Exploring Mobility Data for Meteorological Applications

Mohamed Imthiyas Abdul Rasheeth
Student ID: 201771669

Supervised by Dr. Nadhir Ben Rached (University of Leeds), Dr. Faye Wyatt (UK Met Office) & Dr. Joanne Robbins (UK Met Office)

Submitted in accordance with the requirements for the module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

September 2024

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

UNIVERSITY OF LEEDS

School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name Mohamed Imthiyas Abdul Rasheeth

Student ID 201771669

Acknowledgements

I am deeply grateful to my supervisors, Dr. Nadhir Ben Rached from the University of Leeds, Dr. Faye Wyatt and Dr. Joanne Robbins from the UK Met Office, whose continuous support, guidance, and encouragement have been instrumental throughout my research and the writing of this dissertation. Their insightful feedback and wise counsel have been invaluable, helping to shape this work into what it is today.

A special note of thanks goes to Dr. Faye Wyatt from the UK Met Office. Her dedication, thoughtful involvement, and genuine passion for my research have been a source of great inspiration. Dr. Wyatt's invaluable suggestions and personal commitment to my progress have significantly elevated the quality of this dissertation. I am profoundly thankful for the way she has gone above and beyond to ensure that I had the support and resources I needed to succeed.

I also want to extend my heartfelt appreciation to Belén González Guío and the entire team at Vodafone. Their unwavering support with the mobility data and their willingness to assist at every step made a substantial difference in the progress of my work.

Finally, I am deeply indebted to my parents, Abdul Rasheeth and Shameem Begum, whose unwavering support and love have been the cornerstone of my academic journey. My sincere thanks extend to Hari Loganathan for his steadfast encouragement, which has been a constant source of motivation. I also wish to acknowledge my family — Ghousiya, Zaim, and Salman — whose patience and understanding have been invaluable. Lastly, I am thankful to my friends for their continued encouragement and companionship. This work would not have been possible without the collective support of these remarkable individuals.

Abstract

As climate change accelerates, the need for timely and accurate weather warnings becomes increasingly critical. This dissertation investigates how different weather conditions impact human mobility patterns across the United Kingdom, aiming to enhance early warning systems and disaster preparedness strategies. By integrating mobility data with meteorological information, this research bridges the gap between weather forecasts and real-world human behavior, providing actionable insights that can improve public safety, urban planning, and emergency response efforts.

The study employs both statistical models, such as regression analysis, and machine learning techniques, including Random Forest and Gradient Boosting, to analyze the influence of weather on mobility. The findings reveal that weather variables such as temperature, precipitation, and solar radiation significantly affect mobility patterns, with notable differences observed between weekdays and weekends. For instance, adverse weather conditions like wet or uncomfortable days have a more pronounced impact on mobility during weekends compared to weekdays.

This research also explores regional differences in how weather impacts mobility, demonstrating that more localized analysis can improve model accuracy. Additionally, the study addresses the challenges of data integration, geographic biases, and the limitations of the models used, providing suggestions for future research. These insights contribute to a deeper understanding of the interplay between weather and human movement, with implications for urban planning, infrastructure resilience, and public safety strategies in the face of adverse weather conditions.

Contents

List of Figures

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

In an era where climate change is accelerating, the necessity for timely and effective weather warnings has never been more vital. Accurate early warnings are pivotal in safeguarding lives, infrastructure, and economies, particularly as extreme weather events become more frequent and intense. Global initiatives, such as the **Sendai Framework for Disaster Risk Reduction** (2015–2030) and the United Nations' **Early Warnings for All initiative**, emphasizes the need for comprehensive disaster preparedness. These frameworks aim to reduce the risks and impacts of disasters by ensuring that vulnerable populations receive accurate and actionable warnings that they can respond to before disasters strike (Aitsi-Selmi et al., 2015) (Tupper et al., 2023).

The Sendai Framework, in particular, advocates for a shift from reactive responses to hazards toward a proactive approach that includes assessing vulnerabilities, building resilience, and reducing exposure. It emphasizes the integration of early warning systems with public health, urban planning, and disaster management, promoting stronger disaster risk governance and multi-hazard early warning systems that reach all segments of the population (Aitsi-Selmi et al., 2015). Complementing this, the Early Warnings for All initiative, launched in 2022, emphasizes the need to expand early warning coverage, aiming to ensure that every person is protected by early warning systems by 2027 (Tupper et al., 2023).

Despite these global efforts, a major gap remains in the translation of warnings into real-world action. The effectiveness of early warnings is often compromised by incomplete data on how individuals actually respond to warnings and navigate their environments under varying weather conditions. Mobility data presents a significant opportunity to bridge this gap. The patterns of human movement — whether people choose to move or stay in a location — are influenced by many factors, with weather being one of the most significant. Analyzing these patterns can offer valuable insights into how populations react to weather events, thereby enhancing situational awareness during critical periods (Aitsi-Selmi et al., 2015).

Mobility data provides real-time information on human behavior during weather events, offering an exceptional opportunity to improve traffic management, urban planning, and emergency response strategies. Understanding these patterns can refine early warning systems, making them more targeted, actionable, and reflective of actual behavior of the people in response to weather threats. The integration of mobility data with early warning systems presents a unique opportunity not only to respond more effectively to imminent threats but also to improve planning for future potential events (Tupper et al., 2023).

This dissertation explores how different weather conditions influence mobility patterns across the United Kingdom and investigates how this data can be leveraged to strengthen early warning systems. By gaining a deeper understanding of these patterns, the research aims to improve weather warning communication, making it more actionable and attuned to the specific needs of diverse communities (Tupper et al., 2023). Ultimately, this study seeks to find the significance of the weather variables in mobility.

1.2 Problem Statement

While it is widely recognized that weather conditions influence human behavior and mobility, this relationship has not been extensively quantified using empirical data (Harrison et al., 2021). Despite the availability of extensive datasets, there exists a substantial gap in understanding the specific impacts of meteorological variables — such as temperature, precipitation, etc,. — on mobility patterns. This gap is particularly concerning in areas such as emergency response, urban planning, and evacuation planning, where actionable insights into weather-related mobility shifts could greatly enhance data-driven decision-making.

In the context of urban planning, a deeper understanding of how adverse weather conditions affect mobility patterns can inform the design of resilient infrastructure. For example, if certain weather conditions are found to significantly reduce mobility, this knowledge could guide the development of adaptive urban designs that accommodate these changes, thereby minimizing disruptions to transportation networks and public services. In emergency management, precise mobility data can greatly improve evacuation protocols by providing real-time insights into population movements during hazardous weather events. This, in turn, can enhance the allocation of resources and improve public safety by directing response efforts to areas with heightened vulnerability (Merz et al., 2020).

Moreover, the lack of systematic studies in this area leaves critical gaps in the ability of cities and rural areas to prepare for and respond to weather-related disruptions. Impact forecasting, which integrates hazard data with exposure and vulnerability assessments, is an emerging field that holds significant promise for improving emergency responses and diminishing the effects of extreme weather. However, the success of these systems is heavily dependent on detailed and accurate mobility data to predict how populations will behave during such events (Merz et al., 2020).

By understanding the gap between meteorological data and real-time human mobility, this research aims to contribute critically to the development of more effective early warning systems and disaster preparedness strategies. Ultimately, this study aims to uncover how weather conditions influence mobility patterns, providing insights to inform future applications in transportation, infrastructure, and emergency management.

1.3 Research Objectives

The primary objective of this study is to analyze the relationship between weather conditions and mobility patterns by comparing relevant datasets. The research aims to address the following questions:

1. How do various meteorological conditions (e.g., precipitation, temperature, solar radiation) influence individual mobility patterns?
2. How do mobility patterns differ in response to weather conditions on weekdays compared to weekends, and what are the key distinctions?

3. How does the impact of weather on mobility patterns differ between the United Kingdom as a whole and specific selected regions within the country?

4. How effective is mobility data in improving situational awareness for weather warnings, and what are the strengths and limitations of using this data to enhance weather-related decision-making and emergency response strategies?

1.4 Structure of the Dissertation

The structure of this dissertation is organized as follows:

Chapter 2 provides a comprehensive review of the existing literature on the relationship between mobility and meteorological conditions, identifying key gaps and contributions from previous studies. Chapter 3 outlines the data sources and structure used in the research. Chapter 4 discusses the pre-processing methods and the framework for integrating mobility and meteorological datasets. Chapter 5 details the methodological approach, including statistical modeling and correlation analysis, explaining how the data was analyzed to reveal insights. Chapter 6 presents the key results of the analysis, highlighting significant correlations and patterns found in the data, while addresses the challenges and study's limitations. Chapter 7 discusses the implications of these findings, and suggests directions for future research.

Chapter 2

Literature Review

The integration of weather forecasting with mobility data represents a significant advancement in predicting and limiting the impacts of extreme weather events. This chapter delves into the foundations of impact-based forecasting (IbF) and explores the role of mobility data, alongside other data sources such as social media, in enhancing forecasting models for practical, real-world applications. To illustrate these concepts, a case study on Storm Eunice is presented, demonstrating how shifts in mobility patterns during extreme weather events can inform and improve future forecasting and planning strategies.

2.1 Overview of Impact-Based Forecasting & Mobility Data

Traditional weather forecasting, regularly provided by national meteorological agencies, offers essential updates on atmospheric conditions such as temperature, wind, and precipitation. While these forecasts have become increasingly accurate, they often fall short in predicting the societal impacts of weather events at a granular level. This limitation has led to the development of Impact-Based Forecasting (IbF), an approach that not only predicts weather phenomena but also anticipates their effects on people, infrastructure, and ecosystems. IbF extends beyond traditional meteorological forecasting by integrating data on vulnerability and exposure, thereby facilitating more informed decision-making in areas such as urban planning, emergency response, and public safety (World Meteorological Organization, 2022).

The evolution of IbF is vital for managing the complexities of modern urban environments. For example, the UK Met Office's National Severe Weather Warning Service (NSWWS) shifted from threshold-based warnings to impact-based warnings in 2011. These updated warnings focus on the potential consequences of weather events, enabling more targeted and actionable alerts for various stakeholders, including local authorities and the general public (Suri and Davies, 2021).

2.1.1 The Need for Impact-Based Forecasting

The critical value of IbF lies in its ability to anticipate not only the occurrence of weather events but also their specific effects on different communities. Traditional meteorological forecasts might accurately predict a storm, but without a nuanced understanding of the vulnerabilities within the affected area, effective planning and response remain challenging. For instance, during Tropical Cyclone Fitow in 2013, although the China Meteorological Administration issued accurate meteorological warnings, the cyclone still resulted in extensive flooding, damaging 97

roads, 900 communities, and parking facilities, with economic losses exceeding RMB 890 million. This significant damage occurred because the warnings lacked detailed guidance on the likely impacts, emphasizing the necessity of IbF in mitigating damage more effectively (World Meteorological Organization, 2022).

The growing demand for more accurate real-time data to predict these impacts has sparked interest in using mobility data from telecom providers, such as Vodafone. This data proves invaluable in understanding movement patterns during disasters, thereby aiding authorities in designing more effective evacuation strategies (Harrison et al., 2021).

By integrating exposure and vulnerability data, IbF offers a more comprehensive understanding of risk, enabling both the public and authorities to prepare more effectively for specific hazards. For example, during Storm Eunice, mobility data provided crucial insights into how people moved during the storm, allowing emergency services to prioritize resource allocation based on real-time information (Robbins et al., 2022).

2.1.2 Evolution of Impact-Based Forecasting

Weather warnings have evolved through several key stages, advancing from basic meteorological forecasts to sophisticated impact-based systems. Initially, forecasts provided general predictions of atmospheric conditions like wind, rain, and temperature, without considering societal impacts. This evolved into warnings based on fixed meteorological thresholds, where alerts were triggered when specific weather parameters exceeded predefined limits. The next stage involved developing warnings in consultation with local authorities, focusing on user-defined thresholds. Subsequently, warnings began incorporating spatial and temporal variations, allowing for more localized and accurate alerts. This led to multi-hazard impact-based forecasting, which considers the potential impacts on people, property, and the environment by combining meteorological data with exposure and vulnerability assessments. The most advanced stage involves impact forecast and warning services, offering detailed, tailored information and response guidance specific to individuals or groups. (World Meteorological Organization, 2022)

IbF systems, like those used in New Zealand, emphasize collaboration between various stakeholders, such as governmental agencies, private entities, and local communities. The development of these systems has allowed for improved communication of risks and enhanced emergency responses, particularly during extreme weather events like cyclones and floods (Harrison et al., 2022).

2.1.3 Fundamentals of Impact-Based Forecasting

IbF focuses on assessing the risks associated with weather events by considering the influence of hazard, exposure, and vulnerability. The risk of impact can be mathematically expressed as:

$$|\text{Risk of impact}(x,t)| \equiv |\text{hazard}(x,t)| \cup |\text{vulnerability}(x,t)| \cup |\text{exposure}(x,t)|^{1}$$

Where,

- **Hazard** – refers to any weather, geophysical, or human-induced threat to life, property, or the environment.
- **Exposure** – is the presence of people, infrastructure, or ecosystems in hazard-prone areas.
- **Vulnerability** – indicates how susceptible these exposed elements are to damage, influenced by factors like population density and socio-economic conditions (World Meteorological Organization, 2022).

[1] $\cup$ represents the union of the level of hydrometeorological forecast uncertainty, the degree of vulnerability, and the level of exposure.

Depending on the application, weather forecasts can be categorized into three distinct paradigms, as illustrated in Table 2.1 below.

Paradigms	**Weather forecasts and Warnings**	**Impact-Based Forecasts and Warnings**	**Impact Forecasts and Warnings**
Includes	Hazard only	Hazard and Vulnerability	Hazard, Vulnerability and Exposure
Definition	Provides information on meteorological variables and their expected changes	Describes the expected impacts of weather conditions	Offers detailed information down to individual or group level
Example	Bora winds are expected tonight with wind speeds of 20 metres per second.	Bora winds are expected tonight which may result in delays or cancellation to ferry services.	Ferry services for the island of Brac will most likely be cancelled tonight due to Bora winds.

Table 2.1: Comparison of forecasting paradigms (Source: (World Meteorological Organization, 2022))

To assist decision-makers and the general public in preparing effectively, IbF often utilizes a Risk Matrix. This matrix combines the likelihood of a hazardous event with its potential impact, creating a color-coded system that simplifies the communication of risk levels. The matrix facilitates a more nuanced approach, allowing authorities to take appropriate actions based on both the severity and probability of a weather event (World Meteorological Organization, 2022).

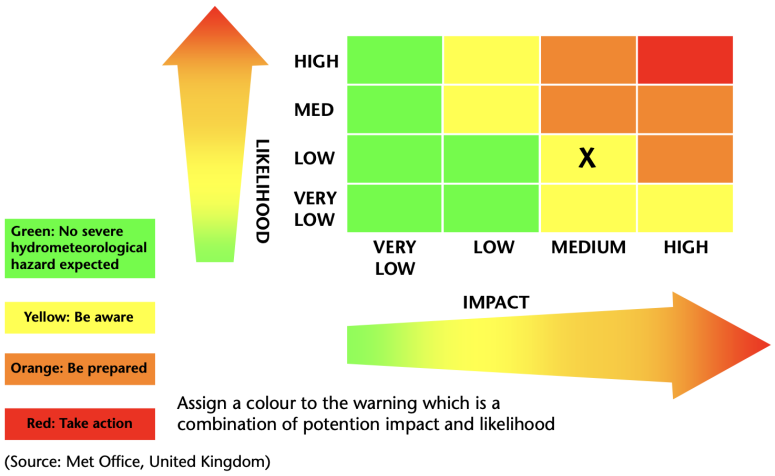


Figure 2.1: Risk matrix used for impact-based forecasting. (Source: (Met Office, 2024b))

In Figure 2.1, the matrix assigns specific colors to different combinations of impact and likelihood, each representing a different level of risk. For example:

- **Green:** No significant hydrometeorological hazard expected.
- **Yellow:** Be aware — minor impacts may occur, and preparation is advised.
- **Orange:** Be prepared — significant impacts are likely, and actions should be taken to mitigate risk.
- **Red:** Take action — severe impacts are highly likely, requiring immediate and decisive responses. (Met Office, 2024b)

This system enables both authorities and the public to respond to potential hazards more effectively, ensuring that appropriate measures are in place to reduce the impact of severe weather events.

Real-world applications of these paradigms are evident in tools like Google Maps or similar mobility and navigation services. For instance, during severe weather events, Google Maps could provide users with detailed information about potential road closures or transit delays based on impact forecasts. Similarly, emergency management agencies can use impact-based forecasts to identify areas that may require evacuation or additional resources, promoting more effective responses during events such as hurricanes or storms (World Meteorological Organization, 2023).

2.1.4 Incorporation of Mobility Data

Effective implementation of IbF relies heavily on collaboration between meteorological agencies and third-party organizations. Partnerships with technology companies and data providers are essential for gathering real-time mobility data, which provides valuable insights into how people respond to weather events. Mobility data from sources such as Bluetooth, Wi-Fi, and mobile apps like Google Maps helps identify disruptions in movement during storms, enabling authorities to prioritize resources and execute more effective emergency responses.

For example, during severe weather events, mobility data can pinpoint the most affected areas, facilitating quicker response efforts and more targeted communications. By integrating mobility data into IbF, authorities can create a more dynamic and responsive system for managing weather-related risks, ultimately enhancing decision-making and improving public safety outcomes (World Meteorological Organization, 2023).

2.1.5 Role of Tech Companies vs. Government

The increasing role of technology companies like Google, Vodafone, and Apple in providing weather warnings raises important considerations regarding public safety responsibilities. While these companies have the ability to quickly spread warnings through their platforms and data, they are not a substitute for government-issued alerts. Governments bear the responsibility of ensuring that all citizens, including those without access to smartphones or the internet, receive weather warnings.

Although tech companies offer valuable data and can amplify warnings, it remains the government's duty to protect citizens through inclusive and equitable systems that do not rely solely on digital platforms. In simpler terms, governments maintain the primary role in public safety, with tech companies serving as complementary partners (Leodolter and Rudloff, 2023).

2.1.6 Biases and Limitations

A significant challenge in IbF is the presence of biases in observational data. Social media data, for instance, often reflects the behaviors and communication patterns of urban, younger, and more affluent populations, potentially leading to skewed perceptions of weather impacts. For example, platforms like Twitter tend to highlight high-profile events, which can result in an overemphasis on dramatic impacts while under-representing more widespread, lower-profile effects (Spruce et al., 2020). This skew in data can distort forecasting models and result in less effective decision-making (Arthur and Williams, 2018).

Another crucial challenge in IbF is the accurate evaluation of forecasts, complicated by biases in impact observation data. The transition from traditional weather forecasting to IbF introduces complexities in measuring forecast effectiveness, as highlighted by (Wyatt et al., 2023). These biases can stem from various sources, including data

collection methods, geographic scope, and the severity of events, making it difficult to consistently evaluate and improve IbF models.

Similarly, mobility data can also be subject to geographic and socio-economic biases. Areas with limited telecommunications infrastructure or lower smartphone penetration may be under-represented, providing an incomplete picture of how populations move during weather events (Harrison et al., 2021). Despite these challenges, the integration of diverse data sources can help reduce biases and enhance the accuracy of IbF models. This includes efforts to balance urban and rural data, improve data collection methods, and ensure that vulnerable populations are accurately represented in forecasts. Furthermore, the need for better governance and standardization of data-sharing protocols across public and private entities is critical to improving the accuracy and reliability of impact forecasts (Harrison et al., 2022).

Additionally, privacy concerns are an important consideration when using mobility data in real-time applications. Ensuring compliance with data privacy regulations, such as the (General Data Protection Regulation (GDPR), 2016), is essential for maintaining public trust and upholding ethical forecasting practices (World Meteorological Organization, 2023). Finding the right balance between these privacy concerns and the valuable insights that mobility data provides is essential for advancing and improving IbF methods.

2.2 Case Study: Storm Eunice and the UK Met Office

2.2.1 Mobility Data and Weather Impacts

In 2023, (Robbins et al., 2023) explored the potential of mobility data in enhancing IbF and situational awareness at the UK Met Office. Their study identified three key areas of collaboration with Vodafone, focusing primarily on ”Mobility analytics”, which leverages the location data of Vodafone’s customers. With Vodafone covering approximately 20% of the UK mobile phone market in early 2023, this data offers a valuable opportunity to provide more personalized weather forecasts and warnings. The integration of mobility data with weather forecasting has emerged as a powerful tool in understanding how severe weather events, such as storms, impact mobility patterns and how populations react to these events.

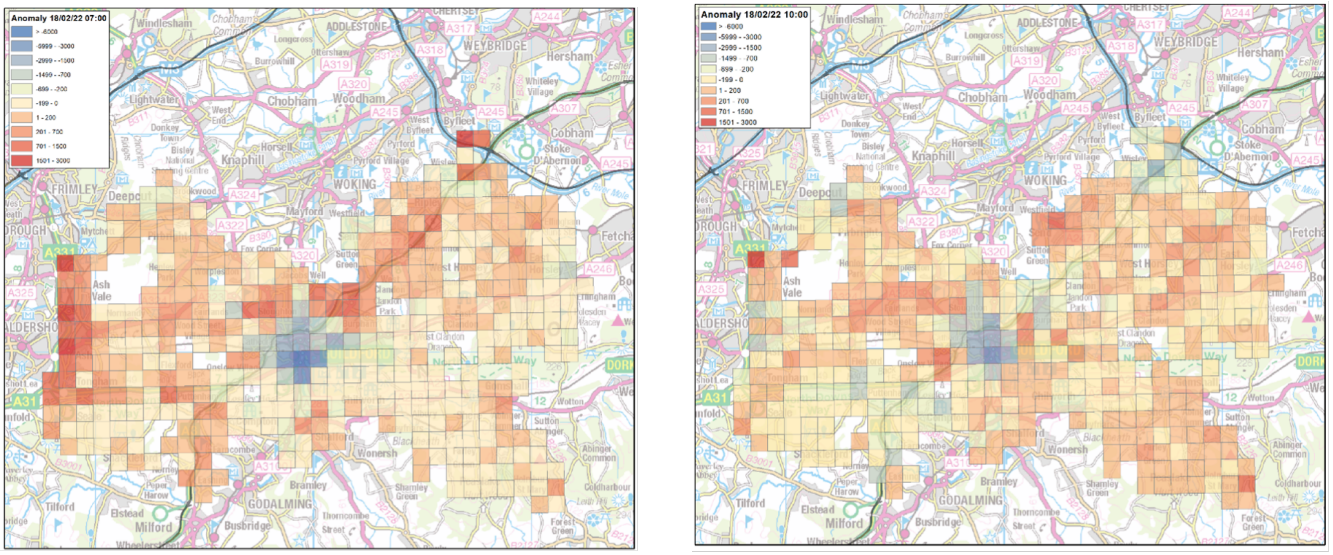
The primary goal of this research is to build on the work of (Robbins et al., 2023) by investigating how various weather conditions influence mobility patterns across different regions. The case study of Storm Eunice serves as a practical demonstration of how real-time mobility data can be used to analyze population movements during extreme weather and support more effective, localized impact-based forecasts.

2.2.2 Storm Eunice: A Practical Application of Mobility Data

Storm Eunice, part of the 2021-2022 European windstorm season, had a severe impact on the UK between 14th and 19th February 2022, with hurricane-force winds causing widespread disruption. Building on the foundational research by (Robbins et al., 2023), the UK Met Office utilized mobility data provided by Vodafone to assess how these extreme weather conditions influenced population movements during the storm. The mobility data included two 14-day periods: 14th-27th February 2020 (a baseline period) and 14th-27th February 2022 (the storm period). This comparative analysis enabled an examination of population movements under normal weather conditions versus during the storm.

The study focused on four regions — Cornwall, Swansea, Guildford, and Brighton — selected for their exposure to

the red weather warnings issued during Storm Eunice. These regions also represent diverse demographic profiles, which are crucial for understanding how different populations respond to severe weather. Mobility data, represented as footfall (the density of people in a given area), was processed using tools such as Python 3, QGIS, and ESRI ArcMap to visualize changes in population movement patterns during the storm. The data was captured at a 1 km x 1 km grid resolution, with each grid cell assigned a unique QuadKey identifier for geospatial analysis.



(a) Footfall at 7am in Guildford

(b) Footfall at 10am in Guildford

Figure 2.2: Footfall anomaly on 18/02/2022 for Guildford.

Figure 2.2 shows the difference in footfall in Guildford at 7 am and 10 am on February 18, 2022, during Storm Eunice, compared to the average footfall on non-stormy days (February 14, 15, 22, 23, 24, and 25) . The time aggregation used in this analysis is hourly, comparing footfall from 7 am to 10 am on February 18 to the hourly average of an entire workday (6 am to 10 pm) for the non-stormy days. The deviation from this baseline is shown as anomalies: red tones indicate a positive anomaly (higher footfall during the storm than average), blue tones indicate a negative anomaly (lower footfall during the storm than average), and yellow tones represent near-average footfall.

The analysis revealed distinct mobility patterns. For instance, in Guildford, the data showed a positive footfall anomaly along major roads at 7 am on 18th February 2022, suggesting increased movement, while central Guildford exhibited a negative anomaly, indicating fewer people than average. By 10 am, footfall had largely returned to near-normal levels, suggesting that the population adjusted their movements in response to the storm's severity. Similar patterns were observed in other regions, highlighting how severe weather can have localized and dynamic effects on population movements (Robbins et al., 2023).

2.2.3 Weather's Broader Impacts

The Storm Eunice case study provides concrete evidence of weather's influence on mobility patterns. However, this research extends beyond a single event to explore the broader relationship between various meteorological conditions and population movements across different environments. This study expands upon the initial findings of (Robbins et al., 2023) in several key areas:

- **Geographic Scope:** The research broadens the geographic scope of analysis to encompass the entire UK, including both urban and rural areas. By comparing mobility responses across these diverse environments

and regions, the study aims to provide a comprehensive understanding of how weather impacts mobility patterns nationwide.

- **Range of Meteorological Conditions:** While the Storm Eunice study primarily focused on extreme wind and storm conditions, this research investigates the impact of a broader range of weather conditions, including temperature, precipitation, and solar radiation. This allows for a more detailed exploration of how different types of weather affect population movements on both weekdays and weekends.
- **Improved Data Processing:** Addressing the challenges identified in the Storm Eunice case study — such as missing data points and inconsistencies in classification—this research employs more advanced data pre-processing techniques. This includes standardizing footfall data, filling in gaps, and creating more accurate baseline measures to ensure robust analysis.

2.2.4 Implications for Future Research

The findings from the Storm Eunice case study underscore the potential of mobility data to enhance IbF by providing real-time insights into population behavior during extreme weather events. Understanding how populations respond to hazardous weather allows authorities to optimize emergency response strategies, resource allocation, and public safety measures.

However, challenges remain, including data inconsistencies, gaps in coverage, and difficulties in classifying mobility attributes such as ”resident” and ”commuter.” Addressing these challenges is crucial for future research to fully realize the potential of mobility data in impact-based forecasting. Further studies should focus on refining mobility data classifications to better account for socio-economic and demographic factors. Additionally, exploring the potential for personalized weather warnings tailored to individual mobility patterns is essential. Continuous collaboration with telecom providers like Vodafone will be vital to ensure that mobility data is accurate, timely, and accessible for use in IbF systems (Robbins et al., 2023).

2.2.5 Conclusion and Link to Research

In conclusion, the Storm Eunice case study provides valuable evidence that weather significantly impacts mobility patterns. By integrating mobility data with meteorological forecasts, this research supports the thesis that weather conditions — particularly extreme weather — cause notable shifts in population movements. The insights gained from this analysis contribute to a growing body of evidence supporting the importance of mobility data in IbF.

Building on this case study, the broader objective of this thesis is to explore how various meteorological conditions affect mobility patterns across different geographic regions. By improving data pre-processing techniques and expanding the scope of analysis to include both urban and rural environments, this research aims to develop more accurate and actionable impact-based forecasts, ultimately improving emergency response, public safety and infrastructure resilience in the face of adverse weather conditions.

The following chapter will explore the data sources and how they are structured, setting the stage for the analysis that follows.

Chapter 3

Data Sources and Summary

This dissertation utilizes two key datasets: Vodafone mobility data and ERA5 reanalysis meteorological data from the ECMWF (European Centre for Medium-Range Weather Forecasts, 2024), produced by the Copernicus Climate Change Service (C3S). These datasets are central to investigating how weather conditions impact mobility patterns across the UK.

3.1 Vodafone Mobility Data

Following its merger with Three UK, Vodafone holds a 32.1% market share of the UK mobile network, making it the largest mobile network operator in the country (Mason, 2023). This significant market presence means that nearly one in three people in the UK are connected to Vodafone's network, providing a robust sample size for analyzing population movements. The Vodafone mobility data is particularly valuable for this research, as it represents a substantial portion of the UK population, allowing for inferences about population-level movement patterns that can be correlated with meteorological conditions to assess how different weather events influence mobility behaviors.

Vodafone's mobility data is collected through mobile phone signals, which track the location of devices over time (Vodafone UK, 2024). This data provides insights into the movement of people within specific geographic areas, making it highly valuable for understanding how weather events affect population movements. The large market share of Vodafone, particularly after the merger, ensures that the mobility data offers a representative sample of the UK population, covering both urban and rural areas. Importantly, this data is anonymized and aggregated to protect individual privacy, yet still provides granular, real-time insights into population behavior that can be directly linked to meteorological data.

3.1.1 Data Summary

The Vodafone dataset consists of eight compressed folders, received on July 11, 2024, each representing a specific time period from August 2023 to April 2024 as tabulated in Table 3.1.

The total compressed data amounts to approximately 52GB, all in CSV format. However, there is a gap in the data for the dates ranging from October 1 to October 22, 2023. This gap is attributed to a custom time standard set by Vodafone, although the specific reason for this omission is unclear.

Folder	Time Period
Folder 1	August 1, 2023 – August 31, 2023
Folder 2	September 1, 2023 – September 30, 2023
Folder 3	October 23, 2023 – November 19, 2023
Folder 4	November 20, 2023 – December 17, 2023
Folder 5	December 18, 2023 – January 14, 2024
Folder 6	January 15, 2024 – February 11, 2024
Folder 7	February 12, 2024 – March 10, 2024
Folder 8	March 11, 2024 – April 7, 2024

Table 3.1: Time periods represented in the Vodafone mobility data folders.

Each folder contains four subfolders with different types of datasets as shown in Table 3.2:

Subfolder	Description
Footfall	Contains quadkeys, the number of unique visitors, and time aggregation.
Footfall Attributes	Similar to Footfall but includes additional attributes like gender, age, and demographics.
Footfall by POI	Adds details about the origin and destination of visitors, along with unique visitors and time aggregation.
Footfall Attributes by POI	Similar to Footfall by POI but with additional demographic details.

Table 3.2: Description of the subfolders in each compressed dataset folder from Vodafone.

Due to the large size of the dataset and time constraints, this study focuses on the ”Footfall” dataset. However, additional variables such as age and gender could be explored in future studies to enhance the precision of statistical modeling.

The dataset derived from the ”Footfall” subfolder across eight separate files (as referenced in Table 3.1), consists of approximately 5,198,283 data points. Each data point is defined by three key attributes: Quadkey, Time Aggregation, and Unique Visitors. This dataset offers extensive geographic coverage across the UK, capturing areas where at least one Vodafone user is active during the specified time intervals. Locations without active Vodafone users, such as remote areas or bodies of water, are excluded from the dataset.

The attributes are defined as follows:

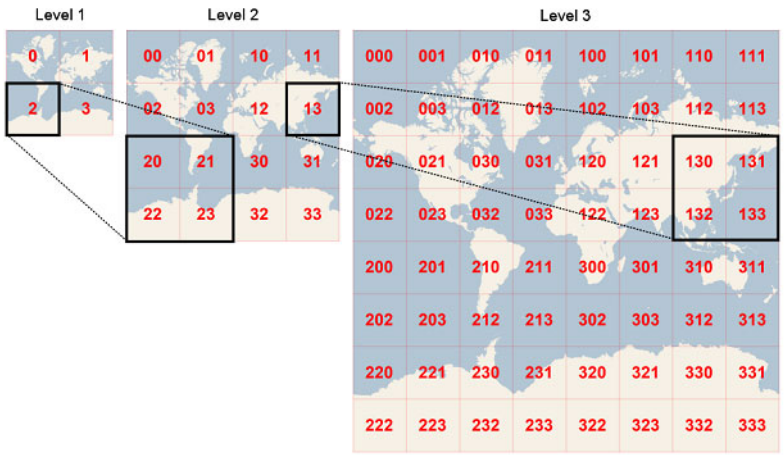


Figure 3.1: Visualization of zoom level of Quadkeys (Source: (Microsoft Bing, 2024))

Quadkey: The term ”quadkey” refers to a quadtree key, which represents a square region in latitude and longitude space, organized by increasing levels of detail. Initially, the entire Earth is divided into four quadkeys, each represented by a single-digit code (0-3). As the zoom level increases, each of these four quadkeys is subdivided further, with an additional digit added to the code. This system allows for progressively finer resolution of geographic data, providing a detailed view of specific regions (Figueira, 2020).

In this dataset, the quadkeys are set at Zoom level 14, corresponding to approximately a 5 km x 5 km grid. Figure 3.1 above illustrates how quadkey zoom levels work.

Number of Unique Visitors: This variable indicates the number of unique visitors to a given location during the specified time period. Each time period is treated independently, meaning that if a visitor travels to Location A in August and returns in September, they will be counted as a unique visitor for each month.

Time Aggregation: The dataset includes different time aggregation categories that provide insights into the movement patterns of individuals across various times of the day and week. These categories are essential for understanding how weather affects mobility at specific times. The following table 3.3 shows the time aggregation:

Time Aggregation	Description
Weekday - all	Unique Visitors across all weekdays (Monday to Friday).
Weekend - all	Unique Visitors all weekends (Saturday and Sunday).
Weekday - night	Unique Visitors between 00:00 and 07:59 on weekdays.
Weekday - workday	Unique Visitors between 10:00 and 15:59 on weekdays.
Weekday - evening	Unique Visitors between 18:00 and 23:59 on weekdays.
Weekend - night	Unique Visitors between 00:00 and 07:59 on weekends.
Weekend - workday	Unique Visitors between 10:00 and 15:59 on weekends.
Weekend - evening	Unique Visitors between 18:00 and 23:59 on weekends.

Table 3.3: Time aggregation categories used in the Vodafone mobility dataset.

A key point to consider is that if a person visits a quadkey (location) multiple times within a specific time aggregation in a month, they are counted only once. This impacts the precision of understanding how often and when people visit specific areas.

For example, the category ”Weekday - workday” aggregates the number of unique visitors to a quadkey (location) between 10:00 and 15:59, across all weekdays in a given month. Figure 3.2 below shows a sample of the footfall dataset for August 2023, where each row corresponds to a specific quadkey and the number of unique visitors.

	qk_id	time_aggregation	n_visitors
0	3113113013203	weekend - all	1664
1	3113131000300	weekend - evening	152
2	3113213320202	weekday - evening	56956
3	3113301300300	weekend - workday	4137
4	3113303010023	weekend - evening	7162
5	3113320212131	weekend - evening	126764
6	3113321113232	weekday - night	2150

Figure 3.2: Sample of Footfall dataset for August 2023, provided by Vodafone.

The quadkey location ”03113113013203” had 1,664 unique visitors during all Saturdays and Sundays in August 2023.

Data Aggregation and Challenges

One significant challenge in this study is that the data is aggregated, meaning it does not offer granular details at the individual trip (temporal) level. While the dataset provides valuable information on population movements, it lacks specifics on how long people stayed, how frequently they visited, or how their movements evolved throughout the day. This limitation affects the precision of the analysis and the conclusions drawn about mobility patterns in relation to weather conditions.

Future research could address these limitations by working with more granular datasets or implementing methods to estimate finer movement details. Nevertheless, the current dataset remains valuable for exploring broad patterns in how populations move during different weather conditions.

3.1.2 Identified issues in Raw (Pre-processed) Dataset

Several issues were identified in the raw dataset:

1. **Mismatch in Zoom Level of Quadkey:** The dataset primarily covers mainland UK regions, where quadkeys are expected to start with 0 or 1, depending on geographic location. However, some quadkeys with leading zeros, since they were stored in a CSV file, when processed in spreadsheet software, these quadkeys were likely interpreted as numerical values, leading to the automatic removal of leading zeros. This unintentional adjustment reduced the zoom level from 14 to 13, resulting in errors in accurately plotting the intended locations (University of Pennsylvania, 2023).

2. **Unique Visitors:** Due to the time aggregation, the data does not capture how often a person visited a specific location within a month. For example, a person who visits Location A *daily* during the first week of a month but not during the remaining three weeks is only counted once, leading to a lack of granularity in visitor behavior.

3. **Lack of Temporal Granularity:** The dataset aggregates footfall data by month and further splits it into subsets of hours on weekdays and weekends. However, this limits our ability to assess the impact of specific events, such as weather conditions or public holidays, on daily footfall patterns.

3.2 ERA5 Reanalysis Meteorological Data

3.2.1 Data Summary

For this study, ERA5 Reanalysis Meteorological Data was sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 is widely recognized in academic and scientific research for its high-resolution, global meteorological reanalysis data, extending from 1940 to the present day. It provides daily meteorological analysis with a resolution of 0.25°, corresponding to an approximately 31 km grid. This fine resolution enables detailed assessments of weather patterns over time and space (European Centre for Medium-Range Weather Forecasts, 2024).

To align with the study's focus on how weather affects mobility patterns, specific geographic boundaries covering mainland UK, between latitudes 49°N to 60°N and longitudes -8°W to 2°E, were selected. The time range spans from August 1, 2023, to April 30, 2024. Due to file size constraints, the data was downloaded in two separate files:

- File 1: August 1, 2023 – December 31, 2023

- File 2: January 1, 2024 – April 30, 2024

Each file, approximately 200 MB in size, is provided in NetCDF format, a common format for array-oriented scientific data. These files contain hourly meteorological data, which will be used to analyze how various weather conditions, such as wind speed, temperature, and precipitation, influence mobility patterns in the UK.

The ERA5 dataset includes several variables relevant to this study, such as wind components, temperature, and gust data, all efficiently processed using xarray and pandas in Python. Table 3.4 provides details of key variables, including their units and metadata.

Variable Name	**Full Name**	**Shape**	**Units**	**Latitude Range**	**Longitude Range**	**Time Range**
longitude	Longitude	(41,)	degrees_east	-	-8.0 to 2.0	-
latitude	Latitude	(45,)	degrees_north	49.0 to 60.0	-	-
time	Time	(Hourly steps,)	N/A	-	-	2023-08-01 to 2024-04-30
u10	10 metre U wind component	(Time, Lat, Lon)	$m\ s^{-1}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
v10	10 metre V wind component	(Time, Lat, Lon)	$m\ s^{-1}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
fg10	10 metre wind gust since previous processing	(Time, Lat, Lon)	$m\ s^{-1}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
d2m	2 metre dewpoint temperature	(Time, Lat, Lon)	K	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
t2m	2 metre temperature	(Time, Lat, Lon)	K	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
i10fg	Instantaneous 10 metre wind gust	(Time, Lat, Lon)	$m\ s^{-1}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
mtpr	Mean total precipitation rate	(Time, Lat, Lon)	$kg\ m^{-2}\ s^{-1}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
ptype	Precipitation type	(Time, Lat, Lon)	code table (4.201)	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
rsn	Snow density	(Time, Lat, Lon)	$kg\ m^{-3}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
sd	Snow depth	(Time, Lat, Lon)	m of water equivalent	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
sf	Snowfall	(Time, Lat, Lon)	m of water equivalent	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
ssr	Surface net shortwave (solar) radiation	(Time, Lat, Lon)	$J\ m^{-2}$	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30

Variable Name	Full Name	Shape	Units	Latitude Range	Longitude Range	Time Range
str	Surface net long-wave (thermal) radiation	(Time, Lat, Lon)	J m^{-2}	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
sp	Surface pressure	(Time, Lat, Lon)	Pa	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
ssrd	Surface short-wave (solar) radiation downwards	(Time, Lat, Lon)	J m^{-2}	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30
tp	Total precipitation	(Time, Lat, Lon)	m	49.0 to 60.0	-8.0 to 2.0	2023-08-01 to 2024-04-30

Table 3.4: Table summarizing variable information, dimensions, and geographical/time coverage for Aug 2023-Apr 2024.

This table summarizes the key variables, their units, and their geographic and temporal coverage. Both files share the same variables and geographic scope, differing only in the time periods they represent. Thus, this table encapsulates the essence of the entire dataset.

By utilizing this detailed meteorological data, the analysis can explore how specific weather variables interact with mobility patterns. These variables form the foundation for understanding the relationship between atmospheric conditions and human movement behavior across different regions of the UK.

3.2.2 Identified issues in Raw (Pre-processed) Dataset

Several issues were identified in the raw ERA5 dataset:

1. **Time Units:** The time is measured in hours since January 1, 1900, at 00:00:00 UTC. This means that each value in the time array represents the number of hours that have passed since that reference date.
2. **Separate Files:** Due to size constraints during download, the data was split into two files, each covering a different time period. These files must be merged for a comprehensive analysis.
3. **File Format:** The ERA5 dataset is in NetCDF format, while the Vodafone mobility data is in CSV format. This discrepancy requires careful processing to ensure compatibility and integration.
4. **Time Aggregation:** The downloaded ERA5 data does not align with the time aggregation used in the mobility dataset, requiring extensive processing, such as summing and averaging across the period, to ensure consistency.

3.3 Supplementary Data Sources

To support the analysis, additional datasets were required beyond the primary mobility and meteorological data.

3.3.1 Quadkey Catalogue

The Quadkey Catalogue is a CSV file in well-known text (WKT) format, providing a reference for Quadkey IDs within the WGS84[1] coordinate system. This file, obtained through a request to Vodafone, is essential for preliminary plotting of boundary lines for each quadkey, as the quadkey itself only provides the center point.

3.3.2 Shapefile

Shapefiles representing the administrative boundaries of the United Kingdom were utilized in this study, downloaded from the GADM (Global Administrative Areas) website. These shapefiles contain essential geographic data, including country borders, regional divisions, and other spatial entities relevant to the analysis. The use of shapefiles allows for accurate mapping and geospatial analysis, which is indispensable for visualizing the impact of weather on mobility patterns within the UK's administrative boundaries.

3.3.3 Daylight Data

The process involved the following steps:

- Grid Setup: Defined latitude and longitude ranges with 0.25° increments.
- Sunrise/Sunset Calculation: Computed sunrise and sunset times for each grid point using the *ephem.Observer* function.
- Exception Handling: Managed cases where the sun does not rise or set by assigning "N/A".
- Daylight Duration: Calculated total daylight minutes by subtracting sunrise time from sunset time.
- Data Storage: Saved the results in a CSV file for further analysis.

This daylight data, when combined with mobility and weather data, will help in exploring the impact of daylight variations on population movement across different regions of the UK.

[1] WGS84 (World Geodetic System 1984) is the standard coordinate system used for mapping and geodesy, using the latitudes (ranging from -90° to +90°) and longitudes (ranging from -180° to +180°).

Chapter 4

Data Cleaning and Manipulation

This chapter outlines the processes used to clean and manipulate the mobility data and ERA5 reanalysis meteorological data. After preprocessing each dataset individually, they will be merged for statistical modeling.

4.1 Preprocessing of Mobility Data

Several steps are necessary to clean, standardize, and prepare the mobility data to ensure compatibility with the ERA5 reanalysis data.

4.1.1 Categorising in Months & Merging

The mobility data is divided into eight separate files, each corresponding to a different time period. The first step involves categorizing the data by the correct month. A new column labeled 'month' is added to each dataset and populated accordingly, as shown in Table 4.1.

Month	File Number	Time Period Covered
August	File 1	August 1, 2023 to August 31, 2023
September	File 2	September 1, 2023 to September 30, 2023
October	File 3	October 23, 2023 to November 19, 2023
November	File 4	November 20, 2023 to December 17, 2023
December	File 5	December 18, 2023 to January 14, 2024
January	File 6	January 15, 2024 to February 11, 2024
February	File 7	February 12, 2024 to March 10, 2024
March	File 8	March 11, 2024 to April 7, 2024

Table 4.1: Categorization of Mobility data into Months

Then, these eight files are combined into a unified dataset, referred to as the '**Mobility Dataset**'.

4.1.2 Correcting Zoom Level to 14

During data processing, certain quadkeys that start with '0' were incorrectly reduced from zoom level 14 to 13 due to how numerical data was interpreted. To correct this, a new column 'corrected_qk_id' was created, adding a '0' in front of affected quadkeys to ensure accurate geographic plotting.

4.1.3 Conversion to Latitude and Longitude Coordinates

Since the ERA5 reanalysis data uses geographic coordinates (latitude and longitude) in the WGS84 coordinate system, it is necessary to convert the quadkeys into corresponding geographic coordinates for compatibility. The method and formulae used to calculate the centroid of each quadkey tile are detailed in Appendix 8. By computing the centroid as the average of the latitudinal and longitudinal bounds of each quadkey tile, the mobility data can be accurately mapped onto the meteorological grid.

4.1.4 Match to the Nearest coordinates of ERA5 Data:

After converting the quadkeys to latitude and longitude coordinates, the resulting coordinates are rounded to two decimal places. However, these rounded coordinates may not exactly match those in the ERA5 dataset. To resolve this, latitude and longitude arrays are extracted from the ERA5 dataset, and a grid of weather coordinates is created. Using a KDTree[1], the nearest weather grid point for each mobility coordinate is efficiently identified (Baeldung, 2023). The latitude and longitude values in the mobility data are then replaced with the closest matching ERA5 coordinates, ensuring consistent spatial alignment between the two datasets.

	qk_id	time_aggregation	n_visitors	Month	corrected_qk_id	latitude	longitude	geometry
0	3113113013203	weekend - all	1664	August	03113113013203	60.00	-0.75	POINT (-0.85000 60.79000)
1	3113131000300	weekend - evening	152	August	03113131000300	60.00	-1.25	POINT (-1.31000 60.19000)
2	3113213320202	weekday - evening	56956	August	03113213320202	57.50	-6.25	POINT (-6.32000 57.45000)
3	3113301300300	weekend - workday	4137	August	03113301300300	58.50	-3.50	POINT (-3.42000 58.40000)
4	3113303010023	weekend - evening	7162	August	03113303010023	58.00	-3.75	POINT (-3.83000 58.04000)
5	3113320212131	weekend - evening	126764	August	03113320212131	56.75	-5.00	POINT (-5.11000 56.82000)
6	3113321113232	weekday - night	2150	August	03113321113232	57.25	-3.00	POINT (-2.93000 57.14000)
7	3113322112033	weekend - night	5544	August	03113322112033	56.50	-4.50	POINT (-4.49000 56.42000)

Figure 4.1: Excerpt of the initial rows from the preprocessed Mobility Dataset.

The final mobility dataset, as shown in Figure 4.1, reflects the correctly mapped latitudes and longitudes. This alignment is verified by comparing the corresponding geometry column, which displays the original converted coordinates.

4.2 Preprocessing of Meteorological Data

This section describes the creation and manipulation of two distinct datasets derived from the raw ERA5 Reanalysis NetCDF files. The first dataset, detailed in subsection 4.2.1, is aggregated on a daily basis to derive secondary weather varibles for later statiscal modelling. The second dataset, discussed in subsection 4.2.2, is formatted to align with the time aggregation used in the Mobility Data to identify general trends and correlations.

4.2.1 Dataset 1: Weather Data (Daily)

Conversion to Pandas Dataframe

Given the differing file formats — mobility data in CSV and ERA5 data in NetCDF — it is essential to convert and manipulate these datasets within a more versatile framework, such as a Pandas DataFrame. Initially, the two

[1] KDTree is a data structure that organizes points in a k-dimensional space to allow for efficient nearest neighbor searches. It is commonly used in spatial data analysis and computational geometry.

NetCDF files, covering the periods from August 1, 2023, to December 31, 2023, and from January 1, 2024, to April 30, 2024, are loaded and converted into xarray Datasets to facilitate the handling of multidimensional data. These datasets are then concatenated along the time dimension, forming a continuous dataset that spans from August 2023 to April 2024. The combined dataset is subsequently converted into a Pandas DataFrame, referred to as 'weather_df', which includes all relevant variables (e.g., temperature, precipitation, wind speed) along with their associated dimensions (time, latitude, longitude). This conversion allows for easier access, transformation, and visualization of the data.

Conversion to Standard time

The dataset's time format, originally representing hours since January 1, 1900, 00:00:00 UTC, is converted into a standard datetime format, as outlined in Appendix 8. The results are stored in a new column labeled 'standard_time', easing further data cleaning and manipulation.

Aggregation by Day and Variable Renaming

The data is currently structured with hourly intervals. To streamline analysis, variables are aggregated or averaged daily, depending on their characteristics. For variables such as temperature and wind speed, which fluctuate continuously throughout the day, a daily mean is calculated. Conversely, variables like precipitation, snowfall, and solar radiation, which accumulate over time, are summed for each day.

To enhance clarity and ease of use, columns are renamed with more descriptive and standardized labels, indicating whether they represent a daily mean or total (e.g., "Temperature (Mean)" or "Precipitation (Total)"). This approach ensures that the dataset is intuitive, allowing users to quickly understand the nature of the data.

The aggregation process is conducted by latitude, longitude, and time period, with each variable being processed according to its specific characteristics. The standard time column from the 'weather_df' DataFrame is utilized in this process. The resulting dataset, titled 'Daily_Weather_Data,' will be employed in subsequent analyses.

	date	latitude	longitude	Temperature (Mean)	Wind Speed U Component (Mean)	Wind Speed V Component (Mean)	Wind Gust (Mean)	Dew Point Temperature (Mean)	Max Wind Gust (Mean)	Total Precipitation Rate (Mean)	Snow Depth (Total)	Snow Density (Total)	Snowfall (Total)
0	2023-08-01	49.0	-8.00	290.935604	9.739130	3.855531	14.548916	289.142042	14.363070	0.000084	2399.999634	0.0	0.0
1	2023-08-01	49.0	-7.75	290.955482	9.658124	3.751059	14.507827	289.119486	14.305549	0.000082	2399.999634	0.0	0.0
2	2023-08-01	49.0	-7.50	290.978828	9.564773	3.644561	14.466180	289.115820	14.214127	0.000091	2399.999634	0.0	0.0
3	2023-08-01	49.0	-7.25	290.991859	9.453436	3.557422	14.458456	289.077854	14.199122	0.000089	2399.999634	0.0	0.0
4	2023-08-01	49.0	-7.00	290.995673	9.343227	3.480240	14.436755	289.045561	14.180353	0.000090	2399.999634	0.0	0.0

Figure 4.2: Excerpt of the initial rows from the preprocessed Weather (Daily) Dataset.

Figure 4.2 provides a visual representation of the processed dataset, offering a comprehensive overview of weather conditions for each geographic location on a given date.

4.2.2 Dataset 2: Weather Data (Aggregated)

Logic for Time Aggregation

To align the weather data with the structure of the mobility dataset, a new GeoPandas DataFrame[2] was created, incorporating geospatial data to match the format of the mobility data. The ERA5 weather datasets, covering the period from August 2023 to April 2024, were processed using the xarray library. These datasets, initially in hourly intervals, were concatenated along the time dimension to form a continuous dataset. To capture broader trends and enhance analysis, the data was aggregated by month. A custom function was employed to classify each timestamp into its respective month period based on the date ranges provided in the mobility data. Table 4.2 summarizes how these dates were categorized into corresponding months:

Time Period	Categorized Month
August 1, 2023 - August 31, 2023	August
September 1, 2023 - September 30, 2023	September
October 1, 2023 - October 22, 2023	October - Not Included
October 23, 2023 - November 19, 2023	October
November 20, 2023 - December 17, 2023	November
December 18, 2023 - December 31, 2023	December
January 1, 2024 - January 14, 2024	January
January 15, 2024 - February 11, 2024	February
February 12, 2024 - March 10, 2024	February
March 11, 2024 - April 7, 2024	March
April 8, 2024 - April 30, 2024	Others

Table 4.2: Categorization of Weather Data into Months

To further refine the analysis, the weather data was categorized into two main groups: ”Weekday - All” and ”Weekend - All.” This time aggregation approach enables a focused analysis of overall weather conditions across weekdays and weekends for each month period, effectively capturing the variations between these periods. Following this classification, the data was converted into a Pandas DataFrame, which was organized by latitude, longitude, month, and the ”Weekday - All” or ”Weekend - All” categories. Depending on the nature of each weather variable, the data was either averaged (for continuous variables like temperature) or aggregated (for cumulative variables like precipitation and radiation) to derive meaningful insights.

Combining with Daylight Data

Following the time aggregation process for the weather data, the focus shifted to the ’Daylight Data’ dataset, which recorded daily daylight minutes. This dataset had not yet been aggregated, so it was first classified into the same monthly periods used for the weather data. Subsequently, the daylight data was further divided into ”Weekday - All” and ”Weekend - All” categories, with average daylight minutes calculated for each group, following a similar methodology to that applied to the weather data in subsection 4.2.2.

After completing this aggregation, the daylight data was integrated with the weather data to create the final ’Aggregated Weather Data’ dataset. During the merging process, month periods such as ”October - Not Included” and

[2]GeoPandas Dataframe inherits from Pandas DataFrame but adds support for geospatial data. It has a special geometry column that contains geometric objects such as points, lines, and polygons. This column enables spatial operations, like distance calculations, intersection checks, and geometric transformations.

"Others" were excluded, due to the absence of corresponding mobility data. This step ensured that the final dataset accurately reflects the appropriate time frames necessary for analysis.

The resulting dataset, organized by latitude, longitude, month, and time aggregation (weekday versus weekend), provides a robust foundation for exploring how variations in weather and daylight influence mobility patterns across different regions and timeframes.

Merging with Mobility Data

To facilitate statistical modeling, the weather (aggregated) dataset from Section 4.2.2 and the Mobility dataset from Section 4.1.4, forming a primary combined dataset. This merged dataset serves as the foundation for the analysis and will be revised as needed throughout the research process.

The merging was carried out using the groupby() function in Python, aligning the datasets by common attributes such as latitude, longitude, month period, and time aggregation. Ensuring consistent column names across both datasets allowed for smooth integration.

As illustrated in Figure 4.3, the resulting dataset includes both the number of unique visitors and relevant weather variables, organized by time period and geographic location. This comprehensive structure provides a robust basis for detailed statistical analysis and modeling.

	Time Aggregation	n_visitors	Month	corrected_qk_id	latitude	longitude	Daylight Minutes (Mean)	Temperature (Mean)	Wind Speed U Component (Mean)	Wind Speed V Component (Mean)	Wind Gust (Mean)	Total Precipitation Rate (Sum)	Snow Dep (Tot
0	Weekend - All	1664	August	03113113013203	60.0	-0.75	935.714583	286.652857	-0.738622	0.480853	6.977686	0.002954	19199.997
1	Weekend - All	3383	August	03113113031000	60.0	-0.75	935.714583	286.652857	-0.738622	0.480853	6.977686	0.002954	19199.997
2	Weekend - All	2361	August	03113113031001	60.0	-0.75	935.714583	286.652857	-0.738622	0.480853	6.977686	0.002954	19199.997
3	Weekend - All	1162	August	03113113033233	60.0	-0.75	935.714583	286.652857	-0.738622	0.480853	6.977686	0.002954	19199.997
4	Weekend - All	2730	August	03113113013221	60.0	-0.75	935.714583	286.652857	-0.738622	0.480853	6.977686	0.002954	19199.997

Figure 4.3: Excerpt of the initial rows from the merged Mobiity and Weather (Aggregated) Dataset.

It is important to note that, due to the conversion of quadkeys to two-decimal coordinates, the dataset contains multiple entries for the same coordinates and time periods, as shown in Figure 4.4. While these entries could be filtered, maintaining the current structure without aggregation is advisable to preserve the dataset's robustness and avoid potential bias in the analysis.

(523, 20)

	Time Aggregation	n_visitors	Month	corrected_qk_id	latitude	longitude	Daylight Minutes (Mean)	Temperature (Mean)	Wind Speed U Component (Mean)	Wind Speed V Component (Mean)	Wind Gust (Mean)	Total Precipitation Rate (Sum)
9790	Weekday - All	1231	August	03113113221213	60.0	-1.25	935.499275	286.008006	0.88409	-1.484068	8.463993	0.010644
9791	Weekday - All	4799	August	03113113203022	60.0	-1.25	935.499275	286.008006	0.88409	-1.484068	8.463993	0.010644
9792	Weekday - All	7535	August	03113113220011	60.0	-1.25	935.499275	286.008006	0.88409	-1.484068	8.463993	0.010644
9793	Weekday - All	3851	August	03113131020131	60.0	-1.25	935.499275	286.008006	0.88409	-1.484068	8.463993	0.010644
9794	Weekday - All	16414	August	03113131003220	60.0	-1.25	935.499275	286.008006	0.88409	-1.484068	8.463993	0.010644

Figure 4.4: Excerpt from the merged Mobility and Weather Dataset, illustrating multiple data points per unique coordinate.

This final dataset, saved as the 'Merged Dataset,' consists of 523 rows, representing distinct data points across various locations and time periods, encompassing multiple quadkeys and months.

Chapter 5

Methodology

This chapter explores and evaluates the models and trends within the dataset, addressing challenges such as lack of time granularity, geographic variations[1], and the short time range of the dataset, which limits the ability to confirm long-term trends. In other words, the lack of historical data from previous years makes it difficult to establish definitive patterns. To overcome these challenges, multiple analytical strategies will be employed to uncover meaningful insights.

5.1 Hypotheses

The hypothesis is as follows:

Null Hypothesis (H_0): Weather conditions do not significantly affect mobility patterns. Specifically, there is no difference in outdoor activity between favorable and unfavorable weather conditions.

Alternative Hypothesis (H_1): Weather conditions significantly affect mobility patterns. People are more likely to engage in outdoor activities during favorable weather conditions, such as dry days with higher temperatures.

To test this hypothesis, statistical methods like regression analysis, and machine learning models such as Random forest and Gradient Boosting will be applied to determine if weather conditions impact human mobility.

5.2 Seasonal Considerations

To analyze mobility patterns across different months, the focus will be on seasonal trends, comparing mobility on weekdays and weekends. The analysis will filter the 'Weekday - All' and 'Weekend - All' values from the Merged Dataset and plot the trends to identify patterns.

By examining these trends, distinct seasonal variations in mobility can be observed in figure 5.1. A notable increase in visitor numbers is seen in August, likely driven by summer tourism and outdoor activities, followed by a decline from September through the winter months. A slight recovery is noted in February and March as the weather improves. This trend is almost similar across both weekdays and weekends.

[1]Geographic variations refer to the importance of each location, such as high mobility in summer at coastal and hiking regions, or increased activity in skiing areas during winter.

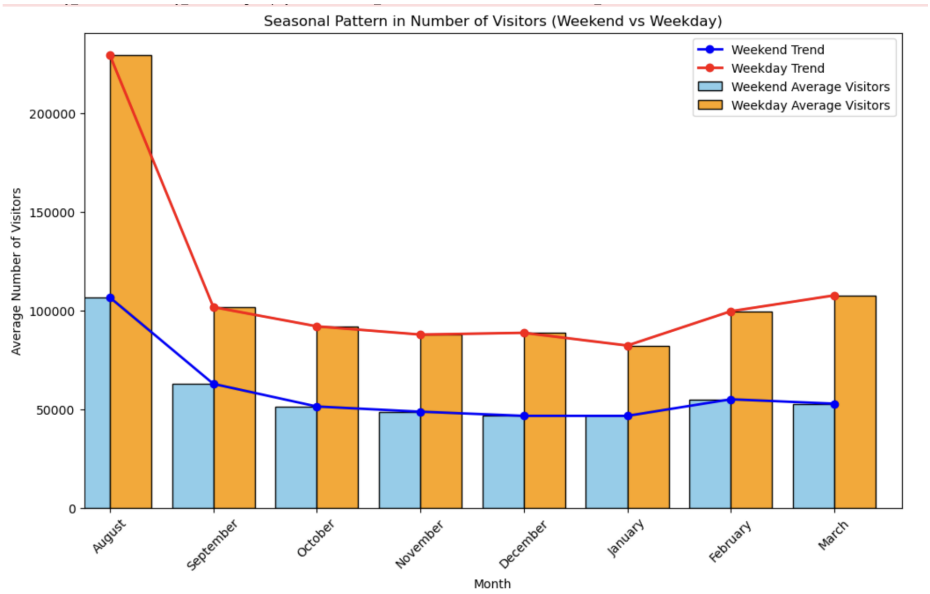


Figure 5.1: Bar plot comparing the mobility trend of the visitors over the weekdays and weekends (2023-2024)

These patterns serve as a foundational stone in understanding whether weather and seasonal factors influence human mobility, thereby providing insights into the impact of weather on daily human mobility.

5.3 Correlation Analysis

Before developing effective data models, it is important to first explore the relationships between weather and mobility variables through correlation analysis. This preliminary step helps identify which factors are most strongly associated with mobility patterns, guiding the direction of further investigation. The complete methodology, formulae and detailed results used in this analysis can be found in Appendix 8.1.

5.3.1 Correlation between Meteorological Variables and Mobility Data across the UK

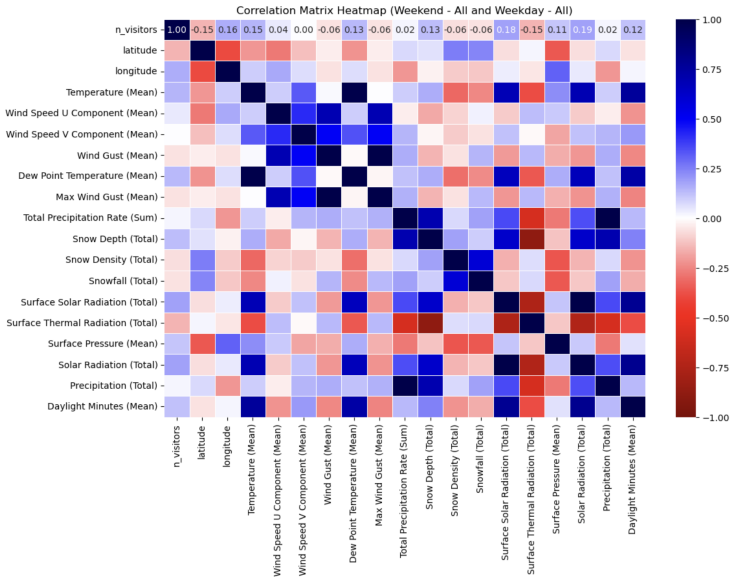


Figure 5.2: Correlation Matrix between Meteorological Variables & Mobility during Weekdays and Weekends

Using the Merged Dataset from section 4.2.2, a correlation matrix was created, focusing on the ”Weekend - All” and ”Weekday - All” categories. These categories were chosen to minimize bias, as including other time aggregations—such as peak work hours or nighttime — could distort the results. Figure 5.2 and table in Appendix 8.1.1 shows that Solar Radiation (Total) and Surface Solar Radiation (Total) has the strongest positive correlations with mobility, while, Surface Thermal Radiation (Total) and latitude shows the strongest negative correlations.

However, the overall correlation values are low, likely due to geographic and temporal variations across the United Kingdom. To gain more precise insights, it may be necessary to focus on a subset of locations.

5.3.2 Correlation between Meteorological Variables and Mobility Data in Specific Regions (Rural & Urban)

To capture geographic variability in mobility patterns, a focused correlation analysis was conducted across selected urban and rural locations. The locations and their corresponding coordinates are detailed below table 5.1:

Table 5.1: Coordinates for Selected Urban and Rural Locations

Location	Latitude Range	Longitude Range
Greater London	51.25 to 51.75	-0.5 to 0.5
Greater Manchester	53.25 to 53.75	-2.75 to -2
Edinburgh	55.75 to 56	-3.5 to -3
Windermere (Lake District)	54.25 to 54.5	-3 to -2.75
St. Ives (Cornwall)	50 to 50.25	-5.5 to -5.25

The correlation matrix in Figure 5.3 revealed that Snow Depth (Total), Solar Radiation (Total), Surface Solar Radiation (Total) and Precipitation (Total) has the strongest positive correlations with mobility, while Surface Thermal Radiation (Total) showed the strongest negative correlation.

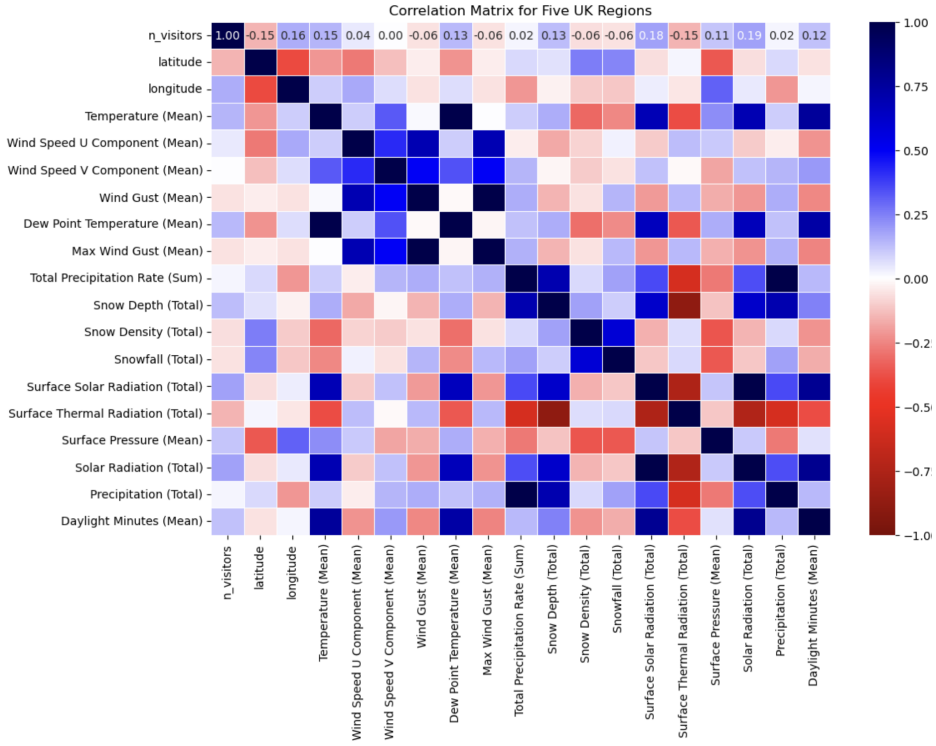


Figure 5.3: Correlation Matrix between Meteorological Variables & Mobility (Urban and Rural)

However, Snow Depth and Solar Radiation may not be the most reliable indicators due to regional variability and inconsistent solar exposure across the UK. Precipitation emerged as a more consistent and relevant variable for analyzing mobility trends across different regions. These findings underscore the importance of considering both geographic and meteorological factors in analyzing mobility patterns.

5.4 Feature Importance using Random Forest

Feature importance is a technique used to identify the most relevant factors to include in statistical modeling across the UK (Built In, 2023) across the UK. The figure 5.4 visualizes the importance of various meteorological variables in predicting mobility patterns, where the length of each bar represents the importance of the corresponding variable, while the black lines indicate the confidence intervals, showing the range of uncertainty in the importance measure.

The plot reveals that latitude and longitude hold the highest importance, suggesting significant geographic influence on mobility. Among the meteorological factors, Temperature (Mean) and Solar Radiation (Total) emerge as key predictors, highlighting the role of favorable weather conditions in increased mobility.

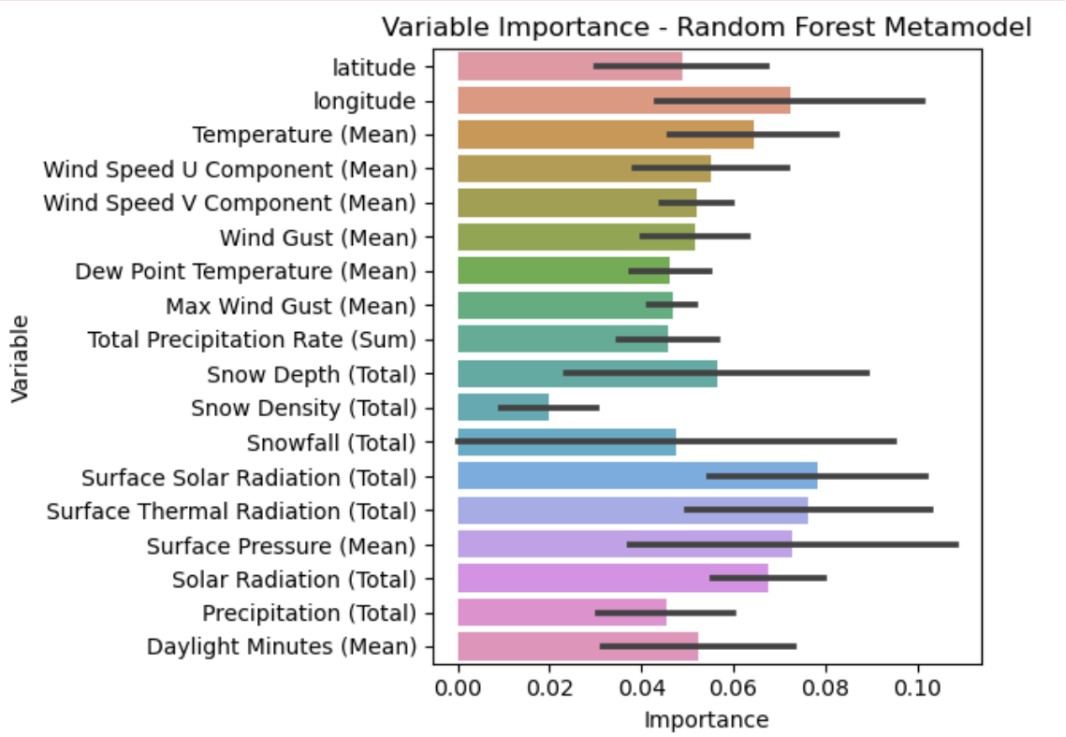


Figure 5.4: Feature Importance using Random Forest across the UK

Other variables, such as Wind Speed Components, Surface Solar Radiation (Total), and Precipitation (Total), also contribute to the model, though with varying degrees of importance. Interestingly, factors like Snowfall (Total) and Snow Depth (Total) show moderate importance, reflecting their relevance in specific regions and seasons.

This feature importance ranking provides valuable insights into which weather variables are most relevant in shaping human mobility patterns, helping to refine predictions.

5.5 Generation of Additional Weather Variables

Based on the feature importance and correlation analysis, key weather variables such as Precipitation, Daylight, Surface Solar Radiation, and Temperature were identified as influential in determining mobility patterns. To formulate the statistical models, secondary weather variables were derived to account for the lack of time granularity in the mobility dataset.

5.5.1 Precipitation - Wet Days & Dry Days:

Days were classified as Dry if daily precipitation was below 0.1 mm and as Wet if it exceeded 25 mm in the *Daily_Weather_Data* dataset from section 4.2.1, as used by the UK Met Office (Met Office, 2024a). These thresholds help differentiate days with minimal precipitation from those with heavy rainfall. The days were further categorized into weekdays and weekends to explore how precipitation impacts mobility across different periods.

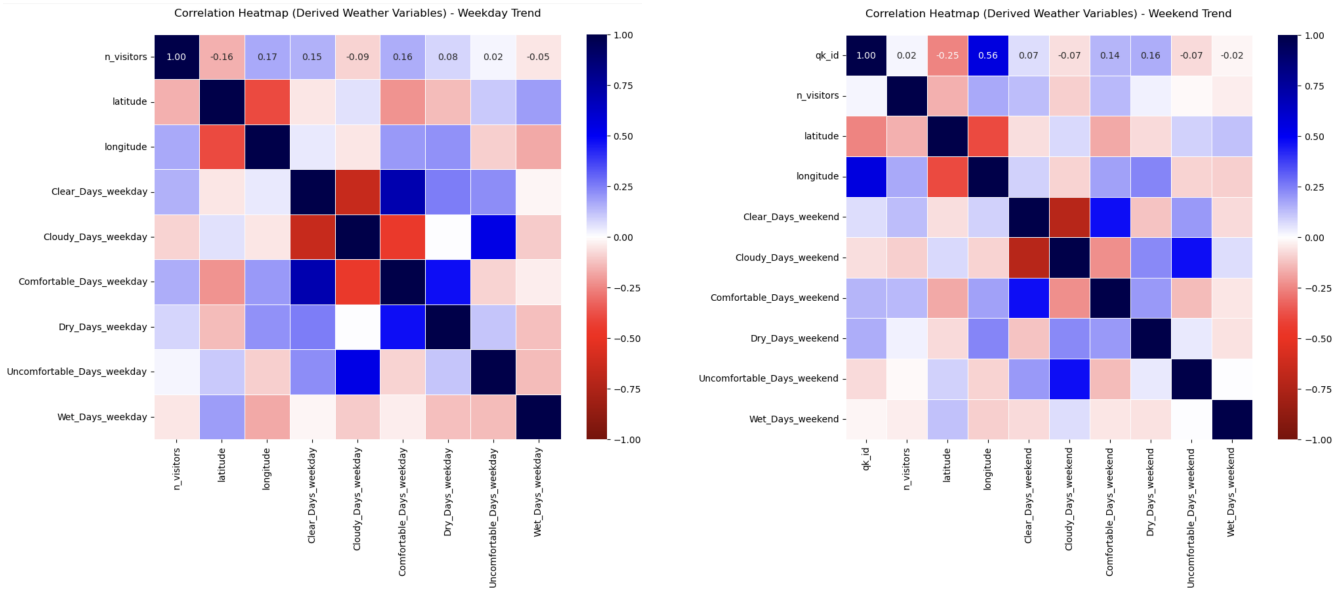
5.5.2 Temperature - Comfortable Days & Uncomfortable Days:

Temperature was analyzed by categorizing days as Comfortable if the mean temperature fell between 18°C and 22°C, as recommended by the NHS (Verasamy, 2023), and as Uncomfortable if outside this range. This classification allows for an examination of how thermal comfort levels influence mobility, with moderate temperatures likely increasing mobility.

5.5.3 Surface Solar Radiation - Clear Days & Cloudy Days:

Days were classified as Clear if surface solar radiation exceeded 10.2 million J/m², representing the 75th percentile of the data distribution. Days with lower solar radiation were classified as Cloudy. This distinction helps identify the influence of sunlight exposure on mobility patterns, with clearer days generally encouraging more movements.

5.5.4 Correlation Analysis of Derived Weather Variables:



(a) Correlation - Trend over the Weekdays (2023-2024)

(b) Correlation - Trend over the Weekends (2023-2024)

Figure 5.5: Correlation plot depicting the mobility trend of the unique visitors over the weekdays and weekends

These derived variables were incorporated into the dataset and analyzed separately for weekdays and weekends.

The correlation matrices in figure 5.5 and table in Appendix 8.1.2 showed positive correlations for variables like longitude, Comfortable Days, Clear Days, and Dry Days, indicating that favorable weather conditions align with increased mobility. Conversely, negative correlations were observed for Uncomfortable Days, Cloudy Days, and Wet Days, suggesting a reduction in mobility under less favorable weather conditions.

These findings set the stage for further statistical modeling to validate the impact of these weather variables on mobility patterns. Before proceeding, visualizing the data on a map will provide additional insights into geographic trends.

5.6 Data Visualization: Comparative Mapping

Data visualization is essential for uncovering patterns and trends in complex datasets. In this analysis, geographic maps were plotted using the shapefile from section 3.3.2 to compare the relationship between meteorological variables and mobility trends across the UK. The focus here is on comparing Dry Days, Wet Days, and Mobility Data, as these visualizations offer a clear comparison of how different weather conditions correlate with mobility patterns.

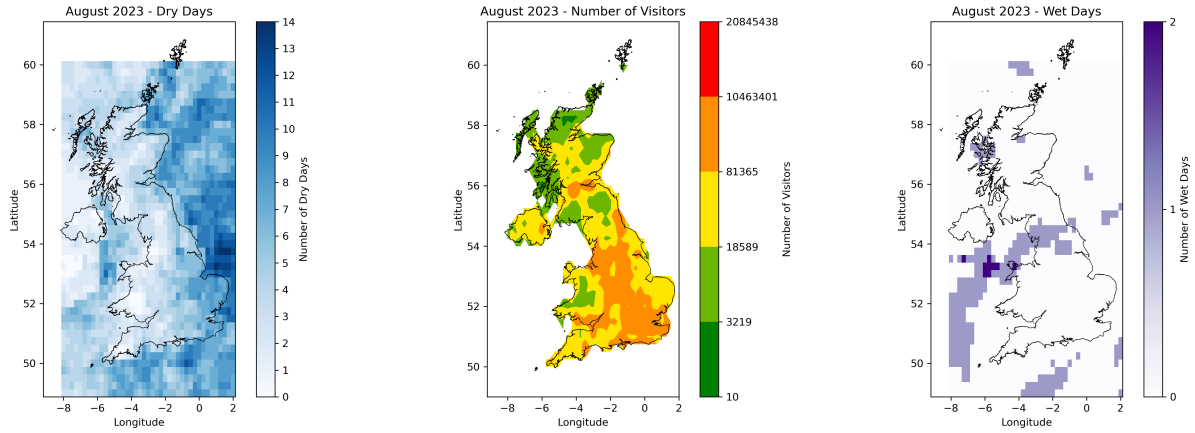


Figure 5.6: Comparative plot for August 2023

Figure 5.6 reveals that regions with more dry days are associated with higher mobility, as indicated by the greater number of unique visitors, whereas areas with more wet days tend to experience reduced mobility, suggesting a comparative influence of weather conditions on mobility in the UK during August.

Additional monthly comparative visualizations are provided in Appendix 9.6 for further analysis.

5.7 Statistical Modeling Across the UK

In this section, various regression models were employed to assess the influence of derived weather variables on mobility across the UK, including supplementary variables such as latitude, longitude, and daylight minutes. The models tested ranged from simple linear regression to more advanced techniques like Random Forest and Gradient Boosting.

- **Random Forest**: An ensemble method that combines the predictions of multiple decision trees to enhance accuracy and handle complex, non-linear relationships (IBM, 2024).

- **Gradient Boosting**: A sequential model that builds decision trees to correct the errors of previous ones, effective in capturing subtle patterns in the data (Displayr, 2024).
- **Generalized Additive Model (GAM)**: A model that extends linear regression by allowing non-linear relationships through smooth functions, offering a balance between interpretability and complexity (Towards Data Science, 2024).

Given the nationwide scope of the dataset, the models generally exhibited weak performance due to the varied geographic locations and lack of time granularity. The dataset primarily captures aggregated trends for weekdays across the UK. The table 5.2 below summarizes the key performance metrics of the models, underscoring the challenges of modeling mobility across such a diverse geographic area:

The table below summarizes the key performance metrics of the models:

Model	R-squared	Mean Absolute Error	Root Mean Squared Error
Random Forest	0.3533	1.3821	1.7497
Gradient Boosting	0.2660	1.4860	1.8641
GAM (Generalized Additive Model)	0.2063	1.5492	1.9358
Polynomial Transformation	0.204	-	-
Linear Regression	0.059	-	-
Log Transformation	0.137	-	-

Table 5.2: Summary of Model Performance Across the UK

The Random Forest model emerged as the best performer with an R-squared value of 0.3533, indicating a moderate level of explanatory power. Gradient Boosting followed, demonstrating similar but slightly lower accuracy. The GAM model provided valuable interpretability, though it performed slightly less effectively than Random Forest and Gradient Boosting. Polynomial Transformation and Linear Regression showed weaker results, with the latter achieving an R-squared value of just 0.059. Log Transformation, while slightly improving performance, did not offer substantial gains in accuracy.

These outcomes underscore the complexities of modeling mobility across the UK's diverse geography, where variations in location and the absence of detailed temporal data can impact model performance. Later sections will narrow the focus to more specific regions to enhance accuracy and interpretability.

5.8 Statistical Modelling: Urban and Rural locations

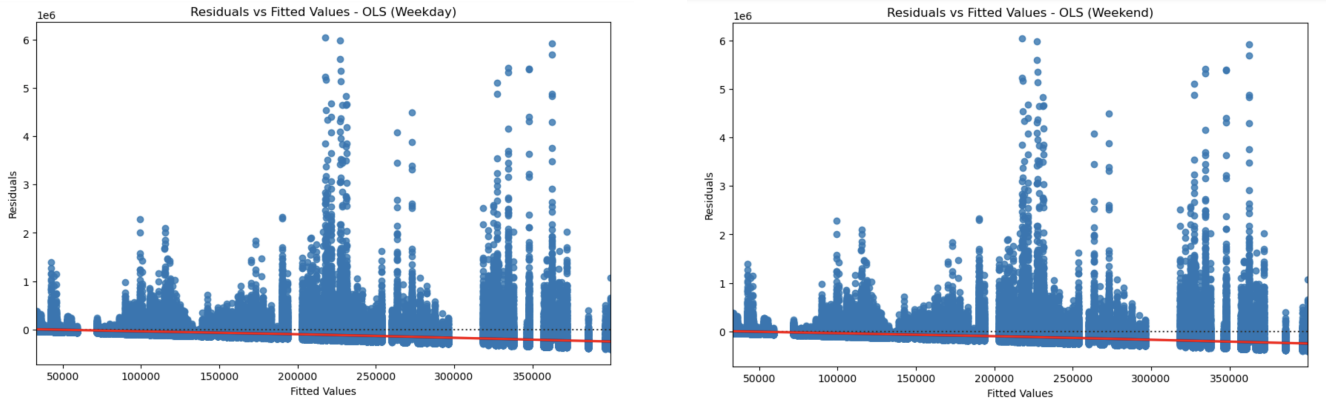
After finding that nationwide models performed poorly, likely due to the diversity of geographic locations and a lack of time granularity, a more focused approach was adopted. This section narrows the analysis to specific urban and rural areas — London, Manchester, Edinburgh, St. Ives, and Windermere — where distinct geographic and demographic factors may provide clearer insights.

5.8.1 OLS Regression Models: Linear Regression

The Ordinary Least Squares (OLS) regression method was chosen as an initial approach to model the relationship between geographic and meteorological factors and visitor numbers. The OLS models for weekdays and weekends offered a baseline understanding but also revealed significant limitations. These limitations were particularly evident in the models' ability to handle non-linear relationships and heteroscedasticity.

Weekday Model: The OLS regression for weekdays explained only 9.6% of the variance in visitor numbers, indicating a relatively weak model fit. Significant predictors included latitude, longitude, and the number of wet days. Specifically, longitude had a positive impact on visitor numbers, while latitude and wet days were associated with a decrease in visitor counts. However, the residuals vs. fitted values plot (see Appendix Figure B.1) demonstrated clear signs of heteroscedasticity, with the spread of residuals increasing alongside higher fitted values. This violates a key assumption of OLS regression. Additionally, the Q-Q Plot (Appendix Figure B.2) revealed non-normality of residuals, and the residual distribution (Appendix Figure B.3) was heavily skewed, indicating that a linear model might not fully capture the complexity of the data.

Weekend Model: The OLS model for weekends performed even less robustly, explaining only 8.7% of the variance in visitor numbers. Predictors such as latitude, longitude, and wet days remained significant, with similar effects as observed in the weekday model. The residuals vs. fitted values plot (Appendix Figure B.4) for the weekend model showed an even more pronounced pattern of heteroscedasticity compared to the weekday model. The Q-Q Plot (Appendix Figure B.5) and residual distribution (Appendix Figure B.6) further reinforced the inadequacies of a linear approach, particularly in accounting for variability in visitor numbers across different conditions.



(a) Residual vs Fitted Values - Weekday

(b) Residual vs Fitted Values - Weekend

Figure 5.7: Linear Regression - Residual vs Fitted Values plot

Given these observations, it was clear that a more sophisticated approach was necessary to address the non-linearity and heteroscedasticity present in the data.

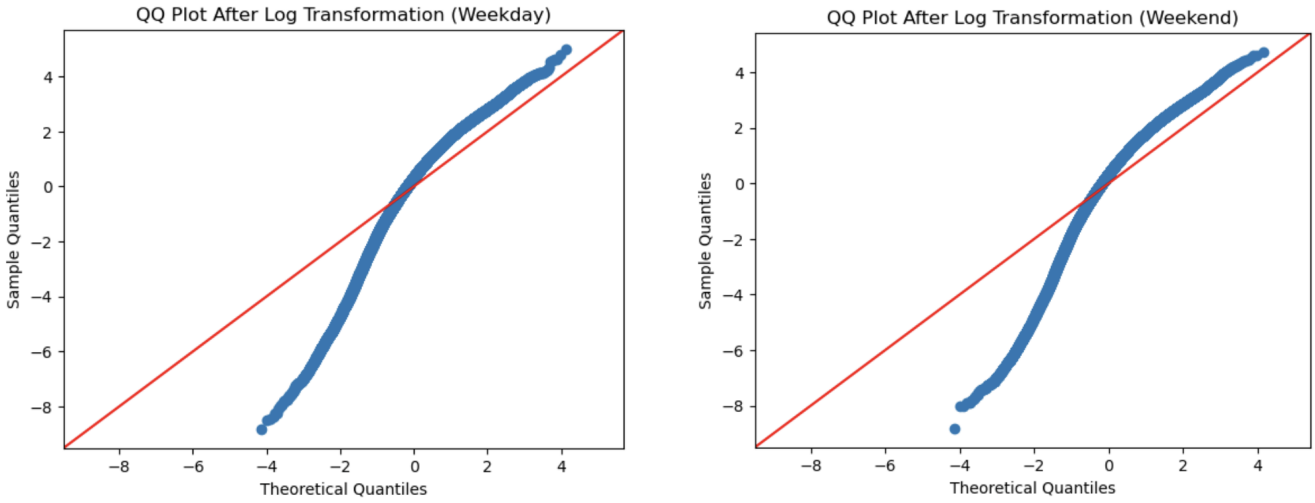
5.8.2 OLS Regression Models: Log-Transformed Regression

To better understand the relationship between geographic and meteorological variables and visitor trends, OLS regression models were applied to log-transformed visitor numbers. The log transformation was intended to address the issues of heteroscedasticity and non-linearity observed in the initial models. While some improvements were noted, the transformation did not fully resolve these issues.

Weekday Model: The log-transformed OLS regression for weekdays explained 13.2% of the variance in the logarithm of visitor numbers. Key predictors included latitude, longitude, and the number of wet days. Longitude positively influenced visitor numbers, while latitude and wet days had a negative impact. Despite these improvements, the Q-Q Plot (Figure 5.9a) revealed significant deviations from normality, particularly in the tails, indicating that the log transformation did not completely resolve the non-normality of residuals. The residuals vs. fitted values plot (Appendix Figure B.7) further supported this, showing persistent heteroscedasticity where the variance of residuals increased with the fitted values. The residual distribution (Appendix Figure B.9) also highlighted that

while the residuals are more symmetrically distributed after the log transformation, significant issues remain in their spread.

Weekend Model: The weekend model explained 15.4% of the variance in the logarithm of visitor numbers. Similar to the weekday model, latitude, longitude, and wet days were significant predictors. However, the Q-Q Plot (Figure 5.9b) for the weekend model also exhibited deviations from normality, particularly in the extreme values, indicating persistent issues with the model fit. The residuals vs. fitted values plot (Appendix Figure B.8) displayed even more pronounced heteroscedasticity compared to the weekday model, indicating that the variability in visitor numbers during weekends is not fully captured by the log-transformed model. The residual distribution (Appendix Figure B.10) similarly reflected these challenges, suggesting that despite the log transformation, the weekend model struggles to account for the underlying variance in the data.



(a) Log Transformed QQ Plot - Weekday

(b) Log Transformed QQ Plot - Weekend

Figure 5.8: Log Transformed Linear Regression - QQ plot

Diagnostic Plots: The Q-Q Plots (Figure 5.9) indicate significant deviations from normality, especially in the tails, suggesting that residuals are not perfectly normally distributed even after log transformation. This is a strong indication that further modeling adjustments or alternative approaches are necessary. The residuals vs. fitted values plots (Appendix Figures B.7 and B.8) illustrate persistent patterns of heteroscedasticity, and the residual distributions (Appendix Figures B.9 and B.10) emphasize the distribution issues within the residuals, reflecting a spread that indicates the presence of heteroscedasticity and non-normality. The analysis suggests that while the log transformation provided some improvement, significant diagnostic issues remain, particularly with heteroscedasticity and non-normal residuals. These issues imply that more advanced modeling techniques, such as non-linear models or machine learning approaches, may be necessary to better capture the complexity of the data.

5.8.3 Generalized Additive Model (GAM) Analysis

To address the limitations observed in the Ordinary Least Squares (OLS) regression models, particularly regarding their handling of non-linear relationships and heteroscedasticity, Generalized Additive Models (GAMs) were employed. GAMs offer a flexible modeling approach by incorporating non-linear smooth functions of the predictors, better capturing the complex interactions between geographic and meteorological variables and visitor numbers.

Weekday Model: The Gradient Boosting Model for weekdays explained 36.8% varia

Latitude: Non-linear effects were evident, with some latitudinal bands showing significant deviations due to local factors. Longitude: Certain longitudes were associated with significantly higher visitor numbers, indicating geographic clustering of popular destinations. Clear Days: The number of clear days showed a positive but complex relationship with visitor numbers, with diminishing returns after a certain point. Cloudy Days: A non-linear relationship was observed, with moderate cloudiness sometimes correlating with higher visitor counts. Comfortable Days: The number of comfortable days had a strong positive effect, though this effect plateaued at higher values. Uncomfortable Days: An increase in uncomfortable days led to a significant drop in visitor numbers, highlighting the impact of adverse weather. Wet Days: Visitor numbers declined as wet days increased, but this effect varied depending on the frequency of wet weather, with a more nuanced impact observed at higher frequencies. Dry Days: Dry days showed a positive relationship with visitor numbers, though with some variability indicating other interacting factors. The Residual Distribution plot (Appendix Figure B.16) for the weekday model indicated good overall performance but also highlighted some challenges in handling outliers.

Partial Dependence Plots (PDP) illustrate the effect of each predictor variable on the response variable while holding other predictors constant. These plots are critical for understanding the specific influence of geographic and weather variables in the GAM models.

Weekday Model: The PDPs for the weekday model (Appendix Figure B.12) show non-linear effects of latitude, longitude, and weather variables, with noticeable fluctuations across the ranges. Weekend Model: The PDPs for the weekend model (Appendix Figure B.13) highlight similar trends to the weekday model, with notable differences in the effect of weather variables like wet days.

Weekend Model: For weekends, the GAM explained 36.6% of the variance. The model struggled slightly more with weekend data due to increased variability in visitor numbers:

Latitude: Non-linear relationships were observed, with certain latitudinal bands showing unexpected patterns, possibly due to regional attractions or events. Longitude: The positive influence of longitude was consistent with weekday findings, again highlighting specific longitudes associated with higher visitor numbers. Clear Days: Clear days continued to show a positive impact, though the effect was more variable on weekends. Cloudy Days: The relationship with cloudy days was more pronounced on weekends, suggesting that visitors may be less deterred by moderate cloudiness during their leisure time. Comfortable Days: Comfortable days had a strong positive effect, similar to weekdays, but with some additional variability. Uncomfortable Days: The negative impact of uncomfortable days was more significant on weekends, likely due to the increased expectation of favorable weather. Wet Days: The model captured the variability in how wet weather affected visitor numbers, with a more pronounced decline compared to weekdays. Dry Days: Dry days showed a positive impact, but the relationship was less consistent on weekends. The Residual Distribution plot for the weekend model (Appendix Figure B.13) reveals that while the model provided robust predictions, capturing all aspects of visitor behavior, particularly during weekends, remains challenging.

Model Diagnostics: The Q-Q plots (Appendix Figures B.14 and B.15) for both the weekday and weekend models showed deviations from normality, particularly in the extremes, suggesting some persistent issues with model fit. The residuals vs. fitted values plots (Appendix Figures B.16 and B.17) displayed heteroscedasticity, indicating that variability in visitor numbers is not fully captured by the model. The residual distributions (Appendix Figures B.18 and B.19) further reflected these challenges, emphasizing the need for continued refinement of the models to better handle the data's complexity.

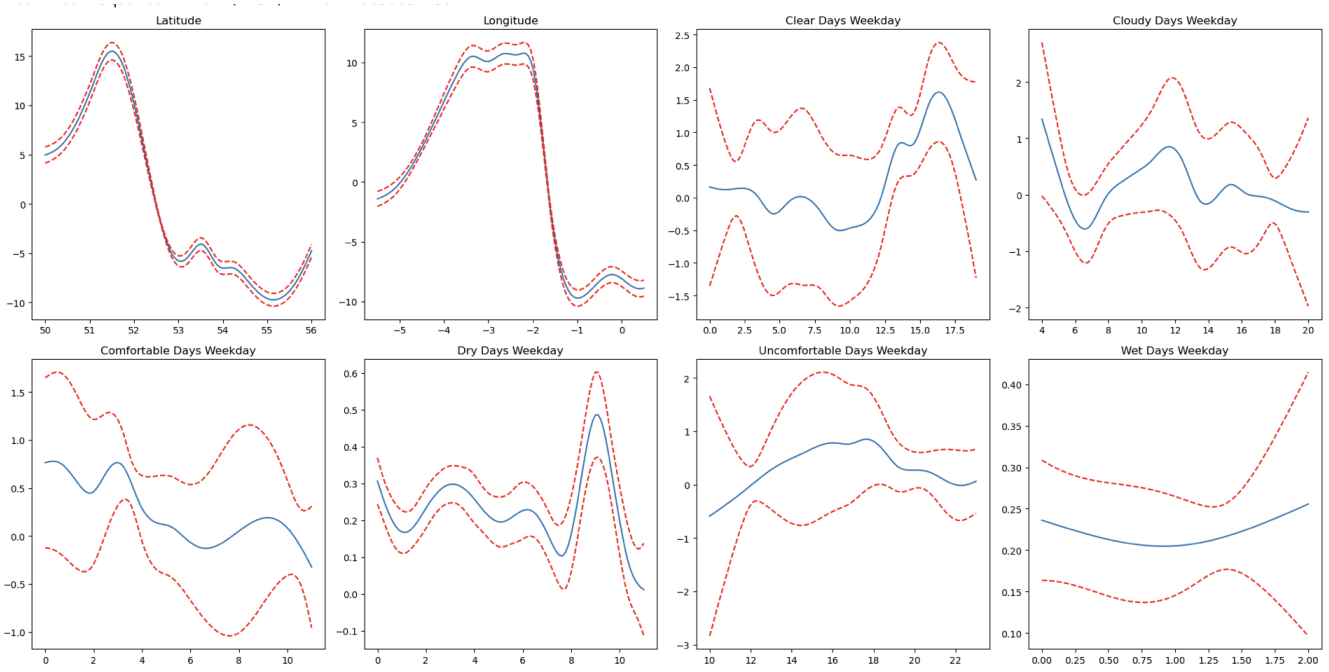


Figure 5.9: Partial Dependence Plots for Weekday Data

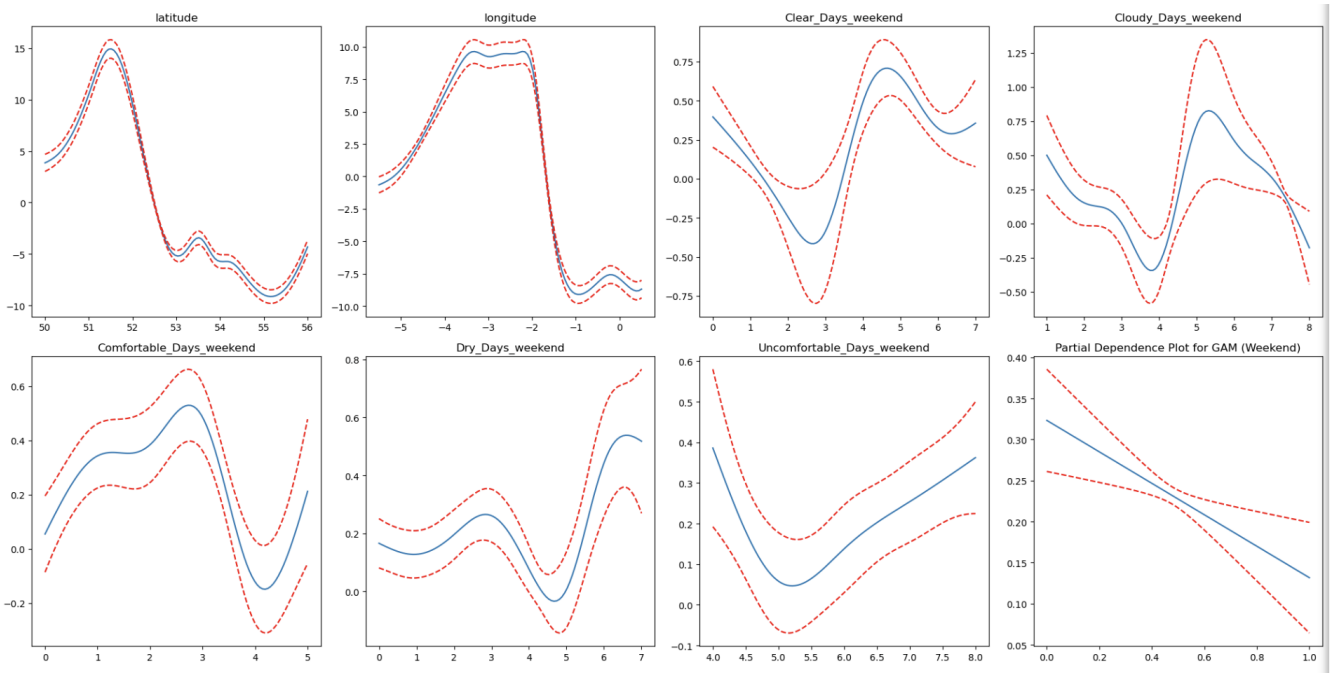


Figure 5.10: Partial Dependence Plots for Weekend Data

These models revealed that non-linearities in the relationships between predictors and visitor numbers are crucial for understanding visitor behavior, particularly the geographic and weather-related factors. The GAM models provided a more accurate depiction of these relationships, capturing complexities that linear models could not.

5.8.4 Random Forest Regression

To further delve into the complex relationships between geographic, meteorological variables, and visitor numbers, Random Forest models were applied. Unlike linear OLS regression or non-linear smoothing GAMs, Random

Forests use ensemble decision trees to capture intricate data patterns, potentially leading to more accurate predictions.

Weekday Model: The Random Forest model for weekdays explained 37.1% of the variance in visitor numbers, closely aligning with the performance of the GAM model. This model effectively captured non-linear relationships across several variables.

The Residuals vs. Fitted Values plot (Appendix Figure B.14) indicated that while the Random Forest model reduced the heteroscedasticity observed in the OLS models, some residual variance remained unexplained. This suggests that the model performed reasonably well across different levels of predicted visitor numbers, although some under- and over-predictions were still present.

Weekend Model: The Random Forest model for weekends explained 36.9% of the variance in visitor numbers, slightly lower than the weekday model but still a significant improvement over the OLS model's performance. This model captured non-linear and interaction effects between predictors more effectively than the OLS model:

The Residuals vs. Fitted Values plot for the weekend model (Appendix Figure B.15) indicated better handling of residuals compared to the OLS models, though some non-normality persisted, particularly at the extremes of the data.

Model Diagnostics: The Q-Q plots (Appendix Figures B.16 and B.17) for both the weekday and weekend models showed deviations from normality, particularly in the extremes, suggesting some persistent issues with model fit. The residuals distributions (Appendix Figures B.18 and B.19) further reflected these challenges, emphasizing the need for continued refinement of the models to better handle the data's complexity.

Weekday Model Metrics:

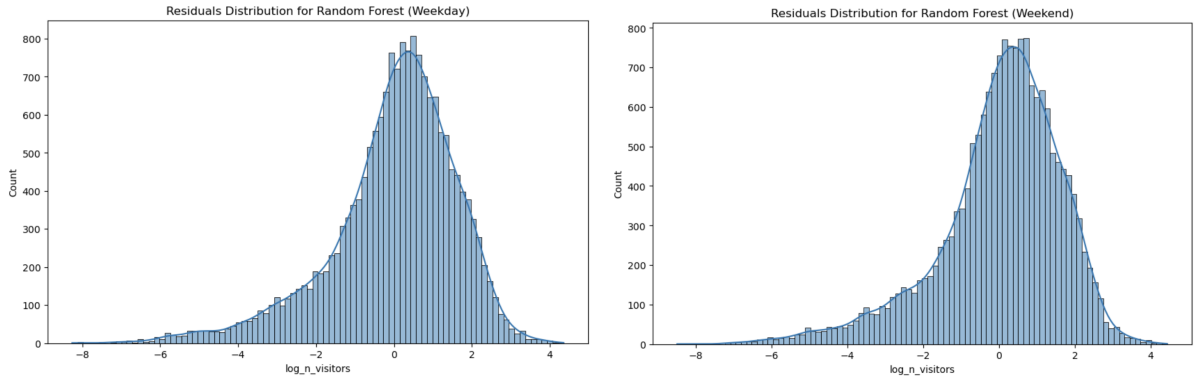
- **R-squared:** 0.3710
- **Mean Absolute Error (MAE):** 1.2206
- **Mean Squared Error (MSE):** 2.6326
- **Root Mean Squared Error (RMSE):** 1.6225

Weekend Model Metrics:

- **R-squared:** 0.3689
- **Mean Absolute Error (MAE):** 1.2317
- **Mean Squared Error (MSE):** 2.6694
- **Root Mean Squared Error (RMSE):** 1.6338

5.8.5 Gradient Boosting Regression

Gradient Boosting Models (GBMs) were employed to enhance predictive accuracy by combining multiple weak predictive models into a robust predictive model, effectively handling both linear and non-linear relationships. This



(a) Residuals Distribution (Weekday)

(b) Residuals Distribution (Weekend)

Figure 5.11: Residuals Distribution for Random Forest Model (Weekday vs Weekend)

model included a comprehensive set of predictor variables: latitude, longitude, clear days, cloudy days, uncomfortable days, comfortable days, wet days, and dry days.

The Residuals vs. Fitted Values plot (Appendix Figure) indicates that the model struggled to completely elim-inate heteroscedasticity, with residual variance increasing at higher fitted values. The Q-Q Plot (Appendix Figure) shows significant deviations from normality, particularly at the tails. The Residual Distribution (Figure 5.11) reflects these challenges, showing a distribution with heavy tails.

Weekday Model: For weekends, the Gradient Boosting Model explained 36.6% of the variance in visitor numbers. The model struggled slightly more with weekend data due to increased variability.

The Residual Distribution plot (Figure 5.11a) indicates good overall performance, though the model could still be improved in handling outliers and extreme cases.

Weekend Model: For weekends, the GBM explained 36.6% of the variance. The model struggled slightly more with weekend data due to increased variability: The Residuals vs. Fitted Values plot for the weekend model (Appendix Figure) also indicates issues with heteroscedasticity, though to a lesser extent than the weekday model. The Q-Q Plot (Appendix Figure) shows deviations from normality, particularly in the tails, indicating that the model did not fully account for extreme values. The Residual Distribution (Appendix Figure 5.11b) reflects these findings, with heavy tails and a skewed distribution.

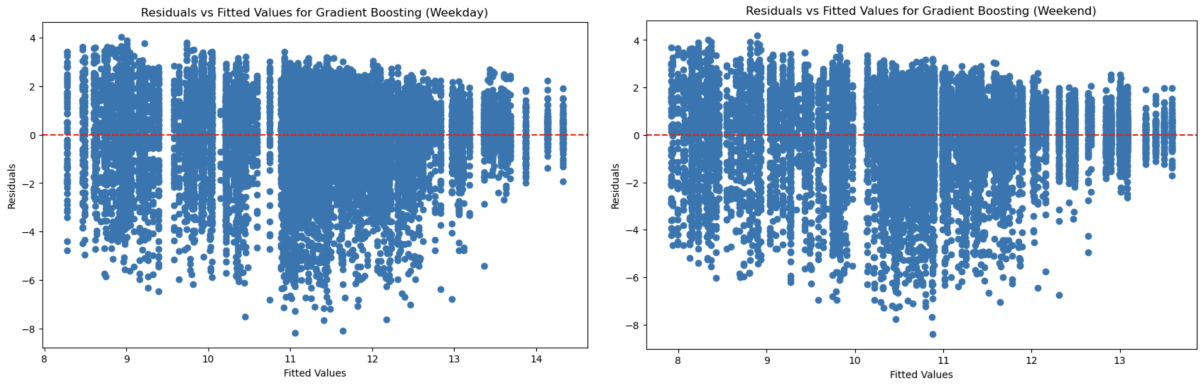
For detailed diagnostic plots, including the Residuals vs. Fitted Values, Q-Q plots, and Residual Distribution plots for both weekday and weekend models, please refer to Appendix.

5.9 Confidence Interval Estimation

5.9.1 Confidence Interval for GAM

This section discusses the process of calculating confidence intervals for the Generalized Additive Model (GAM) to assess the uncertainty around the influence of various predictors on weekend visitor numbers.

Initial Attempt at Confidence Interval Calculation: Using the pygam library, confidence intervals were initially calculated for each term in the GAM, aiming to understand the uncertainty associated with the effects of latitude, longitude, and various weather conditions on visitor numbers. However, the results were unexpected and not typical



(a) Residuals vs. Fitted values for Gradient Boosting (Weekday)

(b) Residuals vs. Fitted values for Gradient Boosting (Weekend)

Figure 5.12: Residuals vs. Fitted values for Gradient Boosting (Weekday vs Weekend)

in such analyses. The confidence intervals were identical across all predictors, suggesting that the method did not adequately capture the true variability in the model's estimates.

For example, the confidence intervals for latitude and longitude were as follows:

Latitude: Lower bound: 11.874, Upper bound: 12.113

Longitude: Lower bound: 11.874, Upper bound: 12.113

The identical bounds across different predictors indicated that the method used was insufficient for providing meaningful estimates of uncertainty.

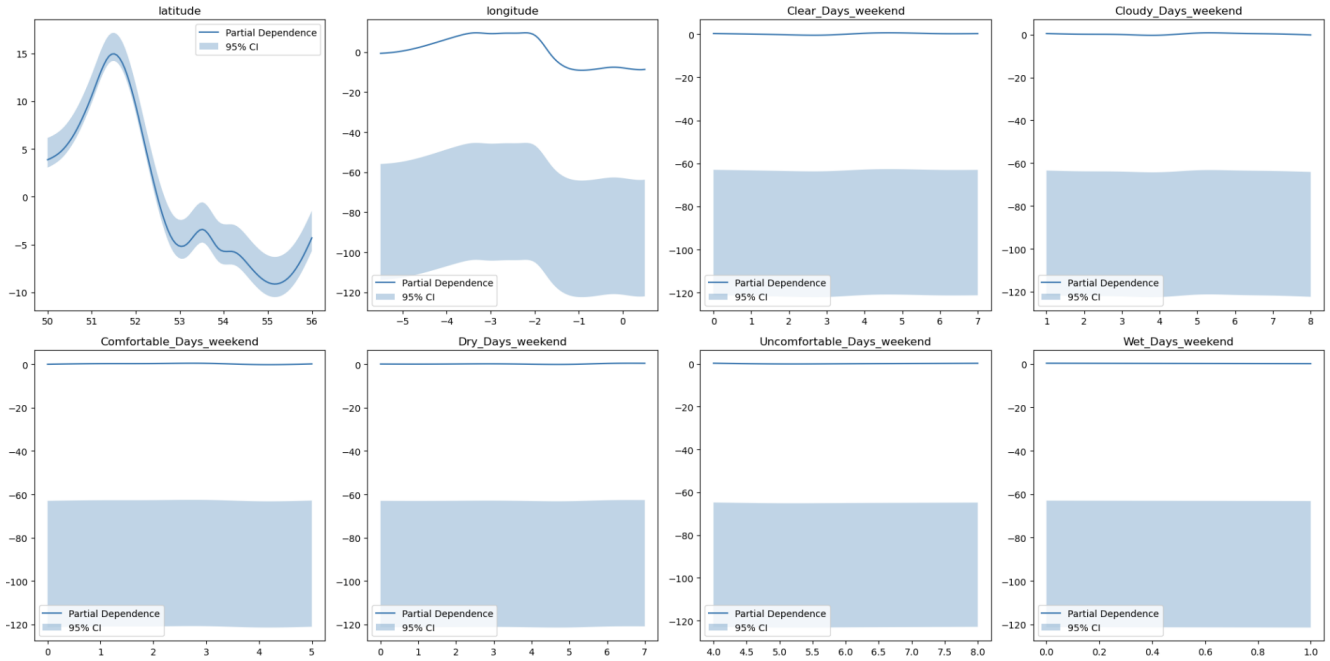


Figure 5.13: Partial Dependence Plots with 95% Confidence Intervals for Weekend Weather Conditions

Partial Dependence Plots: To further explore the effects of each predictor, partial dependence plots were generated, including a 95% confidence interval. These plots provided insights into the non-linear relationships between predictors and visitor numbers.

However, the confidence intervals depicted in these plots were overly broad, indicating a high level of uncertainty in the model's predictions. This lack of precision made the confidence intervals less useful for drawing reliable inferences about the effects of the predictors.

Next Steps: To address these issues, a bootstrap method will be employed for estimating confidence intervals. Bootstrapping involves resampling the data to generate a distribution of the estimator, which should provide more robust and accurate confidence intervals. This approach will allow for better quantification of the uncertainty around each predictor's effect, thereby improving the interpretability and reliability of the GAM for predicting visitor numbers based on weather conditions and geographical factors.

5.9.2 Bootstrap Method

To address the limitations of the initial approach, the bootstrap method was employed. The bootstrap method was chosen because it does not rely on assumptions about the distribution of predictors or residuals. By using empirical data to estimate the variability of the model parameters, the bootstrap method provides more accurate and meaningful confidence intervals, especially in complex models like GAM where traditional methods may fail. This method involved the following steps:

1. Resampling the original data with replacement to create multiple bootstrap samples.
2. For each bootstrap sample, fitting the GAM model and calculating the partial dependence for each predictor.
3. Repeating this process 1,000 times to generate a distribution of partial dependence effects for each predictor.
4. Calculating the 95% confidence intervals for each predictor from the bootstrap distributions.

5.9.3 Hypothesis Testing

The bootstrap confidence intervals for each predictor are as follows:

Predictor	**95% Confidence Interval**	**Statistical Significance**
Latitude	[0.086, 0.175]	Statistically significant
Longitude	[0.860, 1.032]	Statistically significant
Clear Days (weekend)	[0.153, 0.236]	Statistically significant
Cloudy Days (weekend)	[0.193, 0.287]	Statistically significant
Comfortable Days (weekend)	[0.211, 0.265]	Statistically significant
Dry Days (weekend)	[0.185, 0.234]	Statistically significant
Uncomfortable Days (weekend)	[0.171, 0.222]	Statistically significant
Wet Days (weekend)	[0.216, 0.239]	Statistically significant

Table 5.3: Bootstrap Confidence Intervals for Each Predictor

In this case, the bootstrap results indicated that all the predictors, including weather-related variables, had a statistically significant effect on visitor numbers during weekends. The results from this method were more reliable and interpretable, supporting our hypothesis that geographical and weather conditions significantly influence mobility patterns.

The bootstrap-derived confidence intervals confirm that the effects of all predictors are statistically significant. This strengthens our confidence in the model's results and provides a solid foundation for further hypothesis testing and data-driven decision-making.

Chapter 6

Discussion

This chapter delves into the key findings of this dissertation, reflecting on the challenges and limitations encountered during the research.

6.1 Results

6.1.1 Seasonal Trends

The seasonality analysis, depicted in Figure 5.1, reveals a distinct pattern in mobility. There is a noticeable peak during the summer months, particularly in August, followed by a decline as autumn and winter set in, from September through January. Interestingly, a sudden resurgence in mobility is observed in late winter and early spring (February and March). These trends are further illustrated in the figure 5.6 and subsequent figures within the appendix, highlighting the monthly shifts in mobility.

6.1.2 Impact of Weather Variables on Mobility Patterns

Weather variables such as Surface Solar Radiation, Precipitation, and Temperature, alongside geographic factors like Daylight Minutes, Latitude, and Longitude, play a significant role in influencing mobility. This relationship is evident from the correlation matrices and the feature importance plot presented in 5.4.

6.1.3 Trend across the UK vs. Specific Regions

In general, random forest performed the best, capturing about 35% data's variability, other models underperformed significantly, as outlined in Table 5.2. The limitations of these models may be attributed to a lack of temporal granularity, insufficient data on special events such as promotions and public holidays, and the geographic scope.

However, by narrowing the focus to five UK regions and incorporating derived weather variables like Comfortable Days, Uncomfortable Days, Wet Days, Dry Days, Clear Days, and Cloudy Days, model performance improved. In particular, Random Forest achieved an R-squared value of approximately 37% for both weekdays and weekends, while GAM also performed well, explaining about 36% of the variance. These improvements are visualized in Figures 5.9 and 5.10

6.1.4 Trend during weekdays and Weekends

The analysis shows that Dry Days, Clear Days, Comfortable Days, and Latitude positively correlate with mobility, as expected. In contrast, Wet Days, Cloudy Days, Uncomfortable Days, and Longitude tend to reduce mobility. A remarkable finding is that Uncomfortable Days significantly decrease mobility only during weekends, while on weekdays, this variable is not as statistically significant. This difference could be attributed to the fact that, during weekdays, people tend to mobilize for work as long as the weather conditions are not extreme. These insights are supported by the Bootstrap Confidence Intervals presented in Table 5.3.

6.1.5 Weather's Impact on Mobility Patterns: Statistical Hypothesis Testing

Hypothesis testing, based on the confidence intervals derived from bootstrapping samples to fit GAM models, provided answers to several key questions:

How do various meteorological conditions influence mobility patterns? Weather variables such as Precipitation, Temperature, and Solar Radiation, along with geographic factors like Daylight Minutes, Latitude, and Longitude, directly influence mobility patterns. However, refining these results could benefit from incorporating more granular variables.

How do mobility patterns differ in response to weather conditions on weekdays versus weekends? In general, mobility decreases during unfavorable weather conditions like Wet Days and Cloudy Days. However, Uncomfortable Days only significantly impact mobility during weekends, suggesting that weekday mobility, often driven by work, is less affected by weather.

How does the impact of weather on mobility differ between the UK as a whole and specific regions? The Random Forest model, as an ensemble method, provided a more accurate response than other models. However, OLS models performed poorly for the entire UK due to a lack of temporal and spatial granularity, and the absence of external variables such as public holidays and promotions. Focusing on specific regions yielded better model performance.

How effective is mobility data in improving situational awareness for weather warnings? Mobility data proved crucial in understanding the impact of weather on movement patterns. While this data is effective for weather warnings, it should be anonymized and offer greater time granularity to enhance weather-related decision-making.

6.2 Challenges Encountered

6.2.1 Data Integration Challenges

The study faced several challenges related to data aggregation, including geographic and socio-economic biases, and the integration of diverse data formats.

Integrating different data formats — such as NetCDF for weather data, CSV for mobility data, and shapefiles for mapping—presented significant processing challenges. This experience highlights the need for standardized data processing techniques capable of effectively handling diverse data sources. Future research should explore more efficient methods for data integration.

6.2.2 Location Reference Discrepancies

Mobility data used Quadkeys for location referencing, while ERA5 data used two-decimal WGS84 coordinates. Developing a function to convert Quadkeys to coordinates similar to ERA5 was essential for accurate analysis.

6.2.3 Deriving Variables to Strengthen Statistical Models

Since the weather variables cannot be used directly for each day, or week or monthly basis due to the time aggregation in mobility data, deriving new variables from the ERA5 dataset was pivotal to this research. Therefore, daylight minutes, clear vs cloudy days, comfortable vs uncomfortable days and dry vs wet days have been derived to make sense of the time aggregation.

6.3 Study Limitations

6.3.1 Geographic and Socio-Economic Biases:

The reliance on mobility data from Vodafone's network introduced potential biases, particularly in rural areas where smartphone penetration and telecommunications infrastructure may be less robust. These biases may have affected the representativeness of the findings, especially in less-connected regions. Addressing these biases in future research is critical for ensuring that mobility studies accurately reflect the broader population.

6.3.2 Weekday vs. Weekend Patterns

The analysis revealed significant differences in mobility patterns between weekdays and weekends, with weekends showing greater sensitivity to weather conditions. This finding suggests that weekend mobility is more discretionary and influenced by weather, while weekday mobility, often driven by work, is less variable. This insight is crucial for emergency preparedness and public safety strategies, which need to account for these variations in response to weather events.

6.3.3 Impact of Data Aggregation:

Data aggregation by day and location, while necessary for managing large datasets, limited the study's ability to capture finer details of mobility patterns. This likely contributed to lower R-squared values in regression models and may have obscured more nuanced trends occurring at finer temporal and spatial scales. Future studies could benefit from more granular data collection and analysis to better understand intraday and individual-level mobility behaviors.

6.3.4 Counting Unique Visitors

In the mobility data, if a person visits the same location multiple times within a specific time aggregation (e.g., within a month), they are only counted once. This approach impacts the precision of understanding how frequently and when people visit specific areas.

6.3.5 Focus on Aggregate Time Periods

The focus on broad time aggregations like "Weekend - All" and "Weekday - All" provided a general sense of mobility trends but omitted finer distinctions, such as differences between nighttime and evening mobility. Future

research could explore these finer temporal patterns to provide more detailed insights.

6.3.6 Data Granularity

The lack of time granularity — such as hourly or daily (continuous) data — limited the analysis to broader seasonal mobility trends rather than more precise weather-related trends. Additionally, the ERA5 dataset's spatial resolution, covering 0.25 degrees, may have introduced biases when averaged across smaller mobility locations. More granular data would allow for a more precise analysis.

6.3.7 Five regions selection

The selection of only five UK regions — London, Manchester, Edinburgh, St. Ives, and Windermere — provided geographic and demographic diversity but was a limited sample of the broader UK population. Including more regions or categorizing locations by geographic characteristics (e.g., urban, suburban, coastal, forest) could yield more effective models and insights.

Chapter 7

Conclusion and Future Works

7.1 Conclusion

This study has provided significant insights into how weather conditions influence mobility patterns across the UK. By examining variables such as solar radiation, precipitation, temperature, latitude, and longitude, the research underscores the considerable impact these factors have on human movement. Notably, the findings reveal a marked difference in how weather affects mobility during weekdays versus weekends, with the latter showing greater sensitivity to adverse conditions.

Through this analysis, the study contributes to a deeper understanding of the interplay between weather and daily life. These insights have practical implications for urban planning, infrastructure development, and public safety, emphasizing the need to consider weather variations when designing resilient transportation networks and emergency response strategies.

7.2 Broader Significance of Findings

The broader implications of this study extend to our understanding of human behavior in response to weather conditions. The significant influence of weather on mobility patterns highlights the importance of adaptive urban planning, resilient infrastructure, and informed public policies that can mitigate the effects of adverse weather conditions.

7.2.1 Contributions to Urban Planning

The insights gained from this study can inform urban planning efforts, particularly in designing transportation networks that are resilient to weather variations. Understanding the specific weather conditions that most significantly affect mobility can help planners prioritize infrastructure improvements and design more weather-resilient urban spaces.

7.2.2 Implications for Public Safety and Emergency Preparedness

The study's findings on the impact of weather on mobility have critical implications for public safety and emergency preparedness. By understanding how mobility patterns change in response to weather conditions, emergency

response strategies can be better tailored to ensure effective resource allocation and public communication during adverse weather events.

7.3 Directions for Future Research

While this study has made significant contributions to understanding weather-related mobility patterns, several areas remain for future exploration:

7.3.1 Expanding the Range of Variables

Initially, the models were built with only a few variables, such as Unique Visitors, Wet Days, and Dry Days, resulting in low performance. However, by incorporating additional derived weather variables, the model's performance improved significantly, with the R-squared value increasing from 3% to approximately 37%. Future research should focus on adding even more variables, particularly granular ones, to develop more robust models.

7.3.2 Granular Data Analysis

Future studies should aim to collect and analyze more granular data that captures finer temporal and spatial details. This includes exploring intraday variations in mobility and individual-level behavior, which were not fully captured in this study due to the data aggregation methods used. Such granularity would provide a deeper understanding of when and how weather most significantly affects mobility.

7.3.3 Socio-Economic Factors

The study identified potential biases due to the reliance on data from Vodafone's network, which may not fully represent the broader population. Future research should aim to integrate data from multiple sources, such as navigation services like Apple Maps, Google Maps, and social media platforms, to ensure a more representative sample. Additionally, exploring how socio-economic factors like income levels and access to transportation interact with weather conditions to influence mobility would be valuable.

7.3.4 Integration of Real-Time Mobility Data

The integration of real-time mobility data into impact-based forecasting systems presents a promising avenue for future research. By combining weather forecasts with live mobility data, researchers could develop dynamic public safety strategies that adapt in real time to changing conditions.

7.3.5 Longitudinal Studies on Mobility Trends

Longitudinal studies that track mobility trends over several years could provide insights into how long-term changes in climate and weather patterns influence mobility. This could include studying the impacts of increasingly frequent extreme weather events due to climate change and assessing how populations adapt over time.

7.3.6 Geographic Factors

While the study confirmed that geographic variables like latitude and longitude are significant, there is potential to explore more detailed geographic factors. Future research could examine how different landscapes, such as urban,

suburban, coastal, and forested areas, interact with weather conditions to affect mobility. For example, movement patterns in snowy conditions might differ significantly between cities and coastal regions.

7.3.7 Interdisciplinary Approaches

An interdisciplinary approach, combining expertise from urban planning, meteorology, data science, and public health, could greatly enhance future research. Such collaborations would lead to more comprehensive strategies for mitigating the impacts of adverse weather on mobility and public safety.

7.4 Summary of Implications and Future Directions

This study highlights the intricate relationship between weather conditions and mobility patterns across the UK, with important implications for infrastructure development, urban planning, and public safety. The findings emphasize the need for more resilient infrastructure, particularly in rural areas, and more effective emergency preparedness strategies. Future research should build on these insights, addressing the challenges identified and exploring new methods to deepen our understanding of weather-related mobility patterns.

Bibliography

Amina Aitsi-Selmi, Shinichi Egawa, Hiroyuki Sasaki, Chadia Wannous, and Virginia Murray. The sendai framework for disaster risk reduction: Renewing the global commitment to people's resilience, health, and well-being. *International Journal of Disaster Risk Science*, 6(2):164–176, 2015. doi: 10.1007/s13753-015-0050-9. URL https://doi.org/10.1007/s13753-015-0050-9.

Rudy Arthur and Hywel T. P. Williams. The human geography of twitter. *CoRR*, abs/1807.04107, 2018. URL http://arxiv.org/abs/1807.04107.

Baeldung. What are k-d trees?, 2023. URL https://www.baeldung.com/cs/k-d-trees. Accessed: 2024-09-01.

Built In. What is feature importance?, 2023. URL https://builtin.com/data-science/feature-importance#:~:text=Feature%20importance%20refers%20to%20techniques,to%20predict%20a%20certain%20variable. Accessed: 2024-09-01.

Displayr. Gradient boosting: The coolest kid on the machine learning block, 2024. URL https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/. Accessed: 2024-09-01.

European Centre for Medium-Range Weather Forecasts. Era5 reanalysis (ecmwf). https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5, 2024. Accessed: 2024-08-27.

João Paulo Figueira. Geospatial indexing with quadkeys. Published on *Towards Data Science*, https://towardsdatascience.com/geospatial-indexing-with-quadkeys-d933dff01496, 2020. Accessed: 2024-08-19.

General Data Protection Regulation (GDPR). General data protection regulation (gdpr) – official legal text, 2016. URL https://gdpr-info.eu/. Accessed: 2024-09-01.

Sara E. Harrison, Sally H. Potter, Raj Prasanna, Emma E.H. Doyle, and David Johnston. 'where oh where is the data¿: Identifying data sources for hydrometeorological impact forecasts and warnings in aotearoa new zealand. *International Journal of Disaster Risk Reduction*, 66:102619, 2021. ISSN 2212-4209. doi: https://doi.org/10.1016/j.ijdrr.2021.102619. URL https://www.sciencedirect.com/science/article/pii/S221242092100580X.

Sara E. Harrison, Sally H. Potter, Raj Prasanna, Emma E.H. Doyle, and David Johnston. 'sharing is caring': A socio-technical analysis of the sharing and governing of hydrometeorological hazard, impact, vulnerability, and exposure data in aotearoa new zealand. *Progress in Disaster Science*, 13:100213, 2022. ISSN 2590-

0617. doi: https://doi.org/10.1016/j.pdisas.2021.100213. URL https://www.sciencedirect.com/science/article/pii/S2590061721000739.

IBM. Random forest, 2024. URL https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,Decision%20trees. Accessed: 2024-09-01.

Maximilian Leodolter and Christian Rudloff. Learning mobility profiles: an application to a personalised weather warning system. Technical report, AIT Austrian Institute of Technology GmbH, Vienna, Austria, 2023.

Analysys Mason. Vodafone and three merger, 2023. URL https://www.analysysmason.com/contentassets/7ed39076831d471ab600fb70b03da822/analysys_mason_vodafone_three_merger_jun2023_rdmm0_rdmb0_rdmv0_rma18.pdf. Accessed: 2024-08-19.

Bruno Merz, Christian Kuhlicke, Michael Kunz, Massimiliano Pittore, Andrey Babeyko, David N. Bresch, Daniela I. V. Domeisen, Frauke Feser, Inga Koszalka, Heidi Kreibich, Florian Pantillon, Stefano Parolai, Joaquim G. Pinto, Heinz Jürgen Punge, Eleonora Rivalta, Kai Schröter, Karen Strehlow, Ralf Weisse, and Andreas Wurpts. Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics*, 58(4):e2020RG000704, 2020. doi: https://doi.org/10.1029/2020RG000704. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020RG000704. e2020RG000704 2020RG000704.

Met Office. Climate change in the uk, 2024a. URL https://www.metoffice.gov.uk/weather/climate-change/climate-change-in-the-uk#:~:text=Increased%20rainfall%20and%20risk%20of%20flooding&text=In%20winter%2C%20it%20could%20increase,to%20issue%20flash%20flood%20alerts. Accessed: 2024-09-01.

Met Office. Weather warnings guide, 2024b. URL https://www.metoffice.gov.uk/weather/guides/warnings. Accessed: 2024-08-19.

Microsoft Bing. Bing maps tile system. https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system, 2024. Accessed: 2024-08-19.

Joanne Robbins, Isabelle Ruin, Brian Golding, Rutger Dankers, John Nairn, and Sarah Millington. *Connecting Hazard and Impact: A Partnership between Physical and Human Science*, pages 115–147. Springer International Publishing, Cham, 2022. ISBN 978-3-030-98989-7. doi: 10.1007/978-3-030-98989-7_5. URL https://doi.org/10.1007/978-3-030-98989-7_5.

Joanne Robbins, Helen Roberts, and Mark Harrison. Using mobility data from vodafone to improve upon our warnings. Confidential report, not publicly available, March 2023.

M. Spruce, R. Arthur, and H. T. P. Williams. Using social media to measure impacts of named storm events in the united kingdom and ireland. *Meteorological Applications*, 27(1):e1887, 2020. doi: https://doi.org/10.1002/met.1887. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.1887.

Dan Suri and Paul A. Davies. A decade of impact-based nswws warnings at the met office. *The European Forecaster*, 31:30–36, November 2021.

Towards Data Science. Generalised additive models, 2024. URL https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a. Accessed: 2024-09-01.

A. Tupper, C.J. Fearnley, and I. Kelman. Translating warnings into actions: How we can improve early warning systems to protect communities. Technical report, UCL Warning Research Centre, London, 2023. URL https://www.ucl.ac.uk/sts/sites/sts/files/ucl_report_-_translating_warnings_into_actions_april2023.pdf.

University of Pennsylvania. How to preserve leading zeros in csv files, 2023. URL https://provider.www.upenn.edu/computing/da/bo/webi/qna/iv_csvLeadingZeros.html. Accessed: 2024-09-01.

Lucy Verasamy. Weather and the nhs: Preparing for the unexpected, 2023. URL https://www.england.nhs.uk/blog/lucy-verasamy/. Accessed: 2024-09-01.

Vodafone UK. Vodafone analytics privacy policy, 2024. URL https://www.vodafone.co.uk/privacy/vodafone-analytics. Accessed: 2024-08-29.

World Meteorological Organization. *WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services*. World Meteorological Organization, Geneva, Switzerland, wmo-no. 1150, 2nd ed. edition, 2022. URL https://library.wmo.int/records/item/54669-wmo-guidelines-on-multi-hazard-impact-based-forecast-and-warning-service

World Meteorological Organization. *WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services*. World Meteorological Organization, Geneva, Switzerland, wmo-no. 1150, 3rd ed. edition, 2023. URL https://library.wmo.int/records/item/57739-wmo-guidelines-on-multi-hazard-impact-based-forecast-and-warning-service?language_id=13&back=&offset=.

Faye Wyatt, Joanne Robbins, and Rebecca Beckett. Investigating bias in impact observation sources and implications for impact-based forecast evaluation. *International Journal of Disaster Risk Reduction*, 90:103639, 2023. ISSN 2212-4209. doi: https://doi.org/10.1016/j.ijdrr.2023.103639. URL https://www.sciencedirect.com/science/article/pii/S221242092300119X.

Appendix A

Notes

Conversion of Quadkeys to Latitude and Longitude Coordinates

To ensure compatibility between the ERA5 reanalysis data and the mobility data, it was necessary to convert the quadkey coordinates to geographic coordinates (latitude and longitude) in the WGS84 coordinate system. The centroid of each quadkey tile was computed by deriving the latitudinal and longitudinal bounds of the tile. The centroid is calculated as the average of the northernmost, southernmost, easternmost, and westernmost bounds of the tile.

Let the geographic bounds of the tile be:

$$\text{latitude}_{\text{south}},\ \text{latitude}_{\text{north}},\ \text{longitude}_{\text{west}},\ \text{longitude}_{\text{east}}$$

The formula for computing the centroid is given by:

$$(\phi_{\text{centroid}}, \lambda_{\text{centroid}}) = \left(\frac{\text{latitude}_{\text{south}} + \text{latitude}_{\text{north}}}{2}, \frac{\text{longitude}_{\text{west}} + \text{longitude}_{\text{east}}}{2} \right)$$

Where, ϕ_{centroid} is the latitude of the centroid and $\lambda_{\text{centroid}}$ is the longitude of the centroid.

Conversion to Standard Time

The time values in the dataset are represented as the number of hours since January 1, 1900, 00:00:00 UTC. To convert these values into a standard datetime format, we use the following formula:

Given that t_{hours} is the time in hours since January 1, 1900, 00:00:00 UTC, the standard datetime T can be calculated as:

$$T = T_{\text{base}} + \frac{t_{\text{hours}}}{24}$$

Where:

- T_{base} is January 1, 1900, 00:00:00 UTC.

- t_{hours} is the time in hours since T_{base}.
- $\frac{t_{\text{hours}}}{24}$ converts the time from hours to days.

Example:

If $t_{\text{hours}} = 87600$, the corresponding datetime is:

$$T = \text{January 1, 1900} + \frac{87600}{24} = \text{January 1, 1900} + 3650\,\text{days} = \text{January 1, 1910, 00:00:00 UTC}$$

After the conversion, this new standard datetime will be stored in a column labeled `standard_time`, allowing us to continue with further data processing and analysis.

A.1 Correlation Analysis

Before delving into detailed statistical modeling, it is essential to examine the relationships between the variables. Correlation analysis serves as a foundational step, allowing us to identify the strength and direction of the relationships between mobility patterns and meteorological variables. This preliminary analysis helps determine which variables are worth further investigation in terms of statistical significance and trend analysis.

The correlation coefficient, often denoted by r, measures the degree of linear relationship between two variables. The value of r ranges from -1 to +1:

- $r = +1$ indicates a perfect positive correlation, where increases in one variable are associated with increases in another.
- $r = -1$ indicates a perfect negative correlation, where increases in one variable correspond to decreases in another.
- $r = 0$ indicates no linear relationship between the variables.

The formula for Pearson's correlation coefficient (Berman, 2016) is given by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual data points of variables X and Y,
- $\bar{x}$ and $\bar{y}$ are the means of X and Y,
- The numerator represents the covariance between the two variables, while the denominator normalizes the covariance by the standard deviations of both variables.

By calculating the correlation matrix for our dataset, we can visually inspect how different weather variables such as precipitation, temperature, and wind speed correlate with the number of visitors. This step helps identify which weather factors have the strongest influence on mobility patterns, thus guiding the subsequent statistical analysis.

The correlation analysis was conducted using Python's pandas library. The `corr()` function computes the correlation matrix for the entire dataset with just a single line of code:

```
correlation_matrix = dataset.corr()
```

This matrix provides a clear overview of how different variables are interrelated, allowing us to proceed with more focused statistical models.

A.1.1 Correlation between Meteorological Variables and Mobility Data across the UK

The table outlines the correlation score of each weather variables against unique visitors.

Variable	**Value**
Solar Radiation (Total)	0.185252
Surface Solar Radiation (Total)	0.183180
Longitude	0.160136
Temperature (Mean)	0.145444
Dew Point Temperature (Mean)	0.134103
Snow Depth (Total)	0.128900
Daylight Minutes (Mean)	0.124793
Surface Pressure (Mean)	0.110303
Wind Speed U Component (Mean)	0.040563
Precipitation (Total)	0.021735
Total Precipitation Rate (Sum)	0.021729
Wind Speed V Component (Mean)	0.001926
Wind Gust (Mean)	-0.056871
Max Wind Gust (Mean)	-0.058310
Snowfall (Total)	-0.058909
Snow Density (Total)	-0.062598
Latitude	-0.150258
Surface Thermal Radiation (Total)	-0.152499

Table A.1: Variables and Their Corresponding Values

A.1.2 Correlation between Meteorological Variables and Mobility Data in Specific Regions (Rural & Urban)

Table A.2: Correlation Coefficients between Visitors and Meteorological Variables

Variable	Correlation Coefficient (r)
Visitors (n)	1.000000
Snow Depth (Total)	0.710376
Surface Solar Radiation (Total)	0.667890
Solar Radiation (Total)	0.667352
Precipitation (Total)	0.493336
Surface Pressure (Mean)	0.303037
Total Precipitation Rate (Sum)	0.267597
Wind Gust (Mean)	0.249237
Max Wind Gust (Mean)	0.247608
Temperature (Mean)	0.196089
Dew Point Temperature (Mean)	0.146510
Wind Speed U Component (Mean)	0.004638
Wind Speed V Component (Mean)	-0.015094
Snowfall (Total)	-0.027876
Snow Density (Total)	-0.068117
Surface Thermal Radiation (Total)	-0.351960

Generating Additional Weather Variables

Variable	Weekday (Score)	Weekend (Score)
n_visitors	1.000000	1.000000
longitude	0.168898	0.170050
Comfortable_Days	0.163415	0.135880
Clear_Days	0.151824	0.130863
Dry_Days	0.084622	0.028290
qk_id	0.017015	0.020400
Uncomfortable_Days	0.015739	-0.012007
Wet_Days	-0.052924	-0.038430
Cloudy_Days	-0.090475	-0.095138
latitude	-0.157663	-0.161185

Table A.3: Correlation of Variables with Number of Visitors during Weekdays and Weekends

Appendix B

Statistical Modelling: Urban and Rural Locations

B.1 OLS Regression Model Analysis

Weekday OLS Model Analysis

The following plots provide a detailed examination of the weekday OLS model:

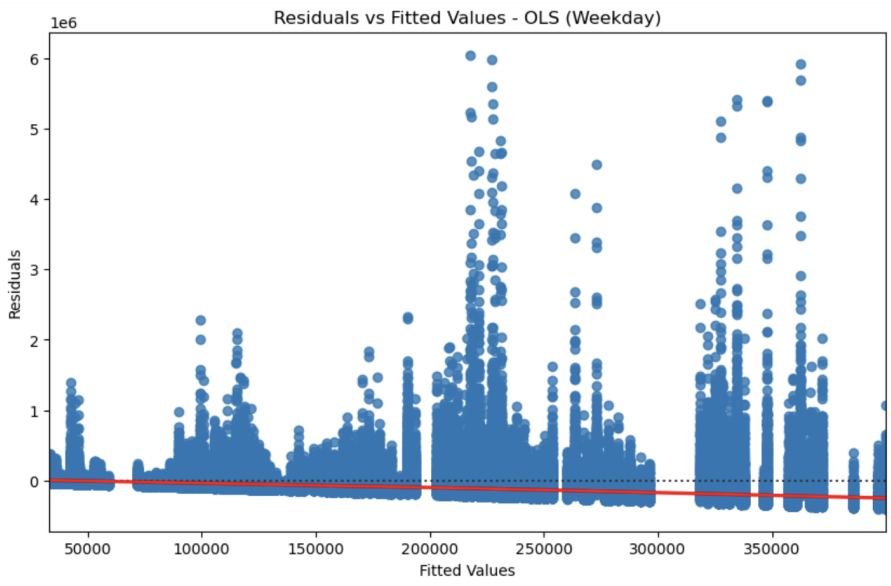


Figure B.1: Residuals vs. Fitted Values (Weekday)

This plot reveals the presence of heteroscedasticity in the weekday model, with increasing residual spread at higher fitted values.

The Q-Q plot illustrates significant deviations from normality, particularly in the upper quantiles, suggesting non-normality in the residuals.

The residual distribution shows a skewed pattern, further confirming the model's limitations in capturing the true distribution of residuals.

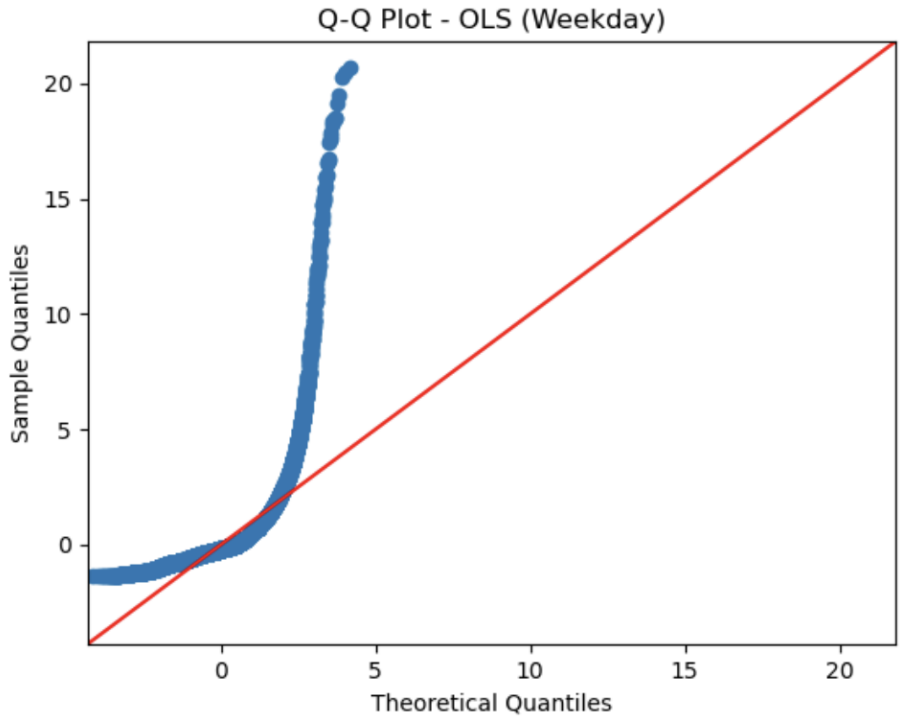


Figure B.2: Q-Q Plot (Weekday)

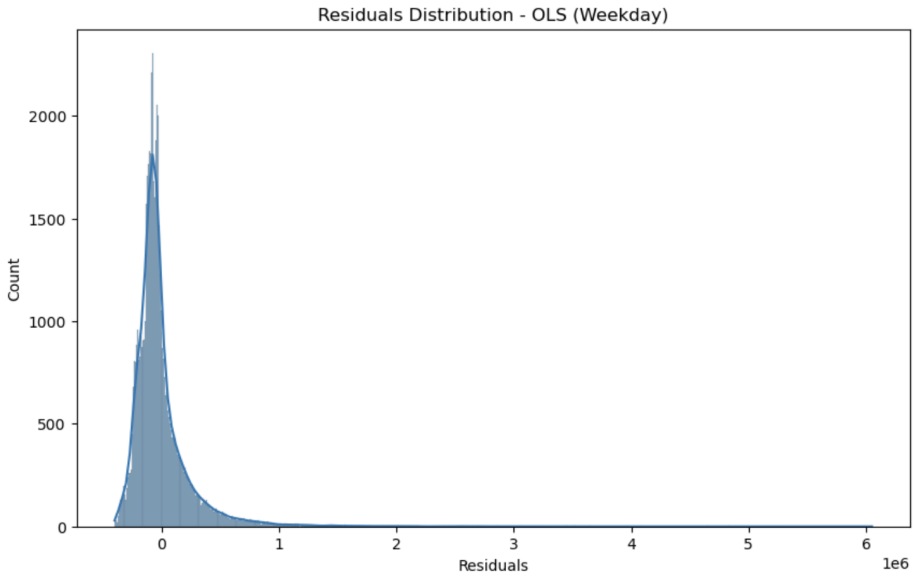


Figure B.3: Residual Distribution (Weekday)

Appendix B.1.2: Weekend OLS Model Analysis

For the weekend model, similar diagnostics were performed:

The weekend model's residuals vs. fitted values plot shows even more pronounced heteroscedasticity compared to the weekday model.

The Q-Q plot for the weekend model also highlights substantial deviations from normality, indicating potential issues with the linear assumptions of the model.

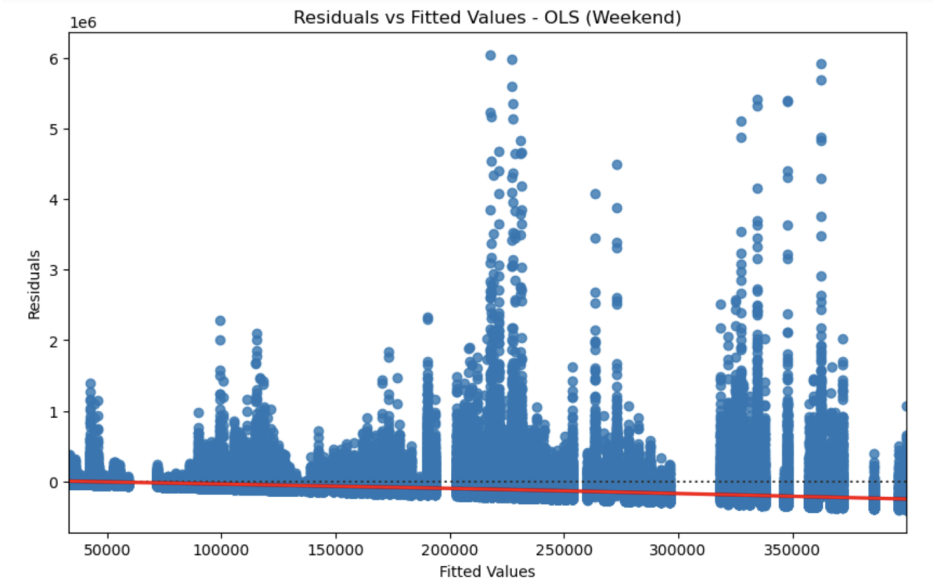


Figure B.4: Residuals vs. Fitted Values (Weekend)

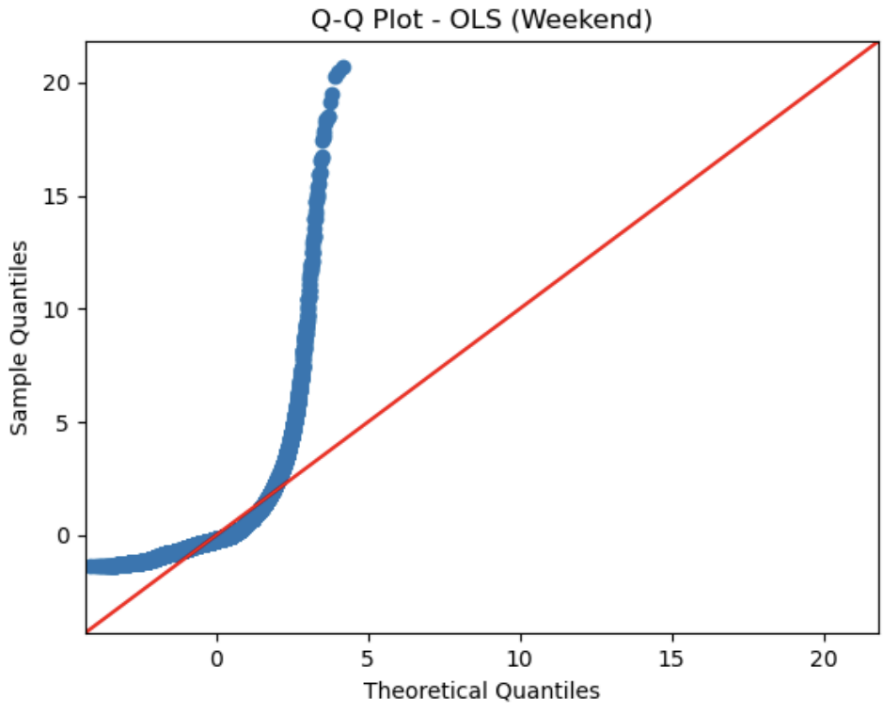


Figure B.5: Q-Q Plot (Weekend)

The residual distribution for the weekend model is similarly skewed, underscoring the need for more advanced modeling techniques to better capture the underlying patterns.

Appendix B.1.3: OLS Regression Results Summary

The tables below summarize the key metrics from the OLS regression models:

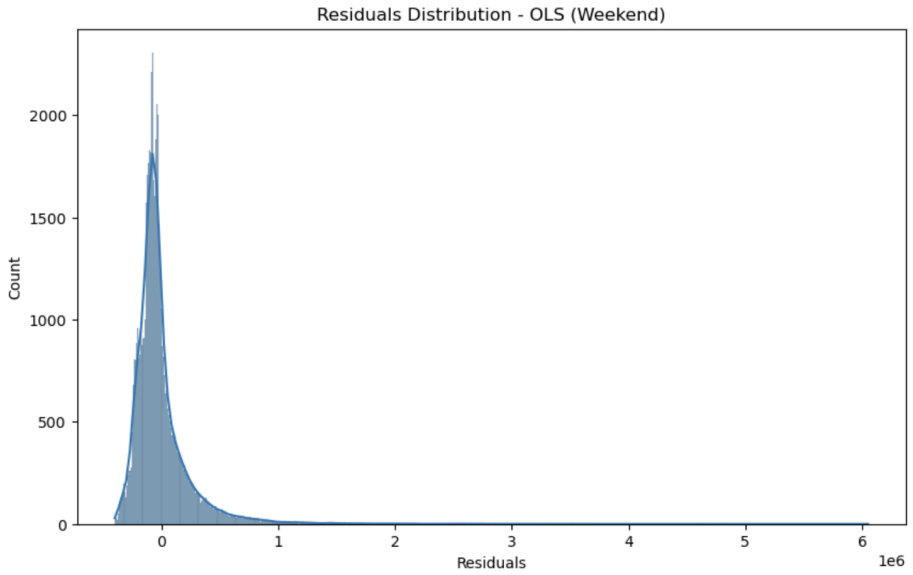


Figure B.6: Residual Distribution (Weekend)

Metric	Weekday	Weekend
R-squared	0.096	0.087
Adjusted R-squared	0.096	0.086
F-statistic	790.5	701.2
Prob (F-statistic)	0.00	0.00
Key Predictors	Latitude, Longitude, Wet Days	Latitude, Longitude, Wet Days
Significant Issues	Heteroscedasticity, Non-normality of residuals	Heteroscedasticity, Non-normality of residual

Table B.1: Summary of OLS Regression Results for Weekday and Weekend Models

B.2 OLS Regression Models: Log-Transformed Regression

B.2.1 Weekday Model

To address heteroscedasticity, a log transformation of the visitor numbers was applied. The log-transformed OLS regression for weekdays explained 13.2% of the variance. Although the transformation improved the fit slightly, significant deviations from normality remained, as indicated by the Q-Q plot. The residuals vs. fitted values plot continued to show patterns of heteroscedasticity, highlighting persistent complexities that the model could not capture.

B.2.2 Weekend Model

The log-transformed model for the weekend explained 15.4% of the variance. Despite these improvements, the model still struggled with heteroscedasticity and non-normal residuals, particularly in capturing variability during weekends, as shown in the Q-Q plot and residuals vs. fitted values plot.

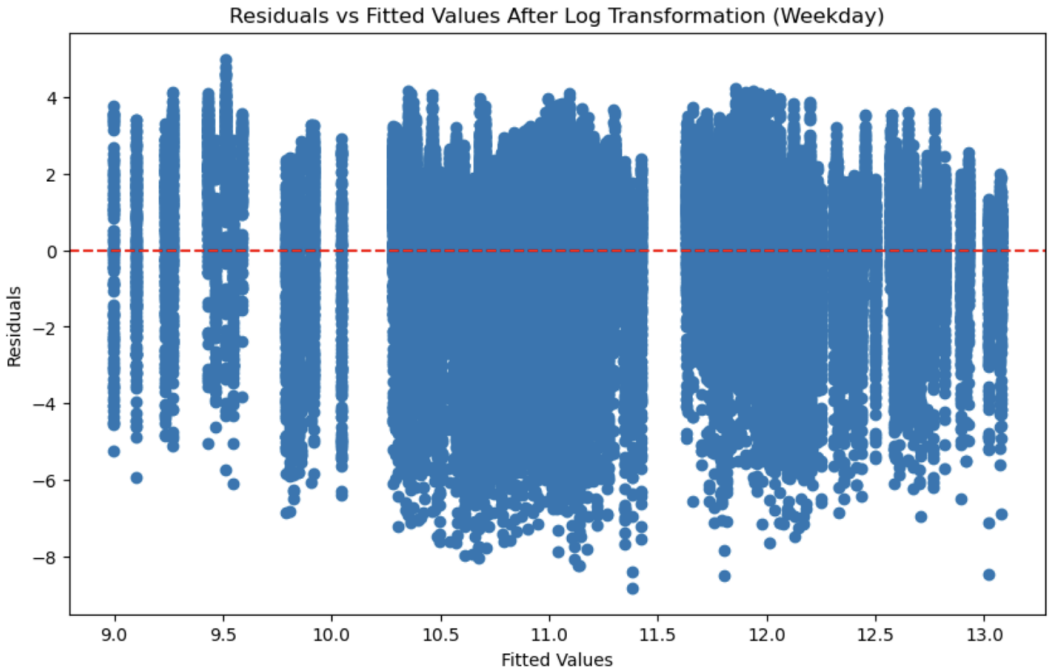


Figure B.7: Residuals vs. Fitted Values After Log Transformation (Weekday)

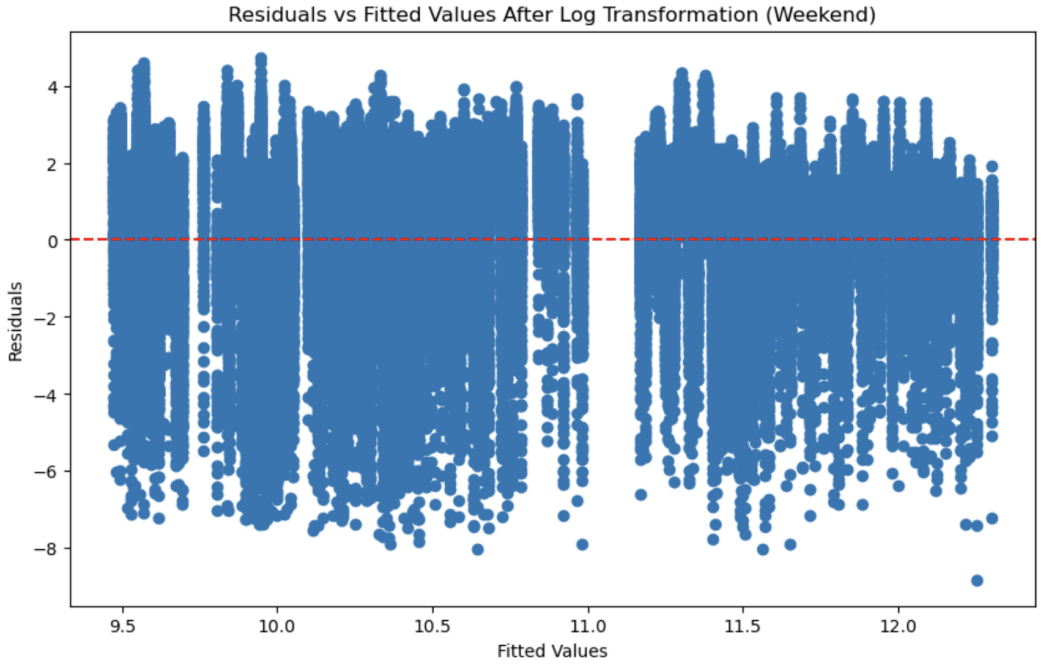


Figure B.8: Residuals vs. Fitted Values After Log Transformation (Weekend)

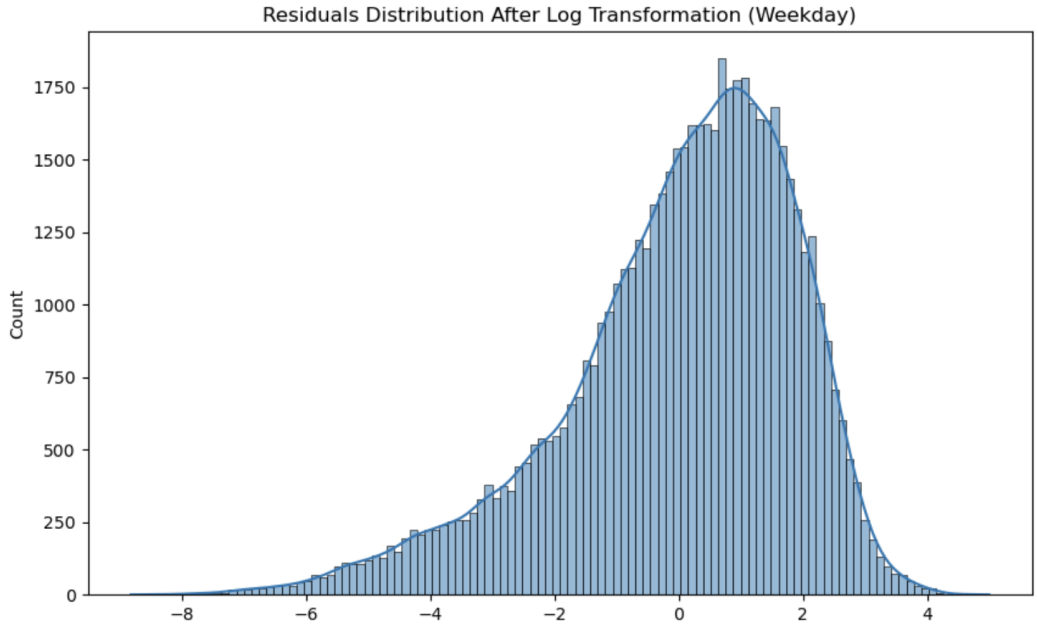


Figure B.9: Residuals Distribution After Log Transformation (Weekday)

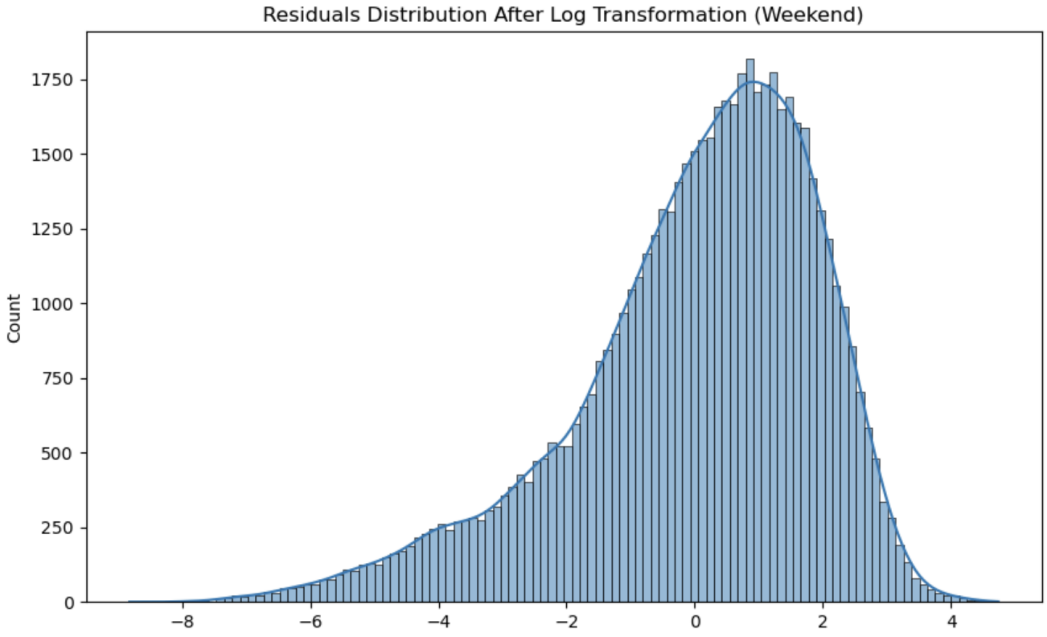


Figure B.10: Residuals Distribution After Log Transformation (Weekend)

B.1.1 Generalized Additive Models (GAM) Analysis

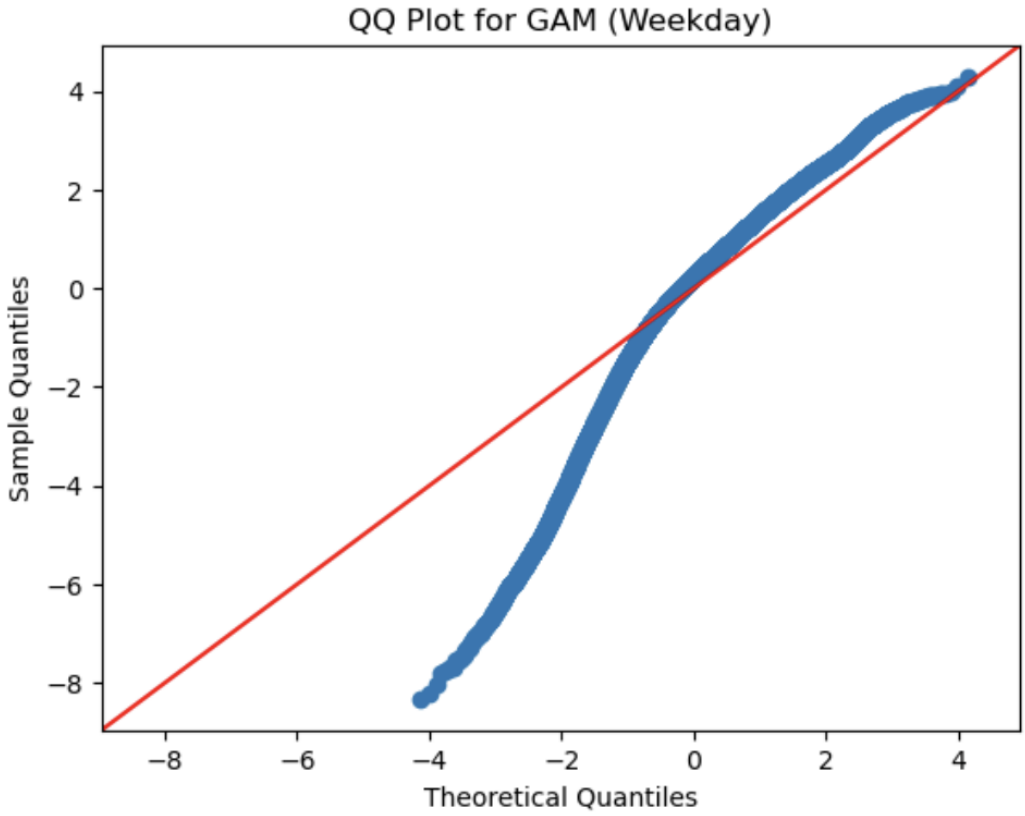


Figure B.11: QQ Plot for GAM (Weekday)

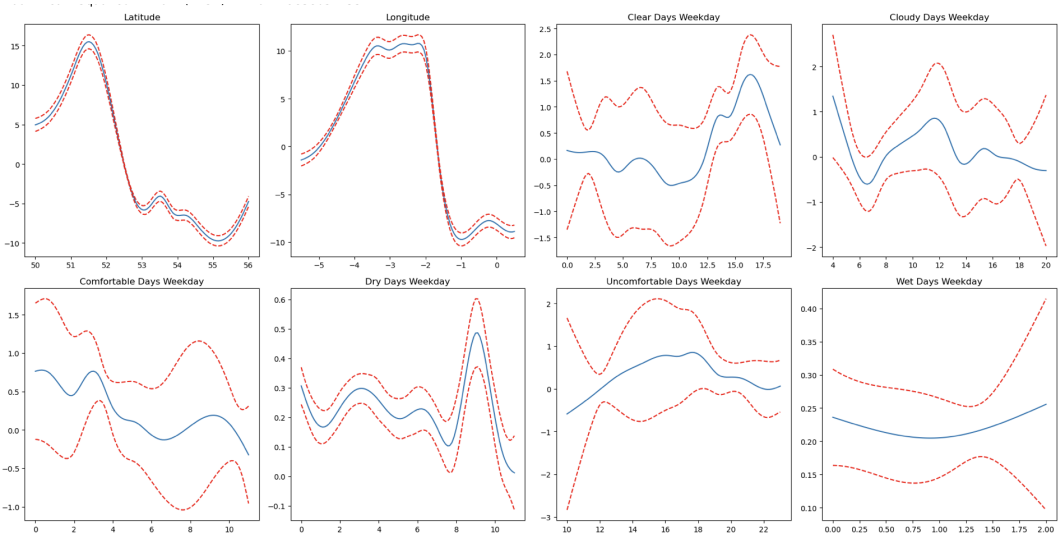


Figure B.12: PDP for GAM (Weekday)

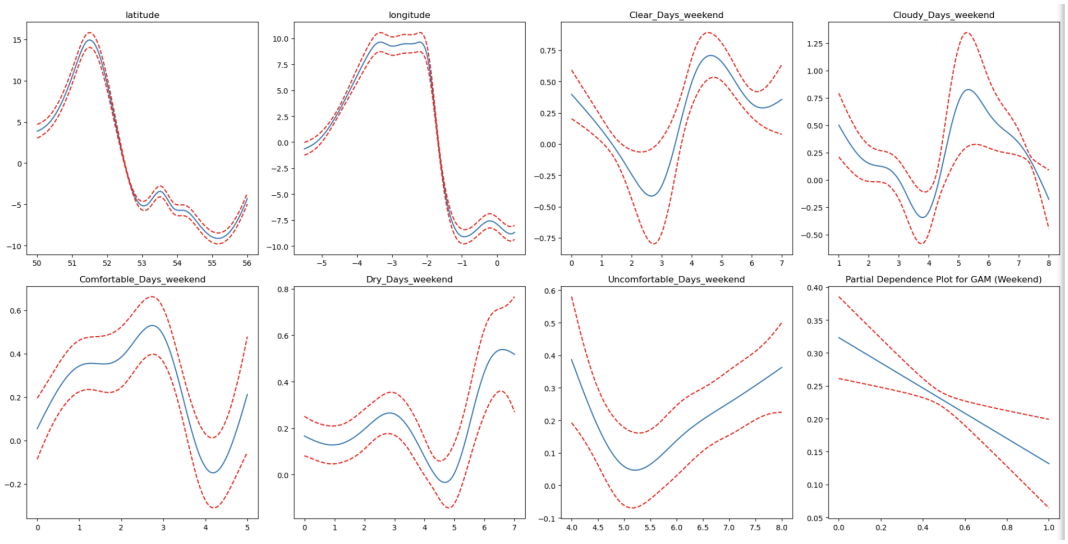


Figure B.13: PDP for GAM (Weekend)

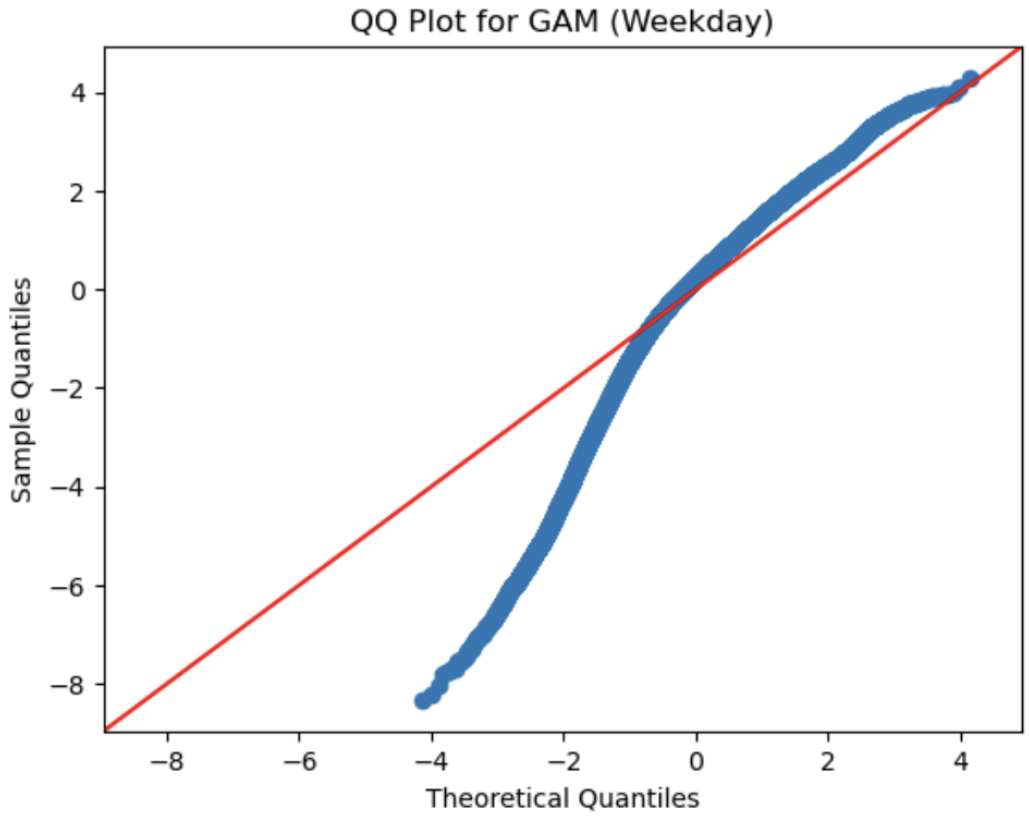


Figure B.14: QQ Plot for GAM (Weekday)

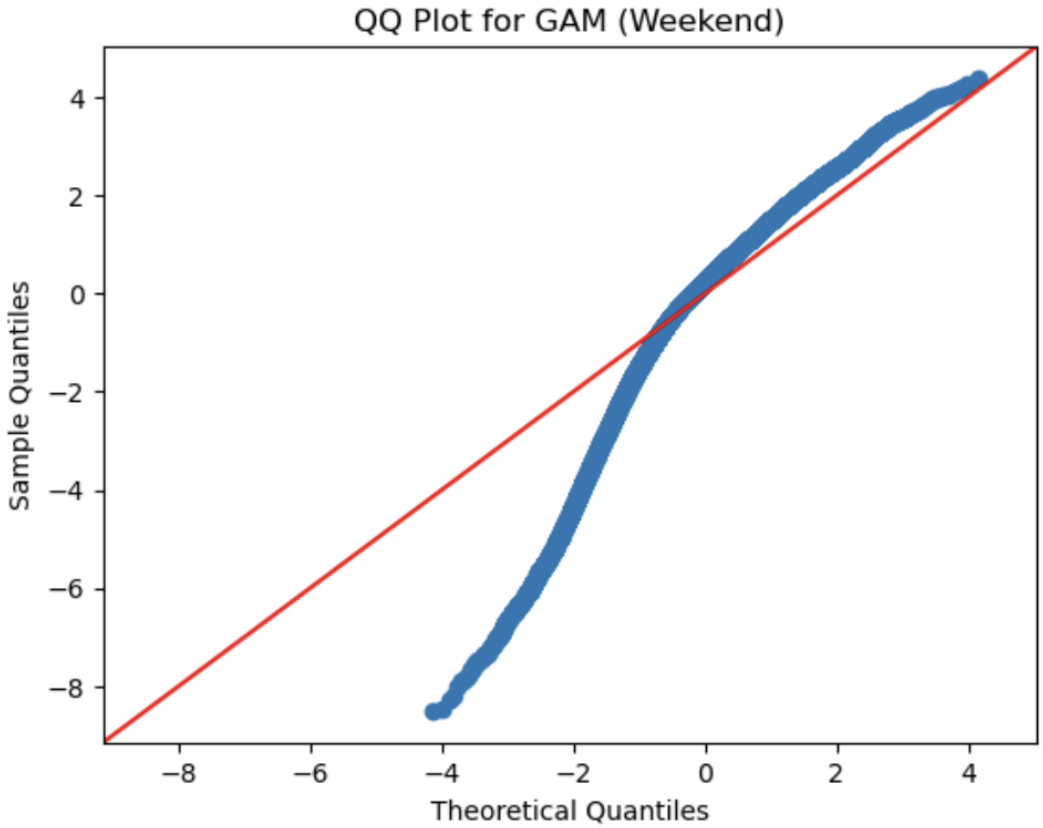


Figure B.15: QQ Plot for GAM (Weekend)

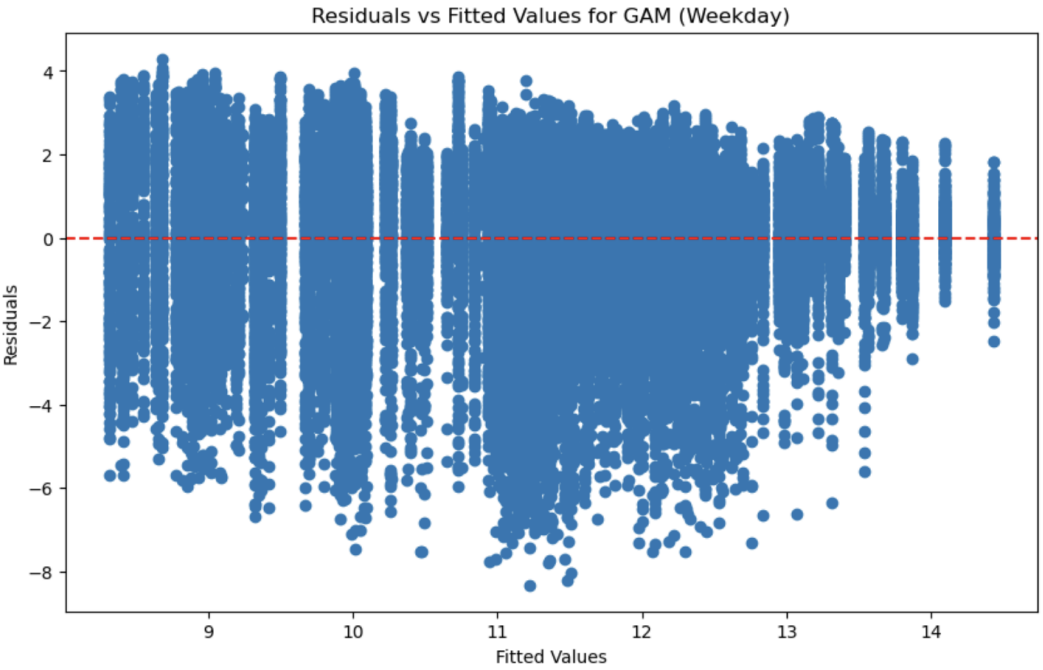


Figure B.16: Residuals vs Fitted Values for GAM (Weekday)

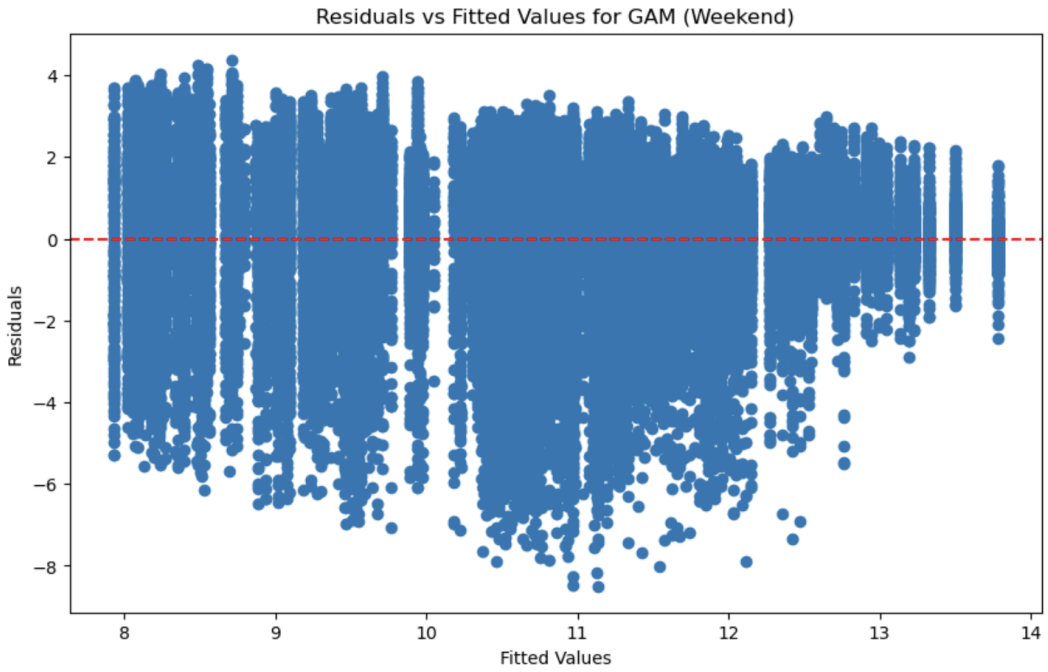


Figure B.17: Residuals vs Fitted Values for GAM (Weekend)

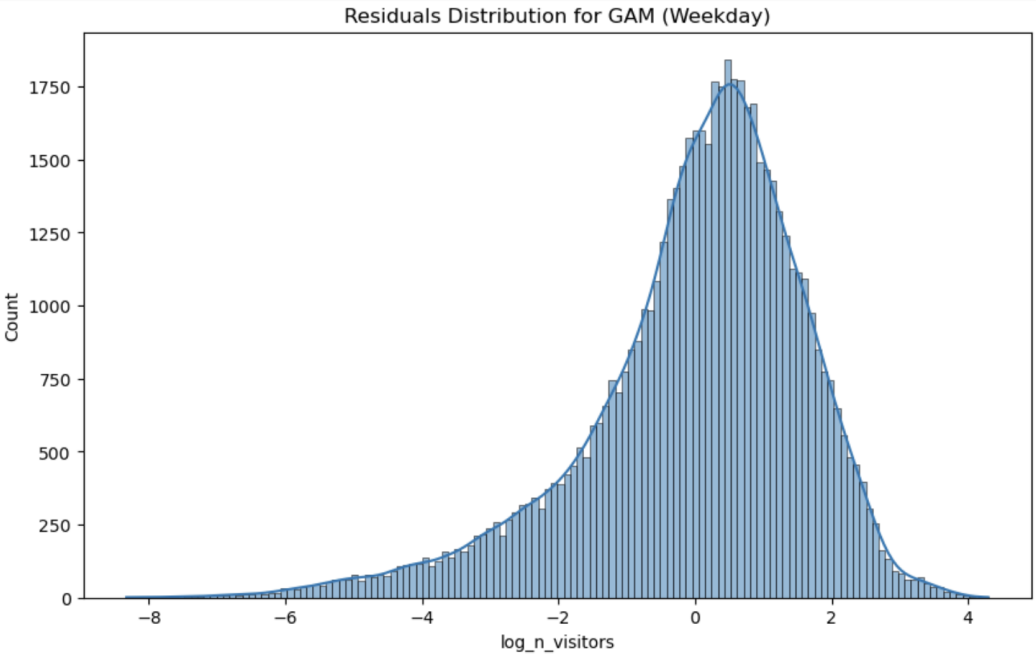


Figure B.18: Residuals Distribution for GAM (Weekday)

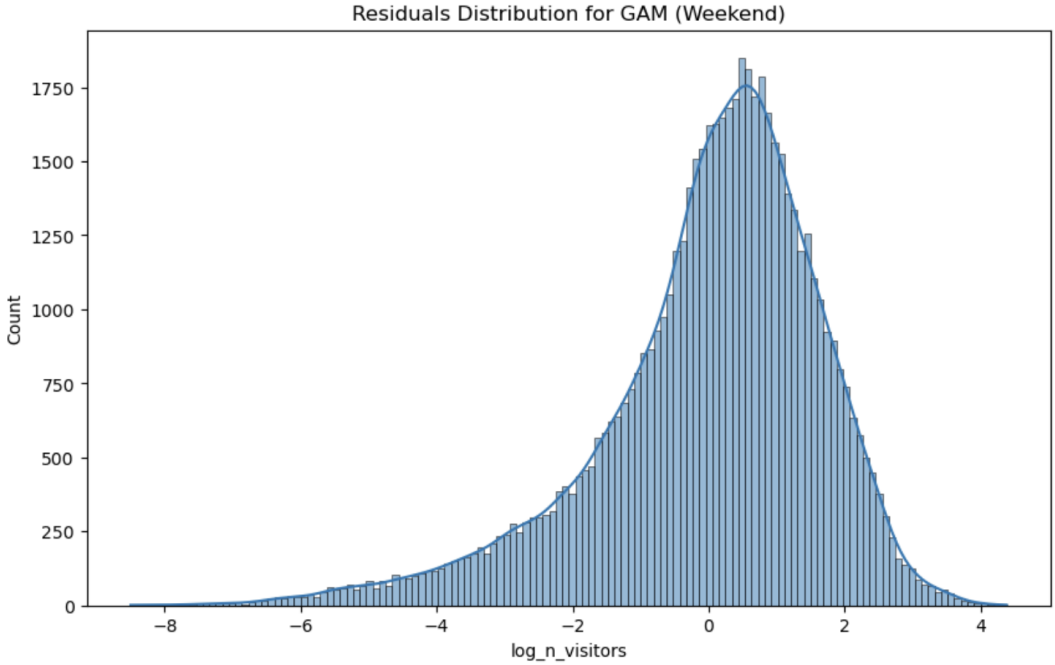


Figure B.19: Residuals Distribution for GAM (Weekend)

B.1.2 Random Forest Regression Analysis

Weekday Model Diagnostics

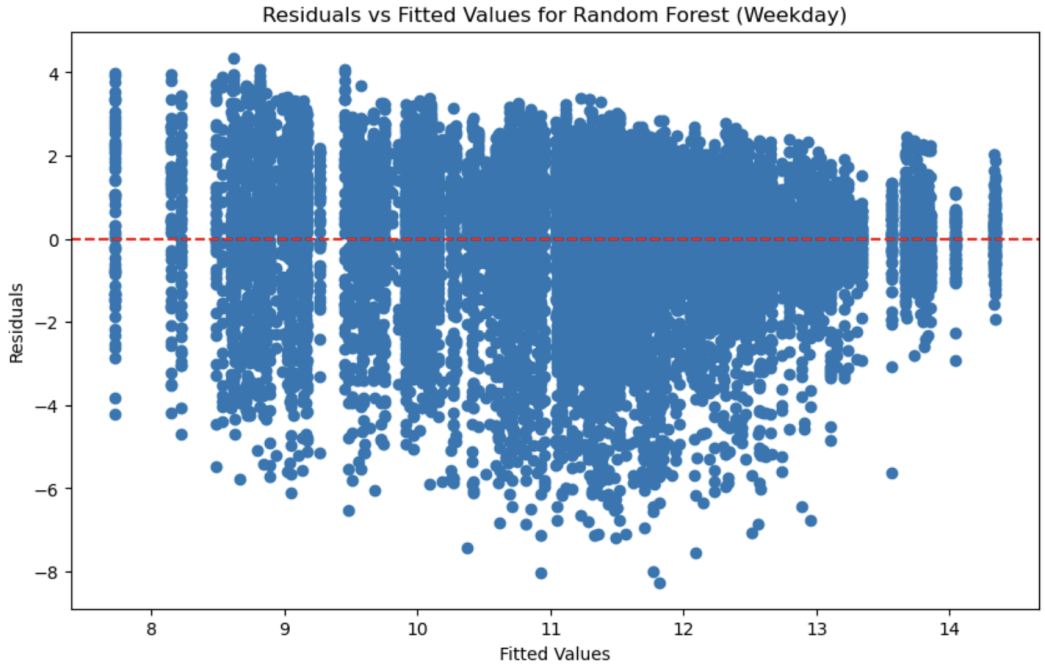


Figure B.20: Residuals vs. Fitted Values - Weekday Random Forest

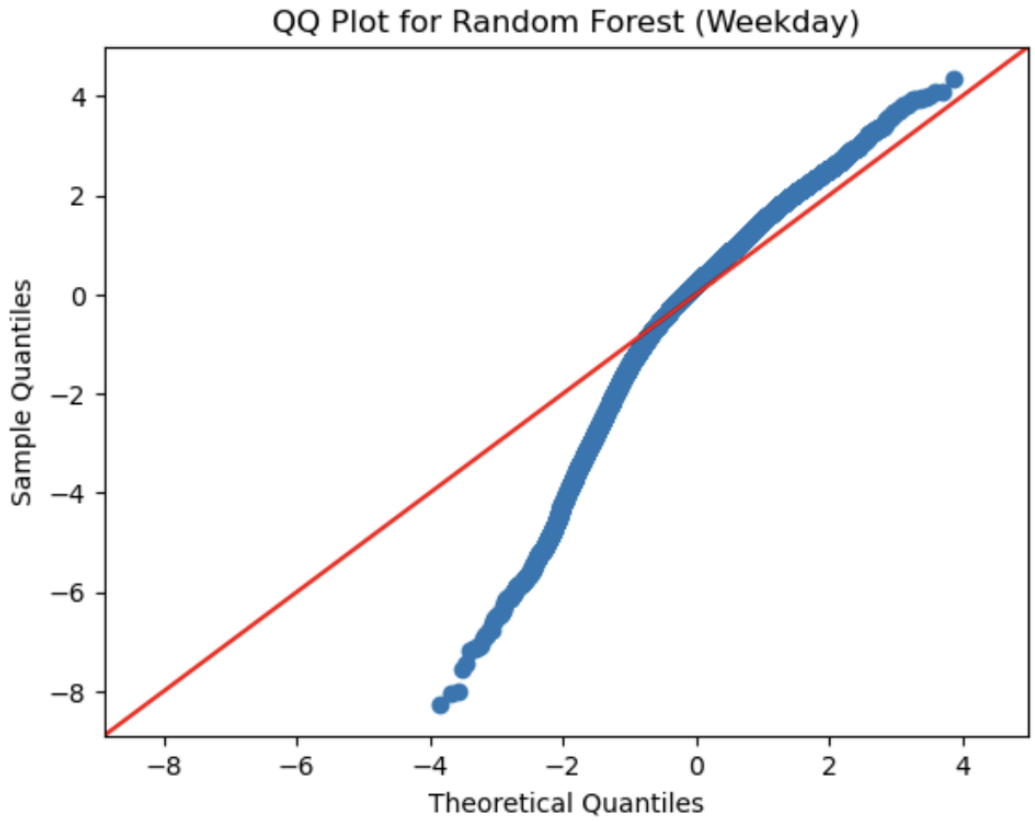


Figure B.21: Q-Q Plot - Weekday Random Forest

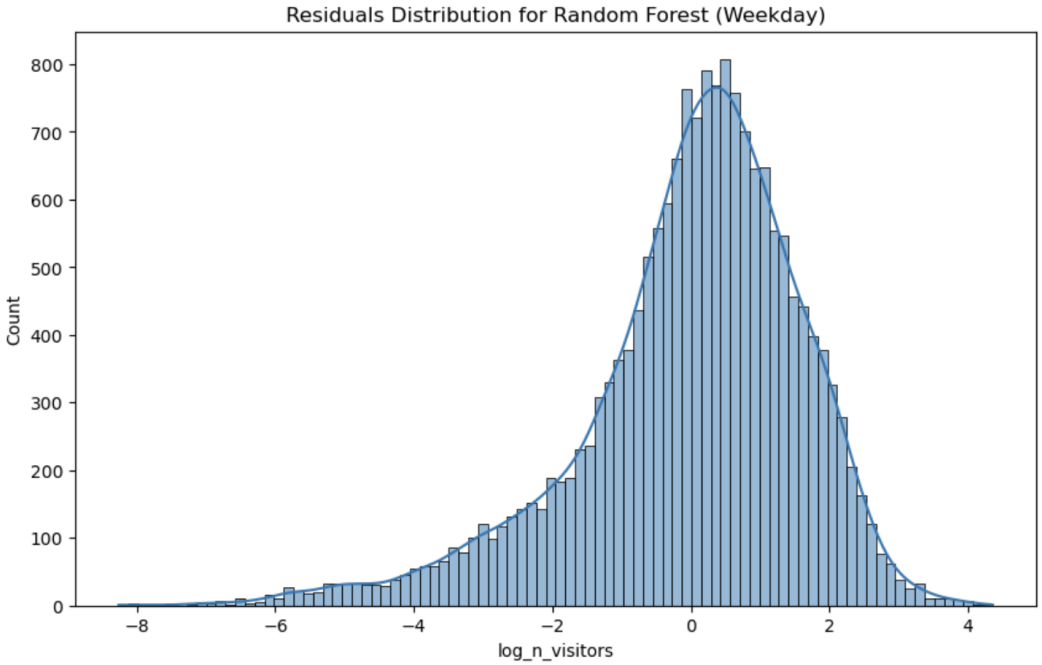


Figure B.22: Residual Distribution - Weekday Random Forest

Weekend Model Diagnostics

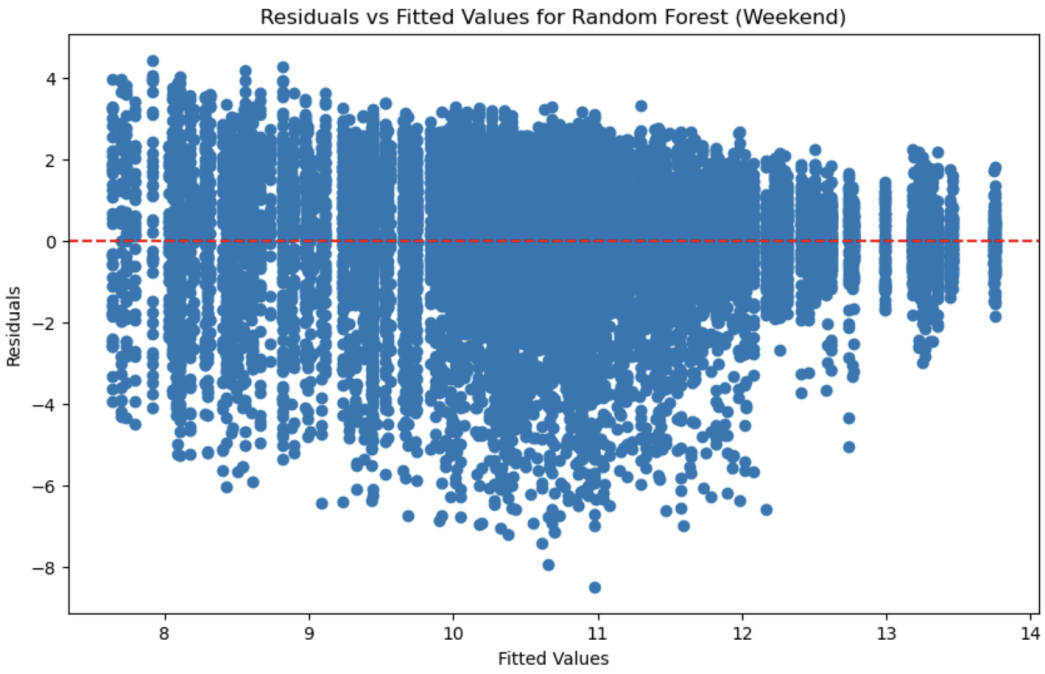


Figure B.23: Residuals vs. Fitted Values - Weekend Random Forest

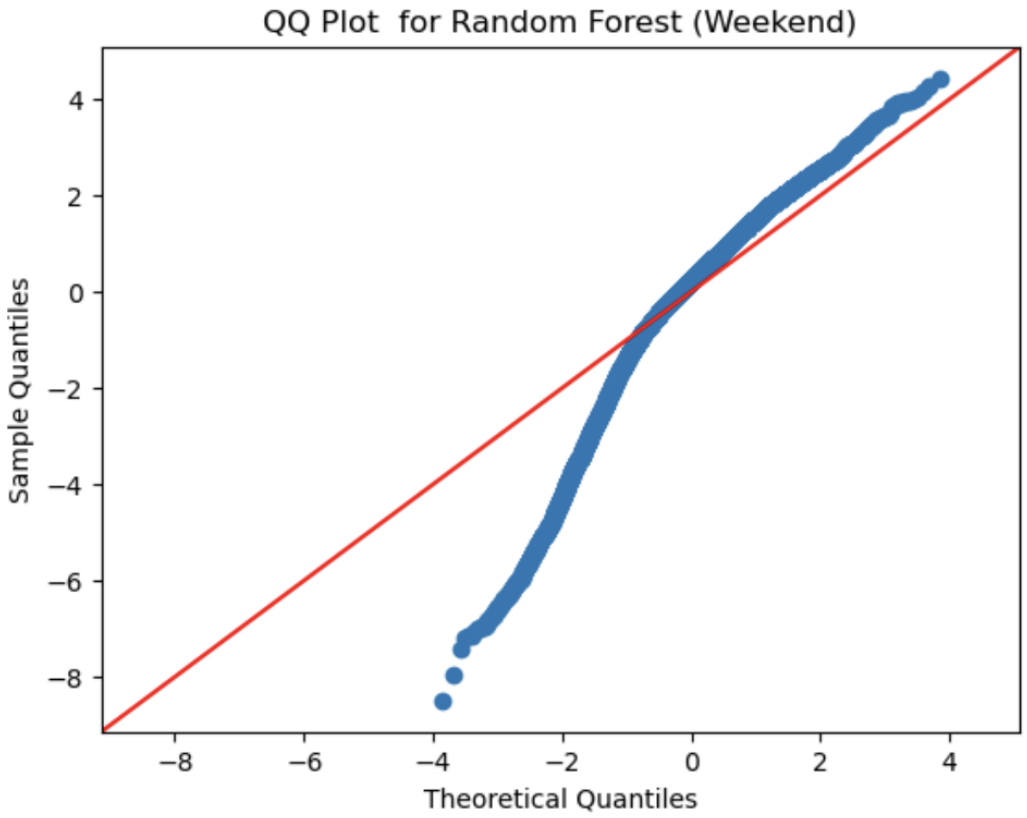


Figure B.24: Q-Q Plot - Weekend Random Forest

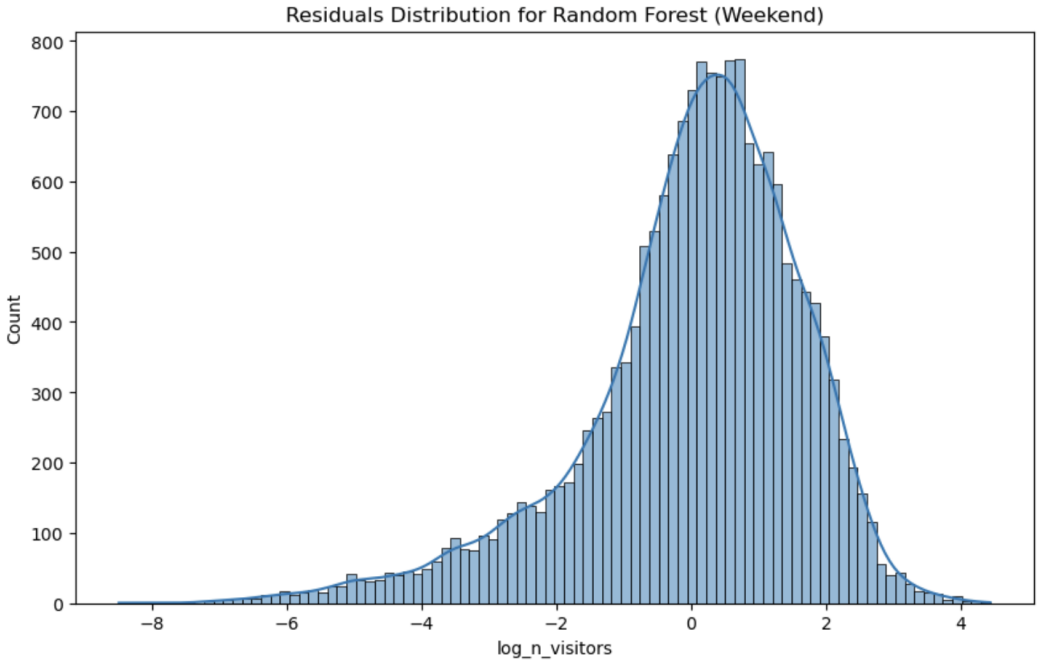


Figure B.25: Residual Distribution - Weekend Random Forest

B.1.3 Gradient Boosting Regression Models

Gradient Boosting Regression: Weekday Model

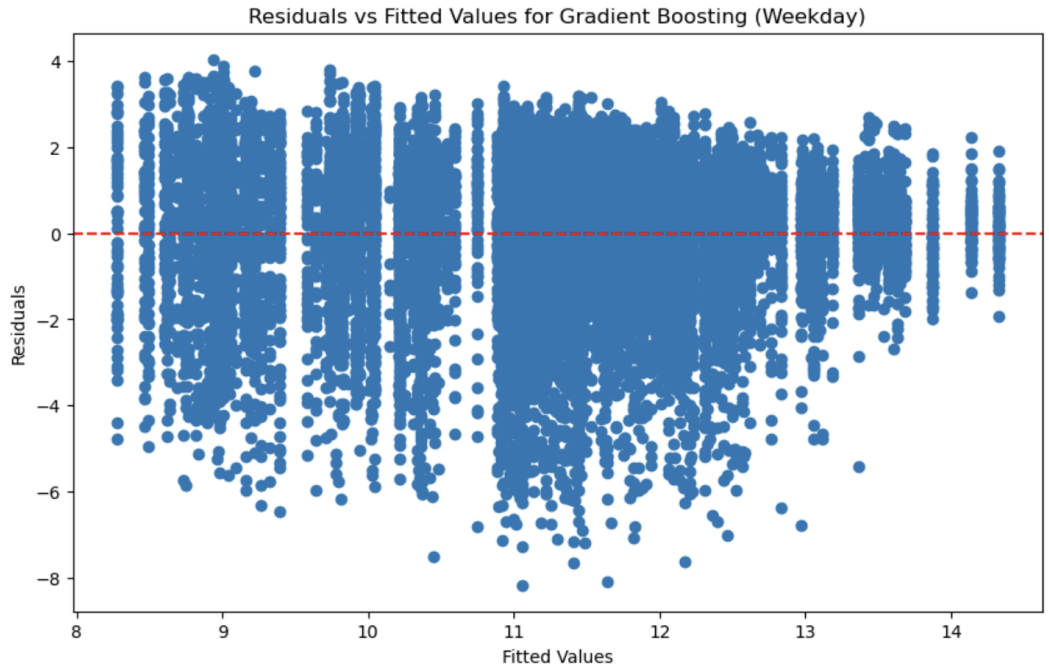


Figure B.26: Residuals vs Fitted Values for Gradient Boosting (Weekday)

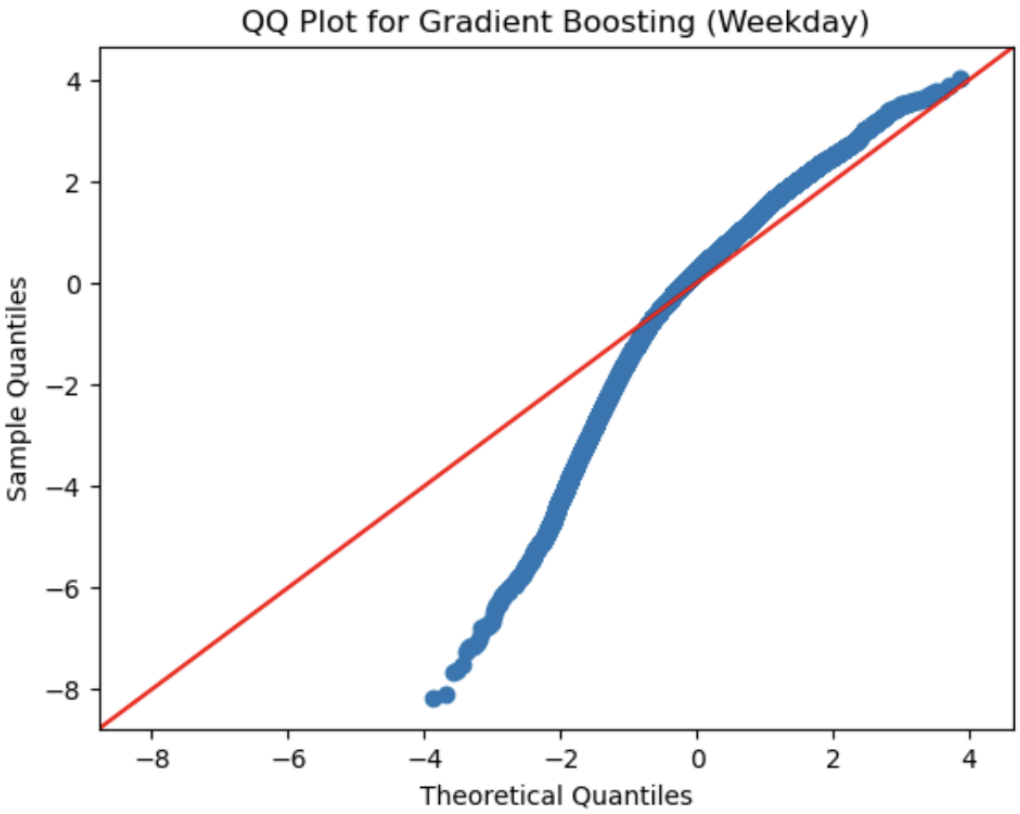


Figure B.27: Q-Q Plot for Gradient Boosting (Weekday)

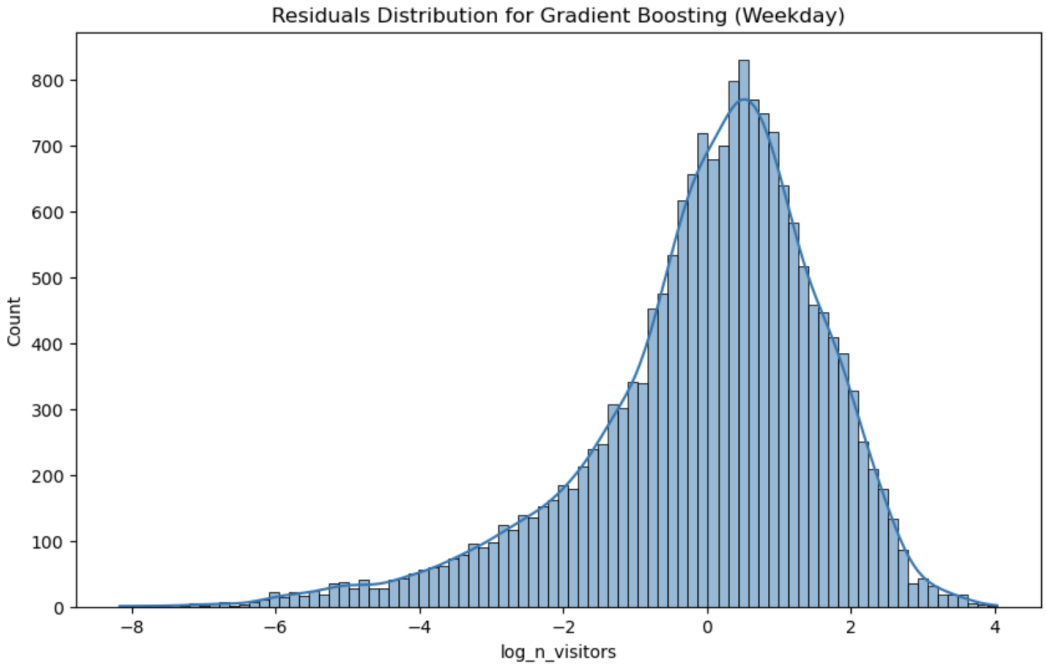


Figure B.28: Residuals Distribution for Gradient Boosting (Weekday)

Gradient Boosting Regression: Weekend Model

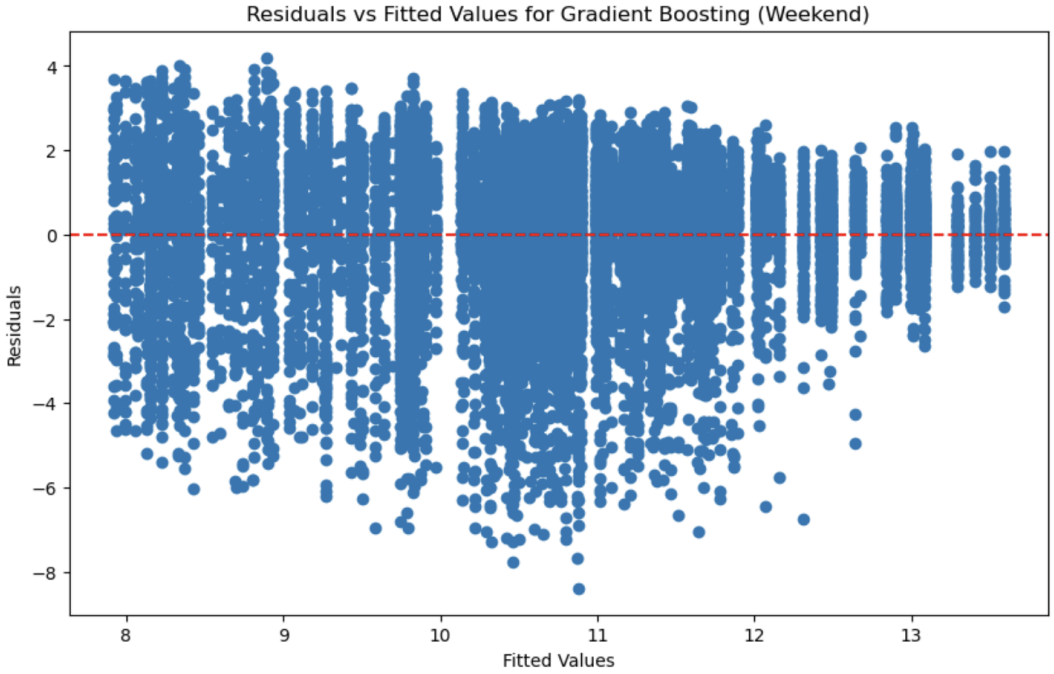


Figure B.29: Residuals vs Fitted Values for Gradient Boosting (Weekend)

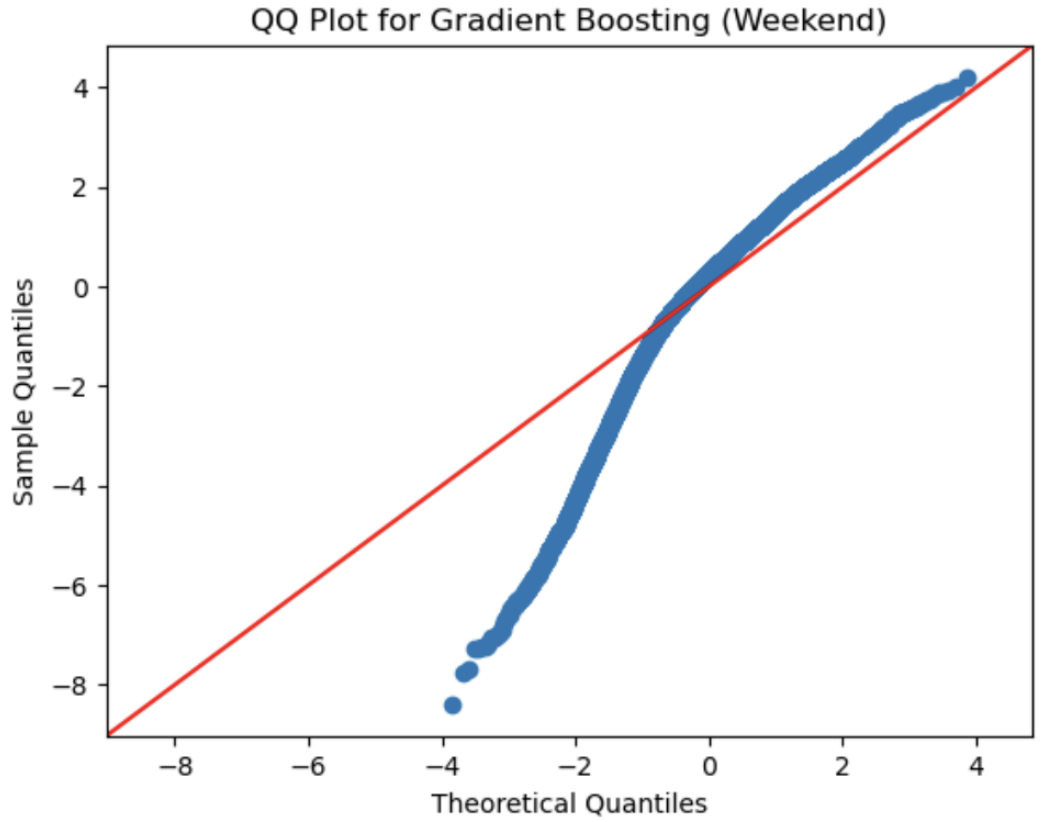


Figure B.30: Q-Q Plot for Gradient Boosting (Weekend)

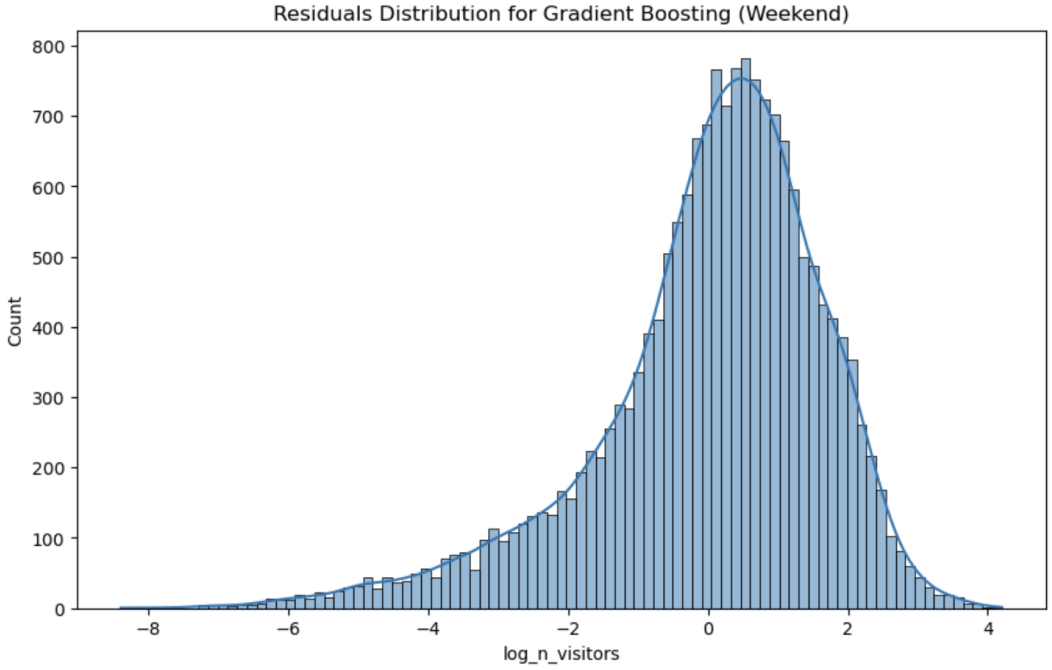


Figure B.31: Residuals Distribution for Gradient Boosting (Weekend)

Gradient Boosting Regression Metrics

Metric	Weekday Model	Weekend Model
R-squared	0.3682	0.3660
Mean Absolute Error (MAE)	1.2259	1.2352
Mean Squared Error (MSE)	2.6443	2.6817
Root Mean Squared Error (RMSE)	1.6261	1.6376

Table B.2: Gradient Boosting Regression Metrics for Weekday and Weekend Models

Appendix C

Data Visualization - Comparative maps

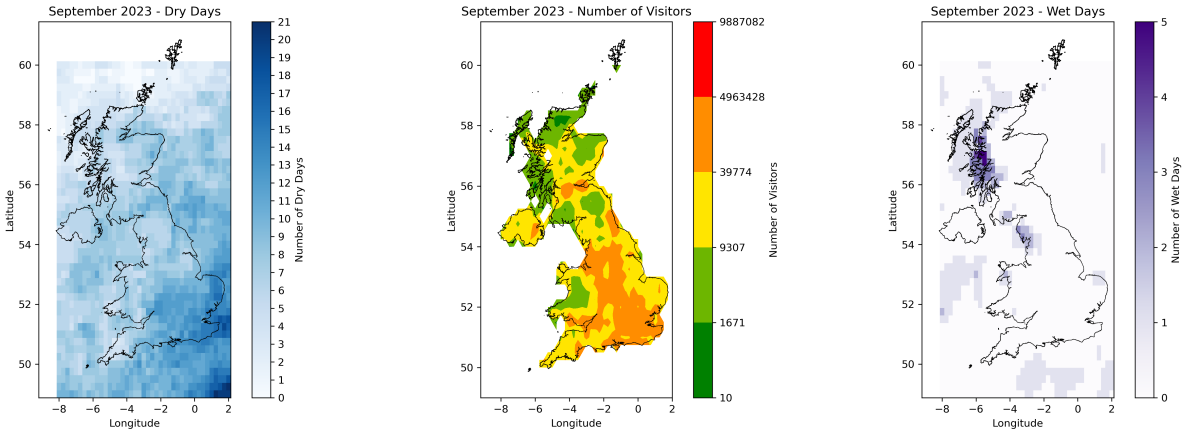


Figure C.1: September comparative plot

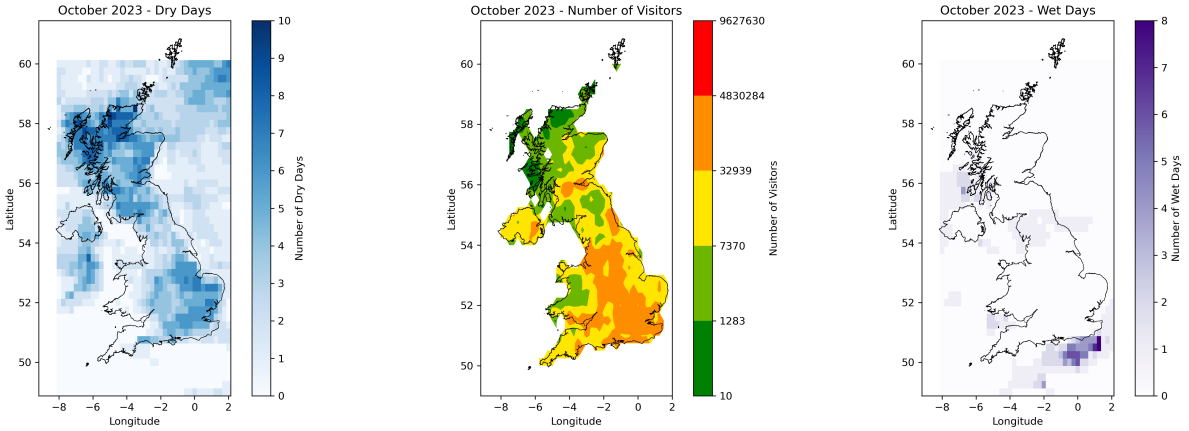


Figure C.2: October comprative plot

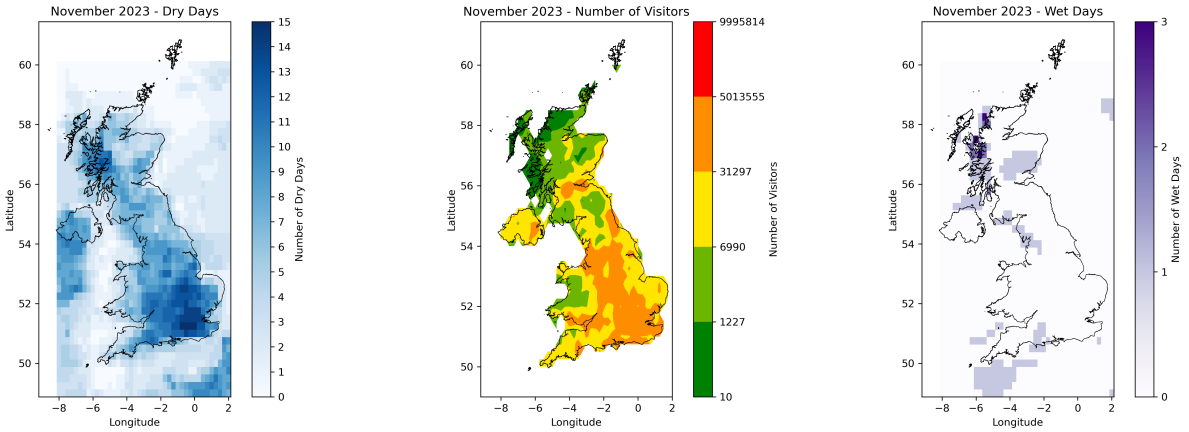


Figure C.3: November comprative plot

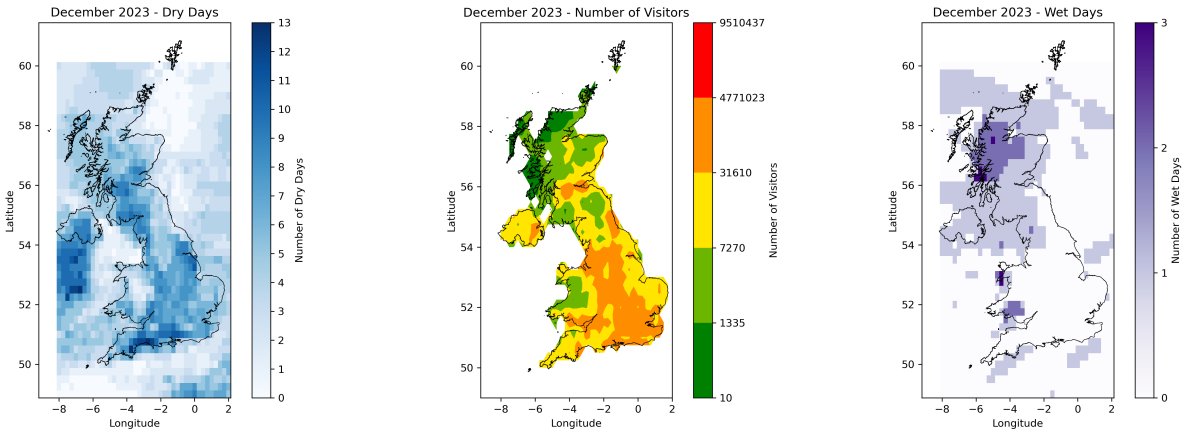


Figure C.4: December comprative plot

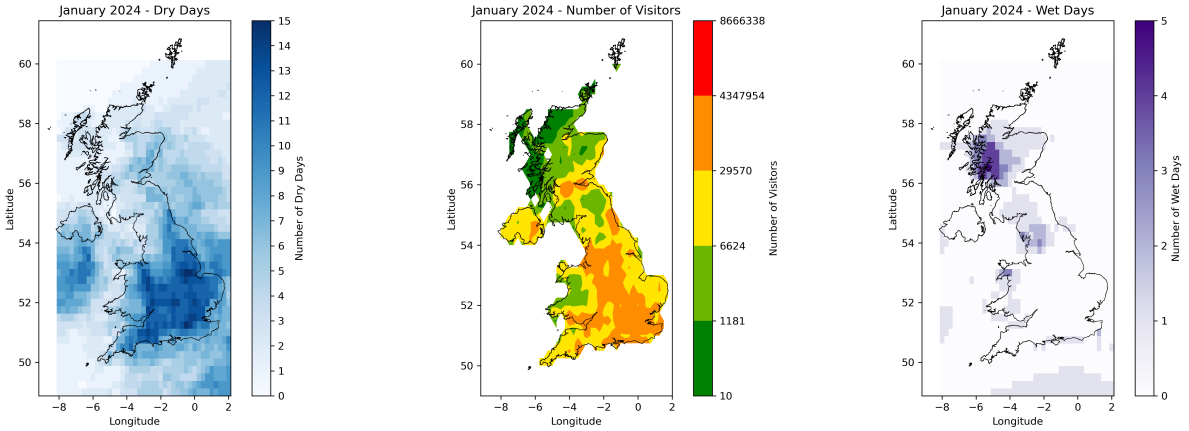


Figure C.5: January comprative plot

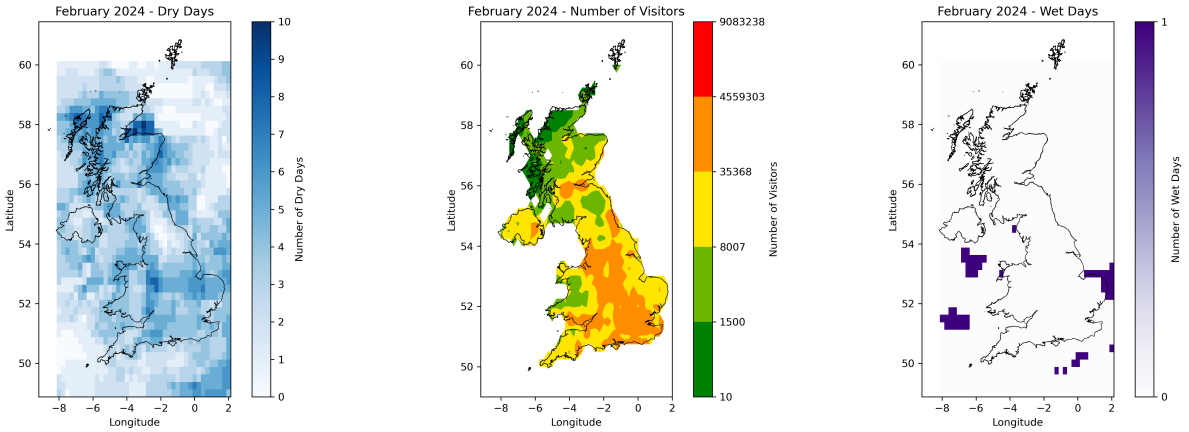


Figure C.6: February comprative plot

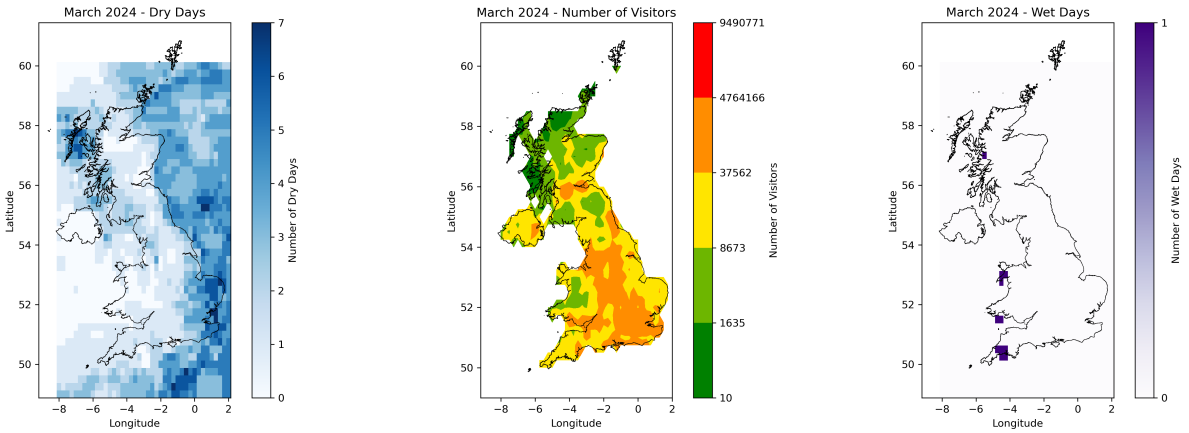


Figure C.7: March comprative plot