy_demand = pd.DataFrame(data=self.
                                                satisfy_demand,
                                                columns=['Satisfied
                                                Percentage'])
            self.utilization_rate = pd.DataFrame(data=self.
                                                utilization_rate,
                                                columns=['Utilization'
                                                ])
            df_concat = pd.concat([self.prpotion, self.sum_type, self
                                                .sum_cost, self.
                                                ev_quantity, \ self.
                                                hunt_count, self.
```

```
                                                satisfy_demand, self.
                                                utilization_rate ],
                                                join='inner', axis=1)
            writer = pd.ExcelWriter(output_dir)
            self.population_details.to_excel( writer, sheet_name="
                                                population_details",
                                                index=False,)
            df_concat.to_excel(
                                writer,
                                sheet_name="train_details",
                                index=False,
                                 )
            writer.save()

        except:
            raise ("Data Output is wrong, please check it again")

    def count_type(self):
        self.population_details = pd.DataFrame(data=self.population,
                                            columns=['Latitude', '
                                            Longitude', 'Type of
                                            charging pile (kW)', '
                                            Economic costs', 'EV
                                            Number of Charging Posts
                                            Serviced', 'Charging Posts
                                             Operating Hours', '
                                            generation'])
        self.type_count.append(list((Counter(self.population_details[
                                            'Type of charging pile (kW
                                            )'])).values()))
        if len(self.type_use) == 6:
            self.prpotion = pd.DataFrame(data=self.type_count,
                                                columns=['7kw', '11kw'
                                                , '22kw', '50kw', '
                                                100kw', '150kw'])
        else:
            self.prpotion = pd.DataFrame(data=self.type_count,
                                                columns=['7kw', '11kw'
                                                , '22kw', '50kw', '
                                                100kw', '150kw', '
                                                200kw'])
        self.sum_cost.append(sum(self.population_details['Economic
                                            costs']))
```

```
    def train(self):
        self.train_total_population = []
        self.train_total_evfood = []
        self.select_items()
        self.population = self.initial_population()
        pdbr = tqdm(range(0, self.generation_num + 1), leave = False,
                                              desc = "training")
        strat_time = time.time()
        self.generate_food(self.ev_quantity)
        for generation_now in pdbr:
            self.hunt()
            is_last_generation = generation_now == self.
                                                  generation_num
            if is_last_generation:
                self.check_list(generation_now)
                self.train_total_population.append(len(self.
                                                    population))
                self.train_total_evfood.append(len(self.ev_food_list)
                                                    )
                self.count_type()
                self.get_utilization()
                self.write_data()
            else:
                self.get_rank(generation_now)
                self.get_offsprings(generation_now)
                self.train_total_population.append(len(self.
                                                    population))
                self.train_total_evfood.append(len(self.ev_food_list)
                                                    )
                self.count_type()
                self.get_utilization()
                if self.constant_ev_quantity:
                    self.generate_food(self.ev_quantity)
                else:
                    self.generate_food(self.train_total_evfood[-1])


input_dir = r"./Train_dataset"

year_num = {
    '2042': 134606,
    '2046': 145345,
    '2048': 146617,
    '2050': 160403,
```

```
}

if os.path.isdir(input_dir):
    files = os.listdir(input_dir)

    for file in files:
        url = os.path.join(input_dir, file)
        file_name = file.split(".xlsx")[0]
        output_dir = f'./result_{file_name}.xlsx'
        year = file_name.split("_")[0]
        ev_quantity = year_num[file_name.split("_")[1]]
        print(f'start training {year}, total EV quantity is {
                                            ev_quantity}')
        lstm_ga = LSTM_GA(
            url=url,
            ev_quantity=ev_quantity,
            generation_num=100,
            output_dir = output_dir,
        )
        lstm_ga.train()
    print(output_dir," finished train")
else:
    print("Please check input dir")
```