INVESTIGATING THE ROLE OF POINTS OF INTEREST IN ESTIMATING MOBILITY PATTERNS IN CITIES

An extended Gravity model - London Rail

KARAN PAPPALA

TU Delft

Investigating the Role of Points of Interest in Estimating Mobility Patterns in Cities

An extended Gravity model - London Rail

by

Karan Pappala

to obtain the degree of Master of Science in Engineering and Policy Analysis,
at the Delft University of Technology,
to be defended publicly on Friday, October 30, 2020 at 15:00.

Student number:	4931440	
Thesis committee:	Prof. dr. ir. A. Verbraeck,	TU Delft, chair
	Dr. ir. T. Verma,	TU Delft, first supervisor
	Dr. Y. Huang,	TU Delft, second supervisor

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Associated code and models are available at https://github.com/karanizer/Estimating-Mobility-POI.

TU Delft

Executive Summary

Urban cities are growing every day due to rising population, vehicular traffic, immigrants looking for opportunities and to accommodate this, cities are expanding and reshaping at an immense rate. People travel within a city for various purposes but the destination locations are the main reason for the movement. It is proved that locations influence a person's travel behavior. This research aims to study the influence of points of interest on human mobility available within a city and estimation of the travel flow between specified origin and destination locations. The city of London is chosen as the case study and the travel network is the London Rail. The research question this study aims to answer is "How to estimate human mobility by using points of interest?".

There are sub-questions developed to guide the process of this study and they are used effectively throughout. In literature, there is not much research regarding intracity mobility and not much regarding estimating mobility with amenities/ points of interest (POI). There is research regarding mobility estimations at scales larger than cities and usually, it is travel flows from multiple origins to a singular destination. This research includes travel flows between both origin and destination locations. There are however implications in research that POI data can be used to quantify the influence of attractions on mobility and might prove to be better than the traditional gravity models.

This research can help estimate the travel flows and aims to quantify the influence of points of interest in London. A new data preparation framework is designed from various data sources that lay the foundation for the data availability problem. The final dataset obtained contains the following variables: origin population, destination population, origin-destination distance, POI categories such as Commercial, Community, Educational, Entertainment, Financial, Government, Healthcare, Sustenance, and Transportation. For the model techniques, Negative Binomial and Poisson regression multivariate models are chosen. A series of experiments were conducted and the final experiments and the models within them were determined with the help of methods such as Akaike information criterion (AIC) and p-value. The model selection process was executed through a series of experiments to determine the most effective models. The model selection methods include the Akaike information criterion (AIC), p-value, Root-mean-square error (RMSE), and Coefficient of determination (popularly known as R-squared or R2). The RMSE and R2 methods were used to determine the best models among the sets of experiments created. Finally, model validation was conducted using the 'Sorensex Similarity Index' (SSI) method to quantify the similarity between the estimated and empirical data.

The models were built separately for each of the days (MTT, FRI, SAT, and SUN) as time could not be included as a variable to the model. First, the process was executed for the day MTT and then it was replicated for the rest of the days. There were not many differences observed between the days as the POIs selected did not contain any difference between the days. If the timings of the POIs were available, then there could be clear differences observed between the days. Limitations and assumptions such as these were elaborated discussed throughout the thesis. The selected models are the first models for both Poisson and Negative binomial regression generalized linear models (GLM) and they perform better than the traditional gravity models. Though the differences observed are minimal, it is important to note that the Negative binomial models performed better on SSI while the Poisson models performed with the model selection methods such as RMSE and R2. The first models performed better using both the model selection and model validation methods. All the first models contain all the variables i.e. all the POI categories including the origin and destination populations along with the distance between them and the Differentiator variable. However, the SSI values are around 0.46 leaving a difference of 0.54 to reach the optimal estimations. This issue is discussed in detail and multiple methodologies including adding additional categorical variables such as economic/demographic indicators are suggested to overcome it.

This research achieves the objectives presented in the paper and answers the main research question. The idea that POIs can be used to estimate mobility and they can be quantified within a city is novel and developing a new model satisfies and extends the research for both academic progression and policy implications. Regarding future work, there are limitless possibilities. The research can be extended for other cities, transport modes and also can be extended to other mobility models. The data availability

problem is evident in the research and if that issue is not prevalent then the study could have been seamless. The POI categories are created using a subjective ideology and that can be improved using modern classification techniques. For policymakers and urban planners, this can help plan the city effectively and efficiently towards a sustainable future solving many problems such as ineffective traffic models, last-mile hike problem, congestion reduction, inequality minimization, irregular spatial designs, livability index, and much more. This study provides the foundation for a new thought process in Urban Science and it also discusses the possibilities this research could provide if extended further. Urban planning policymakers can understand the importance of amenities on human movement through this research and hopefully, allow them to make new creative decisions regarding urban planning.

Keywords: human mobility, urban planning, intracity, amenities/points of interest (POI), gravity model

Acknowledgements

I would like to thank first my committee: Prof.dr.ir. Alexander Verbraeck for guiding me with great advise and bringing clarity to my research, Dr.ir. Trivik Verma for constantly guiding me and supporting me throughout the thesis and introducing me to the world of urban science, Dr. Yilin Huang for patience and highly valuable feedback in terms of research and reporting. I would like to thank Mikhail Sirenko and Yap Jin Rui for their work that enhanced and developed this thesis. I would also like to thank the Computational Urban Science Policy (CUSP) lab for providing me valuable feedback and support, especially during this pandemic. Finally, I would like to thank my family and friends who have always supported and stood by me.

Karan Pappala
The Hague, October 2020

Contents

1 Introduction

This chapter discusses some of the major problems of modern urban sciences such as urbanization, urban planning, and travel demand management. The complexity of the problem requires analyzing the influence of amenities on human movement. However, it is unclear how the concept translates into a quantitative assessment. The proposed method is a combination of open data and machine learning algorithms. This discussion is summarized in the research objectives and research questions sections.

1.1. Context

Urban cities are growing rapidly across the world. It is estimated that 4.1 billion people are currently living in urban areas and projected that almost 7 billion people will live in urban areas by 2050 [100]. That is more than half of the population of the world in urban regions. With high population growth in urban areas, there are numerous challenges up ahead for city planners and policymakers. For urban cities, the major challenges are congestion levels due to increasing traffic, toxic air levels, and integration of sustainable transport, developing towards the future integrating with modern technologies.

Transportation Demand Management or Travel Demand Management (both TDM) is the application of strategies and policies to reduce automobile travel demand or to redistribute this demand in space or in time [110]. In transport as in any network, managing demand can be a cost-effective alternative to increasing capacity. A demand management approach to transport also has the potential to deliver better environmental outcomes, improved public health and stronger communities, and more prosperous and livable cities. The techniques of TDM, applied by government transport agencies, link with and supports community movements for sustainable transport [26]. With the increase in toxic air, congestion levels, and population leading towards sustainable futures, there is great pressure on the Transportation industry for change towards the future. It became self-evident that alternatives to single-occupancy commuter travel needed to be provided to save energy, improve air quality, and reduce peak period congestion [11]. Various solutions are coming up across the world to deal with these issues. Some of the implementations are bike hubs in Netherlands [45], congestion pricing in US [5], and pod taxis in Sweden [3].

To understand the transportation demand, we must understand how people move in the first place. According to the iterative model of trip assignment [89], choice of destination is the priority of an individual, leading to the choice of transport mode. So, for certain, there is a strong relationship between the destination and individuals' movement. It is depicted in 1.1 below. This implies that people travel to or try to travel to places for the destination as the priority. If they are not able to travel to their choice of destination, they might choose an alternative destination with similar services. This might also depend on various other factors but the services that people can experience are the reason for getting out and traveling. When people travel to a location, there might be multiple purposes involved. A person can travel to go watch a movie with his/her family, eat at a restaurant, and finish the day off with some ice cream at the beach. In this process, there are four different activities - the movie, a restaurant, ice cream, and the beach. They can travel to the beach and stumble upon ice cream or travel for the ice cream and stay for the beach. In any case, they were attracted to the services offered by the location and utilized it. People travel towards a location providing services and this is observed

everywhere; Usually, every city has a downtown/city center which is the most attractive and popular part of the city. The curiosity to quantify the attraction of people towards key points of interest is the motivation behind this study.

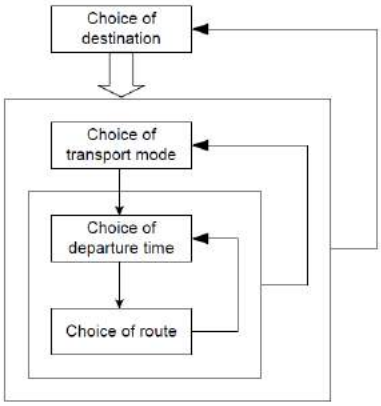

Figure 1.1: Iterative model of trip assignment

1.2. Relationship between Points of Interest and Mobility

In transportation, there has been a long desire to understand the motivation behind people's movements, i.e., what are the factors that drive people to move within a city in-between locations [107], e.g., work, shopping, studying, etc. Some researchers suggest that the motivation behind people's movements generally relates to certain purposes/activities that depend on locations [104]. In simpler terms, people's movements are highly related to the distribution of activities (e.g., school, office, food stall, etc.) within a city.

Understanding the relationship between activities and human mobility is a fundamental research problem in transportation, which is highly useful for estimating trip distribution and travel routes[107]. The research done by [86] proves that human mobility is highly related to POIs and this information could be very useful for urban planning and traffic management. They state that multiple types of research have been carried out for this purpose [112] [76]. Transport policymakers have problems with the demand for transport and keeping up with sustainable solutions.

Public transport such as buses and metros helps people travel large distances within the city for low costs and bring them closer to their destination. This helps people in reducing the cost of their vehicle usage, contributing to a sustainable environment, and avoiding waiting time in the heavy traffic that comes along in the intracity travel. Most people when they exit their travel stop, have their destination set in close proximity. This phenomenon is known as last-mile travel.

Transit networks and the built environments that support and surround them can both facilitate and hinder more sustainable travel behavior, particularly when considering the linkage between main travel modes and their first/last leg connecting journeys [108]. The key concern of last-mile problems is the facilities linking the main mode to the home, workplace, or wider destination, which are often poor and thus discouraging sustainable behavior [73]. The last mile problem, at its core, is quite a simple one public transport does not take us exactly where we need to go, parking is not always available everywhere we go, owning a car or any kind of vehicle is not always possible or even reasonable. And walking is not always the quickest or the most convenient way to move around the city [42]. Some last mile solutions implemented are E-bikes [8], bicycle pods [45], electric vehicles [2] and they are also being used for the sustainable delivery systems [18] [111] [39].

In various parts of the world, the last-mile bike is being implemented that helps people get out of their penultimate travel point and use bikes to get to their respective destinations. It is prominent in Europe and countries including Brazil, Chile, China, New Zealand, South Korean, Taiwan, and the USA began to introduce the bicycle-sharing program [92]. By the end of 2017, more than 23 million shared bicycles were available around the globe and 304 cities in more than 20 countries had implemented bicycle-sharing systems [59].The demand for bike travel is increasing and it can develop sustainable

behavior for cities while solving the last mile problem. The last mile travel demand can be measured by analyzing and predicting mobility related to the facilities from which the last-mile journey begins.

With advancements in technology, data such as POI data, landuse data, and other data related to a city are being captured and used to improve the city in various areas such as traffic models, livability improvement, etc. The POI data captured is limited to the big cities in the Western Hemisphere and few cities in the Eastern Hemisphere. The POI data available for mid-range or lower-tier cities is not adequate at all. For the big cities also, the data captured is not complete as cities contain various systems intertwined in complex networks. With time, more data will be captured and real-time data might be available soon in the future. That is a gold mine for a researcher in urban science. The POI data chosen must be efficient, adequate, and should integrate seamlessly into the model. The type of model will be determined after the POI data chosen is explored. The POI data is the basis for this study and using it to estimate mobility within a city has not been researched, though suggested.

The locations are the reason why people go and fulfill their activities, and transport is a means to get there. The locations are the points of interest (POI). By understanding POI and the attraction they have on peoples' movement, we can understand the weights they have on mobility. They might be influenced by multiple amenities such as a restaurant, a library, and a park. Different amenities attract different people in multiple ways. And people do different activities at different times of the day. On weekdays, the majority of the people usually have a similar routine and go to similar destinations. On weekends, they perform other activities and the travel pattern differs. This is important for the study and time can tell us the necessary differentiation between the days of the week and travel patterns towards the amenities.

1.3. Knowledge Gap

It is understood that there is a relationship between POIs and mobility. There are not many studies regarding quantifying the influence of POIs on mobility. There are research studies that indicate that POI as data points for estimating mobility would yield better results. Using POI to estimate mobility is a new approach and has not been studied in this sense before. It has been suggested by a couple of researchers for estimating mobility but has not been extensively integrated into a model. Due to advancements in technology, data is being captured effectively and can be used efficiently. The POI data available in the world is limited to big cities and even then, the data captured is not complete nor adequate. The city chosen needs to have high standard data quality for the research to be effective and seamless.

By understanding what attracts people at different times and how this affects the transport department to align with people's demands in terms of sufficient/excess transit lines and appropriate last mile bikes will help the city develop a sustainable lifestyle, decrease the congestion levels, and toxic air in the city. By integrating mobility models with POI data, we can assess the effect of POI data on human mobility and be able to estimate the flow between the designated locations. Though a person might travel to a specific destination for their purpose, the surrounding amenities might influence behavior on the mobility of the individual. POI data determines the activities undertaken by people and the activities can determine the mobility flow. The POI data contains various amenities that need to be categorized first.

The previous models though might have the theoretical capability, do not have the data available to perform the empirical experiments. POI data is being captured daily, and maybe one day the data might be available in real-time. Then accurate models can be developed and cities can truly flourish with the means of urban science. The concept of using POI data to estimate mobility is unique and is the core of this study. POI data is fairly new and trying to quantify it while trying to estimate mobility is a novel idea.

1.4. Research Objectives

From the above knowledge gap, it is clear that there is a clear influence of POI on mobility. This research aims to find if mobility can be estimated by using POI as key metrics and how much attraction do they possess on the movement of people within a city. First, the research objective of this thesis is

"1. Quantify the influence of POI on mobility"

By quantifying the weight associated with POIs infunce on mobility, the same can be used to estimate the mobility patterns as well as understand the spatial distribution. Thus, the next research objective this thesis aims to study is

"2. To estimate mobility with the aid of POI"

1.5. Thesis Outline

The thesis is divided into four parts: Introduction to the problem, Case Study, Model development, and Synthesis. The below Table 1.1 outlines the structure of the thesis.

Table 1.1: Thesis outline

Part		Chapter
I	Introduction to the problem	1.Introduction 2.Literature review
II	Case Study	3.Exploratory Data Analysis 4.Methodology
III	Model development	5.Model results
IV	Synthesis	6.Discussion 7.Conclusion

Part I introduces the research problem. Chapter 1 Introduction 1 provides the context for the problem. Chapter 2 2 provides the Literature review necessary to understand and formulate the foundation for the problem. At the end of this chapter, the research questions are formulated.

Part II explores the Methodology section. In the Exploratory Data Analysis chapter 3, the case study is selected and the framework for the preparation of data is presented along with exploratory analysis. The data description and sources are also included. Next, the Methodology chapter 4 ends with the methods used in this research.

Part III showcases model development. The chapter Model results 5 includes Model validation as a part of it. The experiments are discussed and the final models are selected after iterative and rigorous validation. Exploratory analysis presents the analysis and compilation of the data before model development and the critical assumptions undertaken. Model results are the results obtained to achieve the research objectives. It ends with Model validation where the model performance is evaluated.

Part IV discusses the model results, limitations, academic progression, and policy implications along with future work. It ends with revisiting the research question and concludes the research of this thesis with reflection.

1.6. Research Questions

From the above discussion, a few research questions are formulated which will guide the study. This leads to the main research question,

"How to estimate human mobility by using points of interest?"

Following the main question, the following sub-questions are formulated. The sub-question determines the flow of the research and help in guiding the process. The formulation for the sub-questions is explained in detail in the next chapter Literature Review 2.

- Do modern techniques and models exist for estimating mobility using points of interest?
- What are the existing models to estimate mobility?
- What role do points of interest play in assessing the Origin Destination travel flows?
- How does the new model compare and contrast against other existing models?

2

Literature Review

This chapter aims to provide the background and foundations for this study. First, The formulation of the subquestions is discussed in detail. These questions guide the process for this research. Next, mobility models with Points of interest (POI) are explored, followed by the Mobility models section. From this, the Gravity model is selected and discussed. It delves into the Gravity model and how POI data can be used as inputs to the model. Finally, POI sources are discussed and how OpenStreetMap (OSM) is a reliable source to obtain POIs. Thus, this chapter outlines the POI data accessibility and reliability necessary for the gravity model selected and guides the research process.

2.1. Subquestions Formulation

The subquestions were presented in the Introduction chapter 1. The formulation is discussed in detail in this section. First, we know that there is a relationship between POI and mobility; and, as we are trying to estimate mobility using POI, it is natural to find out existing studies regarding estimating/predicting mobility using POI. This leads to the question, "Do modern techniques and models exist for estimating mobility using points of interest?"

After finding out existing studies, it is important to know the different methods in which mobility can be estimated and the research on it. This would determine new methods to incorporate the POI data points and the right model for the estimation. This leads to the next subquestion, "What are the existing models to estimate mobility?"

Once the methods are decided and the results are obtained, it important to understand the importance of the role of point of interest (POI) in estimating mobility and to what extent it is effective. This leads to the question to be answered, "What role do points of interest play in assessing the Origin Destination travel flows?".

From a researcher's perspective, there should be new information from this study. A new model is being built to estimate mobility within a city, along with contemporary concepts such as POI as data inputs and the model should be evaluated to understand it's performance when compared against existing traditional methods. The subquestion for this is developed as "How does the new model compare and contrast against other existing models?"

The purpose of the subquestions is discussed and these questions will guide the study process.

2.2. Mobility models for POI

In the research conducted by [48], they explored the potential of POIs by replacing the population variable with POIs and check-in (population proxy) data. It was deduced that there was an improvement in the performance of all the different models and an increment of 20% of the Sorensex Similarity Index (SSI). This method is discussed and used later in the research. As mentioned above in 2.5, the OSM data has been used for multiple research purposes and can be used for obtaining the POI data.

The study conducted by Camargo, Bright, and Hale [58] proposed the idea that OSM points of interest features could be utilized in the gravity model; "replacing the idea of people gravitating towards other people with the idea of them gravitating towards features which might be of interest to them."

This idea sparked me to study this problem and is the origin behind this research. They also mentioned, "we would like to explore the origins and destinations of movements related to an area of interest."

A study was conducted in China visualizing the relationship between human mobility and points of interest [119]. They designed Singapore massive public transportation data and POIs retrieved from Foursquare. The studies demonstrated that people's movements are highly related to POIs, and some other interesting findings have also been observed. Additionally, they explored the relationship on POIs for movements by different groups of people to figure out what POIs are more attractive to the respective groups. So, there is an attraction of POIs that can be quantified and further, increases the purpose of the study.

There is not much research regarding mobility models with POI as inputs, mobility, and the corresponding estimation models have to be explored. This is discussed in the upcoming section.

2.3. Mobility

Individual human mobility is the study that describes how individual humans move within a network or system [77]. The concept has been researched by several interdisciplinary fields. Understanding human mobility has many applications in diverse areas, including spread of epidemics [117] [56], mobile viruses [114], city planning [57] [102] , traffic engineering [79] [115] and financial market forecasting [65], to name a few.

Regarding predicting mobility, abundant research has been conducted and multiple models have been developed. A taxonomy of data-driven human mobility models was curated [113] and it includes a section 'Population Mobility model' includes gravity model [121], radiation model [103] and intervening opportunities (IO) [106] models. These models focus on the movement of the collective population between two locations - origin and destination. They are used to estimate the migratory flow between the regions. As thef name suggests, these models traditionally use demographic and geographical data, being the population of the regions and the distance between them. Over time, the gravity model, developed by Zipf gained popularity and was used in many studies and applications.

There have been numerous studies comparing these classes of models diagnosing the model performances and other metrics. A study using census commuting data from France, Italy, Mexico, Spain, USA, UK, and additionally, commuting within London and Paris has concluded that the gravity model performs better, by a minute margin, than other models at predicting traffic flows at these spatial scales [83]. According to a recent study by Camargo [58], the models are compared at small spatial scales to predict traffic flows between different wards in the county of Oxfordshire, UK. They concluded that the gravity model performed significantly better than radiation and IO models. All the traditional models use the population as the key metric and distance between the origin-destination locations. This has traditionally been the case for a long time. Though that being the case, the accuracy of prediction is quite low with all models performing poorly at meso scales. They suggest that population as a data metric is not sufficient and granular data such as OSM data providing points of interest could be incorporated into mobility models and this may provide better predictions.

Recent studies have used social media data such as Foursquare as a proxy to population [48]. Foursquare allows its users to check-in to locations as a feature. The check-in data generates Location-based social network (LBSN) data where people publish their location when they make a post on the application. From the LBSN data, they identified the check-in points of interest (POI) categories relating to the location. Though the sample may be small, the data is compared against real data such as taxi data to estimate the number of trips which makes it reliable. Different models such as gravity, radiation, IO, and PWO models are compared across two key factors - population and LBSN, in the city of New York. From the research, it was concluded that the models with LBSN have performed better than the models with the population as the key factor. Also, it can be inferred that gravity and radiation models performed well among the models.

At the city level, the research available is comparatively low to the intercity movement. Most of the models are predicting large scale commuting flows efficiently but perform poorly when it comes to small spatial scales [58]. There is not enough research regarding small spatial scales and the predictions are far from accurate. This problem of poor mobility predictions at small spatial scales has also been addressed by considering the variation in the accessibility of different sites [97] and in the topology of urban spaces which include slums [55]. There is evidence to support the relationship between land use and human mobility in a city but the spatial data is aggregated at a broad level, not providing intricate

details about a city, and with emerging technologies, data is available at a granular level providing accurate details [80]. From this, we can deduce that there is not much research when it comes to the intracity level.

From the above diagnosis, we can confer that the gravity model performs well among the models when it comes to estimating mobility. For this research, the model to be developed is inspired by the gravity model and it is discussed in further sections. But first, we need to understand the data availability of POI.

2.4. Gravity Model

The gravity model is inspired by Newton's law of gravitation. The gravity model assumes that the flow between an origin and a destination is proportional to their attractions (population) and inversely proportional towards distance between the locations [121]. It is a parameterized model and the model is calibrated against empirical travel data to obtain parameter values. It can be used to understand the pull effect the masses have on the movement of entities. As we are looking for attraction towards amenities, this model is a perfect fit to obtain the parametric values associated with POIs.

As mentioned above, the gravity model has been used in various disciplines. Given the nature of the problem containing POI data in the model meaning multiple variables, the gravity model can be designed to have multiple variables. This can apply to the case here. The gravity model is most commonly used by international and regional economists to study trade [70]. This has been used extensively for research over decades and developed slowly [99] [69].

In the simplest form, the model is represented as

$$T_{ij} = G\left(\frac{M_i^x * M_j^y}{D_{ij}^\alpha}\right) \tag{2.1}$$

where T_{ij} represents flow from origin i to destination j, M_i and M_j typically represent the populations for locations i and j, D_{ij}^α denotes the distance between the two locations, and G represents residual variable.

The model has been modified over time by researches, it is estimated to be as following

$$T_{ij} = exp(\beta_0 + \beta_1 M_i + \beta_2 M_j + \beta_3 D_{ij} + \beta_4 x_1 + \beta_5 x_2 + ...) \tag{2.2}$$

In this case, all β are the weights associated with corresponding variables and the value signifies the relationship with the flow variable T_{ij}. The x_1 and x_2 and other additional variables represent the extra factors which influence the mobility.

From a mathematical perspective, the gravity model is a type of regression analysis, a means of comparing sets of variables in search of relationships between them. Most models in this scenario follow multi-variate regression analysis [63]. This is explored further in the Methodology section.

2.5. Points of Interest (POI)

The POI data is the points of interest (entities) available in a landscape. It encompasses all the amenities available such as schools, hospitals, offices, parks, banks, etc. Compared to conventional land use data, the POIs have finer spatial and granular detailed properties. Existing studies have used POIs from check-in data or social media tagged data [48] [84] [91] [90] [98]. The problem with this approach is the incompleteness or bias towards the selected data. People usually check in with social media when they are visiting tourist attractions or restaurants and cinemas rather than hospitals, schools, and metro stations. To overcome this issue, there are other sources to obtain a holistic system of POIs. There are APIs from various companies pertaining to selected regions and they have been used for various research purposes [85] [58] [87] [118].

For this research, OpenStreetMap (OSM) is going to be used. Unlike other data sources that require monetary inputs, OSM is available freely worldwide and does not pertain to certain regions. There are concerns regarding the quality of OSM data as it is compiled by volunteers and does not follow standard industry procedures [68]. A lot of quality research has been produced using OSM data [58] [51] [72] as well as performing quality assessments for credibility [75] [120] [71] [67] [93]. For this research, it can be said that OSM data though not entirely comprehensive, can be used for research purposes and provide useful insights to the study of mobility in urban landscapes.

Given that the POI data is available abundant in big cities, a city in the Western Hemisphere will be chosen for this study. The points of interest in a city will be enormous and will have to be categorized accordingly. Depending on the data selected and coverage of the data, the techniques to categorize the POIs will be determined. Though the gravity model is selected as the model for estimation, if the POI values are not sufficient or are of high quality, then the research would not be effective. Thus, the POI data selected will have to be explored and adjusted accordingly to fit the model techniques.

The OSM POI data contains amenities which determine the corresponding activities occurring at the locations. This needs to be explored and amenities have to be selected depending on the use of activity and their corresponding category. From the above literature, it clear that there is a relationship between POIs and mobility. The foundations for the process are described and the methods used to formulate this process are described in the following section.

Thus, we have covered the research sub-questions and by answering them, we get closer to quantifying the POIs to estimate mobility. We have determined that OpenStreetMaps is a reliable source for the POI data and can be used to explore the relationship with mobility. Regarding mobility, the gravity model is selected and the structure of the model to be applied is determined in the upcoming chapter. The other models such as the Radiation model can be also be used as an alternative model. However, given the time limitations only the gravity model is considered for this research. The last two research questions "What role do points of interest play in assessing the Origin Destination travel flows?" and "How does the new model compare and contrast against other existing models?" are discussed partially in the above sections and will be elaborated in chapter 6 Discussion.

3

Exploratory Data Analysis

In this chapter, initially, the data preparation along with the methods used is discussed and a data preparation framework is developed. At the end of the data preparation framework, the final dataset required for the model is developed. Next, the model development section is discussed including the methods to determine the best models. Finally, the model validation section contains the necessary methods to determine the performance of the model.

3.1. Case Study - London Rail

For this study, the city of London is chosen due to data availability and for its location in Europe. London is the capital and largest city of England and the United Kingdom [10]. It is considered as one of the world's most important global cities and is well known for its historical significance, multicultural integrity, and technological advancements [23] [15]. It is estimated that the mid-2018 municipal population of Greater London was 8,908,081, the third-largest populated city in Europe after Moscow and Paris [9] [22].

According to the 'Travel in London' report generated by the Transport of London 2019 [109], around 26.9 million estimated daily average number of trips occurred in Greater London, 2018 for all the modes including rail, bus, car, cycle and walk. For the rail network, the estimated daily average number of trips is 5.8 million. The public transport share is 36% while the private transport share is 37&. The statistics also indicate that a sustainable lifestyle of travel is being promoted and utilized by the citizens compared to the 2000s. The report discusses in detail the different systems associated with travel in London. The three main themes followed for the future of transport of London and the development of the report are 1) Healthy Streets and healthy people, 2) A good public transport experience and 3) New homes and jobs. This affirms the vision of this research that the urban science is brought into play to improve the livable conditions for a city.

Boasting such rich history and significance in the world, London has been the epicenter of numerous research studies. The availability of data for this research is prominent for the city of London. Before going further, it is important to note that the following analysis is conducted in Python. With an increase in computational capacity and development of open-source programming languages such as Python and R, detailed analysis as well as Machine Learning (ML) techniques can be applied to achieve our objectives. The data sources are discussed below.

The shapefiles are the necessary files that provide the geographic boundaries of the concerned region. The shapefiles are captured in the form of geographic information system (GIS) files. A geographic information system (GIS) is a computer system for capturing, storing, checking, and displaying data related to positions on Earth's surface [88] [60]. The London GIS files can be accessed from the London Datastore and with this, we can outline the city of London [37].

To obtain the points of interest (POI) in London, there are various methods to obtain the same. From the above Literature Review, it is concluded that OpenStreetMaps (OSM) is selected for obtaining the POIs. Geofabrik is a free open data download server and contains data extracts from the OpenStreetMap project which are normally updated every day [14]. The data extract for Greater London is selected for the study. To extract and analyze the data, the Python package pyrosm is utilized. This package is

extremely helpful and processes the data rapidly comparative to the osmnx [54] package in Python.

The population is an important factor in the study. As we are performing analysis within a city, it is important to obtain the population estimates at the lowest level possible. In the case of London, the population estimates are available at a ward level and this is a sufficient meso scale for this study. The population estimates can be found from the Office of National Statistics website [4].

Regarding the travel data, we need the flow of people traveling from one location (origin) to another (destination) within the city using any means of transportation. For this research, the London railway network is considered. The London Underground is the oldest underground railway network in the world [40] and is apt for the research as it contains adequate information. The Origin-Destination (OD) travel data is provided by Transport for London (TfL) [43] and it constitutes of three modes: London Underground (LU), London Overground (LO), and Docklands Light Railway (DLR). There is TfL data but it is indicated that it is part of LU. The three modes are considered for this study. However, the station coordinates are not provided in the same section and are obtained via the right to information provided by the Greater Authority of London [12]. With the station coordinates, we can identify the station's location on the map and the surrounding amenities associated with it.

The table 3.1 below summarises the data sources discussed above.

Table 3.1: Data sources summarised

No.	Data extracted	Data source
1	Shape files	London Datastore [37]
2	Population estimates	Office of National Statistics [4]
3	Points of Interest (POI)	OpenStreetMaps [14]
4	Origin-Destination Travel flow	Transport for London [43]
5	Station Coordinates	Greater London Authority [12]

3.2. Data Description

The data sources are discussed above and from the same, the data is extracted and modified. This section provides an overlay of the information provided in the datasets obtained.

Shape files

The shapefile selected is the `OA_2011_London_gen_MHW.shp`. It was compiled in 2011 by the Greater London Authority for London and the shortest geography is Output Area (OA). The key ID of the dataset is OA11CD is Outer Area code which constitutes of Postcodes and wards and the necessary data for the study: geometry (in the shape of a polygon) is available. A geometric polygon in the shape of a polygon connected by straight lines on a spatial plane, in conceptual terms, it means a region on the map bounded together as a polygon. The dataset also constitutes ward and borough codes and names respectively along with various household variables. It contains 25053 Outer Area codes in total.

Population

The population file obtained is SAPE20DT10a-mid-2017-coa-unformatted-syoa-estimates-London.xlsx and the sheet used is 'Mid-2017 Persons'. It was compiled by the executive office of the UK Statistics Authority and they compile information about the UKs society and economy and provide the evidence-base for policy and decision-making, the allocation of resources, and public accountability. It contains OA11CD as the key ID and constitutes the population for all ages of the region. There are 25053 Outer Area codes in total and then we consider the total population for the respective Outer Area (OA) codes.

Data Coverage

In England and Wales 2011, Census Output Areas (OAs) are based on postcodes as at Census Day, wards (and parishes). The minimum OA size is 40 resident households and 100 resident persons but the recommended size was rather larger at 125 households. These size thresholds meant that unusually small wards and parishes were incorporated into larger OAs. In total there are 181,408 OAs in England (171,372) and Wales (10,036).

OSM POI data

The file `greater-london-latest.osm.pbf` is obtained from the Geofabrik data download server where Great Britain is selected under Europe. By using the library pyrosm containing function `get_pois()`, we can obtain all the POIs for the region. The data extracts are usually updated daily. Regarding the data description, there a dedicated OSM wiki website that explains the structure and information of the dataset [27]. The assumption considered here is that the dataset obtained acts as a proxy for the coverage of the city of London.

Travel data

The Origin-Destination (OD) data available is provided by Transport for London (Tfl) for project NUMBAT. There is a data description file `2018NUMBAT_Definitions.xlsx` which describes all the datasets in the folder. The dataset represents the travel demand on a typical autumn weekday, Saturday and Sunday at all stations and lines of the London Underground, London Overground, Docklands Light Railway, TfL Rail / Elizabeth Line, and London Trams. Data covers every 15-minute period throughout the traffic day and assumes a perfect train schedule being operated. The folder contains various London transport-related data such as OD data, Station Entry/Exit data, rail network information, and much more. The OD data selected is available for the different modes:- London Underground (LU), London Overground (LO), and Docklands Light Railway (DLR), and this is divided across different days of the week separately as MTT (Monday to Thursday), FRI (Friday), SAT (Saturday) and SUN (Sunday). The assumption considered here is the travel dataset is accurate and encompasses the rail network for the study.

Station Coordinates

The station coordinates are not available in the same folder but they are provided by Tfl (Transport for London). The file selected is `Stations_20180921_locations` , complied in 2018, and it contains the latitude and longitude for different stations across the modes - LU, LO, DLR, TfL Rail, and Tramlink. As the OD data is available for LU, LO, and DLR, these specific modes are only considered for the study.

3.3. Methodology Flow Process

The research is divided into 3 phases - Data Preparation, Model results, and Model validation. The flow process is illustrated in figure 3.2 below.

Table 3.2: Methodology Flow process

Phase	Topic
I	Data Preparation
II	Model results
III	Model validations

Data Preparation is the phase where the data is compiled from different sources, explored, and modified to obtain the final OD dataset. In Model results, the final OD dataset is fit to the model and the estimates are obtained along with the predictions of the flow. In Model validation, the results are analyzed and the model is verified against empirical data to obtain the efficiency of the model. The different phases along with the methods used are discussed in further sections.

3.4. Data Preparation Framework

Phase I 'Data Preparation Framework' is discussed in this section. There are 3 sections in this process: POI data, Travel data, and Final OD data. The POI data section is discussed first followed by the Travel data section and finally, the Final OD data section is discussed. The following research flow diagram outlines the different processes undertaken.

Before discussing the above sections, it is important to note that the analysis of this phase is conducted in Python. Python offers many packages to perform specific tasks depending on the nature

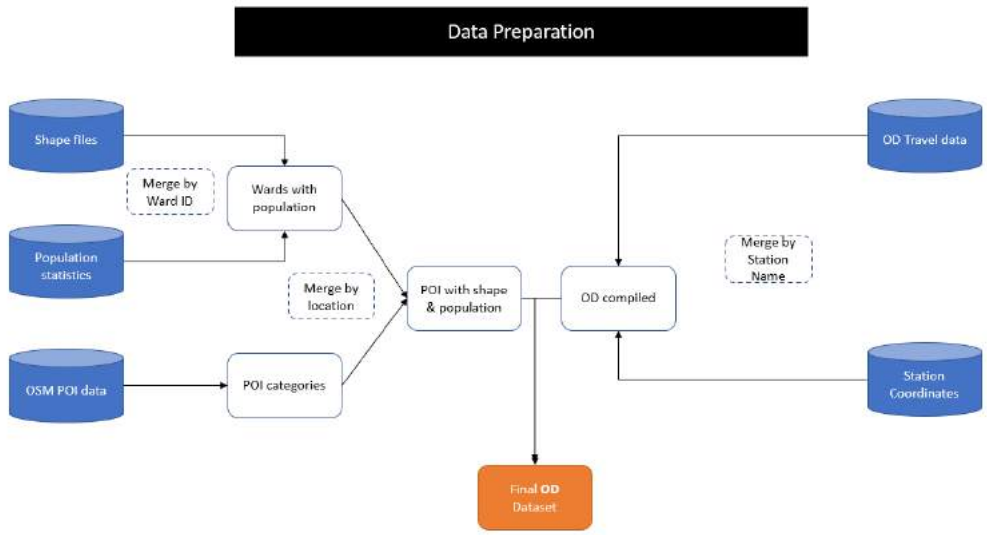

Figure 3.1: Data Preparation Framework

of the problem. The packages used for the analysis in Phase I (Data Preparation Framework) are showcased below in Table 3.3.

Table 3.3: Software implementation of the chosen algorithms in Python

No.	Type of technique	Package name	Reference
1	Data Analysis & Manipulation	pandas	[29]
2	Geospatial Data Analysis	geopandas	[13]
3	Data Extraction	requests, zipfile, io, os	[35], [46], [16] ,[28]
4	Analyzes OpenStreetMaps data	pyrosm	[31]
5	Datetime functions	datetime	[6]
6	Visualization	matplotlib	[21]
7	Statistics and Spatial Geometry analysis	scipy: spatial, stats	[36], [38]
8	Array manipulation	numpy	[25]
9	Spatial Planar Manipulation manipulation	shapely.geometry	[41]
10	Scaling, centering, normalization, binarization methods	sklearn.preprocessing	[1]

POI data

First, the shapefiles and population estimates are combined using the key ID OA11CD. Thus, we obtain the total population estimates for the different Outer Area (OA) codes. The smallest level of a spatial boundary in London is the ward. The ward code and corresponding names are available in the shapefile. The ward level is chosen as it comprises of the railway stations within themselves and it can provide a sound analysis for the points of interest available in the vicinity of the station as well. The population estimates are aggregated at a ward level `"WD11CD_BF"` leading to a dataset comprising of Ward Code along with corresponding spatial polygons and total population at the same level.

Figure 3.2 illustrates the wards of London and the color highlights the population corresponding to the ward. From the figure, it is observed that most of the population of the ward is between 10000 and 25000 citizens.

Next, the POIs data is analyzed. It contains all the various locations of London and the different variables to describe the type of building in the location along with the coordinates (latitude and longitude). The important variables considered are amenity, shop, building, and tourism. Amenity describes the useful and important facilities for visitors and residents and is the critical variable to be considered. Facilities include for example toilets, telephones, banks, pharmacies, prisons, and schools.

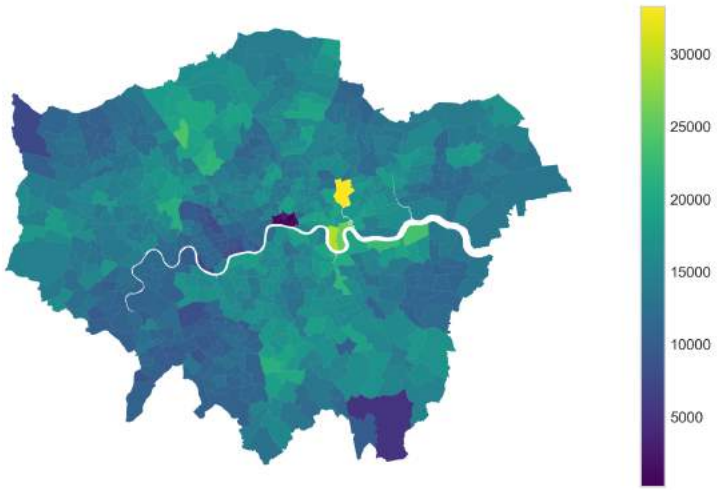

Figure 3.2: Population illustrated across wards of London

Shop contains commercial locations such as retail stores, supermarkets, and convenience stores. The Building variable describes the type of building in the location such as offices, churches, temples, schools, etc. Tourism is for the locations which are availed by tourists such as hotels, museums, and other attractions. A new variable `poi_type` is created which is the same as the amenity column and then the gaps in the column are filled from the remaining columns shop, tourism, and building in the corresponding order. The order is chosen as such to eliminate the gaps and complete the variable `poi_type`. From this, the necessary columns ID, `poi_type`, geometry (latitude and longitude complied together), and name are extracted.

From this extracted dataset, we observe that there are 117295 points of interest obtained for the city of London. This volume of data points is quite high and complex to integrate into a multivariate model and needs to be categorized. The wiki page of OpenStreetMaps for key: amenity showcases the different categories that the amenities are divided into [17]; inspired by this, a subjective categorization is created for all the points of interest. The different categories created for the POIs are Community, Commercial, Educational, Entertainment, Financial, Government, Healthcare, Miscellaneous, Sustenance, and Transportation. The categories are described in table 3.4 below. The category Miscellaneous is removed from the analysis as it contains amenities such as garbage bins, telephone, toilets, bench, etc which are unimportant and do not attract people to travel towards it.

Each POI belongs to one specific category. The motivation behind choosing the category for a POI is the activity behind it and the categorization present in the wiki page of OpenStreetMaps [17]. For example, religious places are primarily for the communities, and London is a multicultural city, and do not require financial consumption. Hotels are primarily for tourists and tourists visit places for entertainment; Hence, hotels are assigned for tourists. The assumption taken here is that this dataset encompasses all the POIs in the city of London and the OSM dataset acts as a proxy for the entire city of London. There are many limitations considered in this section. First, in a city such as London contains a huge proportion of working-class people who travel to offices during the weekdays. This is observed in the Peak timings for the weekdays in section 3.4. However, the data for the office locations, and the levels of office buildings, are not available and the available data in OSM is utilized and categorized as Commercial as it is a commercial land-use space. Each office/business can be considered as a point of interest as it is attracting people towards it irrespective of the purpose and London contains many tall structures containing multiple offices, implying numerous points of interest. There might be other offices such as factories, industries, construction work, etc for the blue-collar citizens who may use public transport frequently. There might be other points of interest that the OSM data does not capture and is a major limitation to consider. Land-use data can be used to add additional information regarding

the POI data but the data was not available in the format required. However, OSM data has the column 'landuse' and is utilized in the categorization of POIs. These are the limitations considered in the spatial aspects. The above-mentioned people/POIs might be missed out from the analysis and this is a limitation in the study to be noted.

The final limitation is the time aspect of the points of interest. The dataset is considered to be the same throughout the week, indicating that all the places are available to people at all times throughout the week. However, this is not the case in real life. Offices are open during certain timings, businesses operate in a vast range of time schedules, schools and government offices follow a particular schedule. This can be categorized for the type of population as well. Having said that, that is not included in the study due to time and data availability limitations and can be explored in future works. If the data is available, then machines with high-performance capabilities would be required and the data would be considered as big data. The difference is also noted between weekdays and weekends. Usually, weekdays have a similar pattern and Saturdays are open longer for some locations while on Sundays lots of locations close early. This pattern is observed in section 3.4. Having said that, these limitations are important to consider and assumptions are made for the research to understand and solve the problem.

Table 3.4: POI categories summarized

No.	POI Category	Count	Description	Examples
1	Commercial	28835	Commercial spaces such as Shopping, Supermarkets and Offices	convenience, clothes, office
2	Transportation	21742	Transportation related amenities	bicycle rental, parking, car sharing
3	Sustenance	20980	Amenities pertaining to food and drinks	restaurant, cafe, fast food
4	Miscellaneous	19972	Random amenities	garbage bins, telephone, toilets, bench
5	Government	7890	Amenities owned by government	post office, police, public buildings
6	Community	5572	Places specific for the communities	place of worship, community center, charity
7	Educational	4473	Amenities involving education	school, library, college
8	Healthcare	3162	Amenities concerning healthcare	pharmacy, dentist, doctor
9	Financial	2524	Banks and ATMs' predominantly	ATM, bank, currency exchange
10	Entertainment	2145	Amenities most attractive for tourists and recreational	art, theatre, museum

Now, we have two datasets; 1) comprising the shape of London at a ward level along with corresponding population estimates and 2) the points of interest with a POI category and corresponding spatial locations. These two datasets are merged using a spatial join using the geopandas library of Python. The spatial join chosen is intersection and with this, the POIs are embedded onto the polygons depending on their locations in the wards. The POIs are aggregated together at a ward level and the counts for different POI categories are obtained per ward. This leads to the desired dataset (POI dataset) with counts of POI categories at a ward along with the total population and spatial features.

This POI dataset is analyzed and illustrated in figure 3.3 below similar to the population represented in London. We can observe that for categories Government, Sustenance, Transportation, Community, and Commercial, the POIs are spread all across London while the remaining categories are sparsely spaced. Though the ranges are different for all categories, the center of London always seems to contain the highest proportion of amenities. This gives us an overview understanding of the distribution of POIs across the city of London.

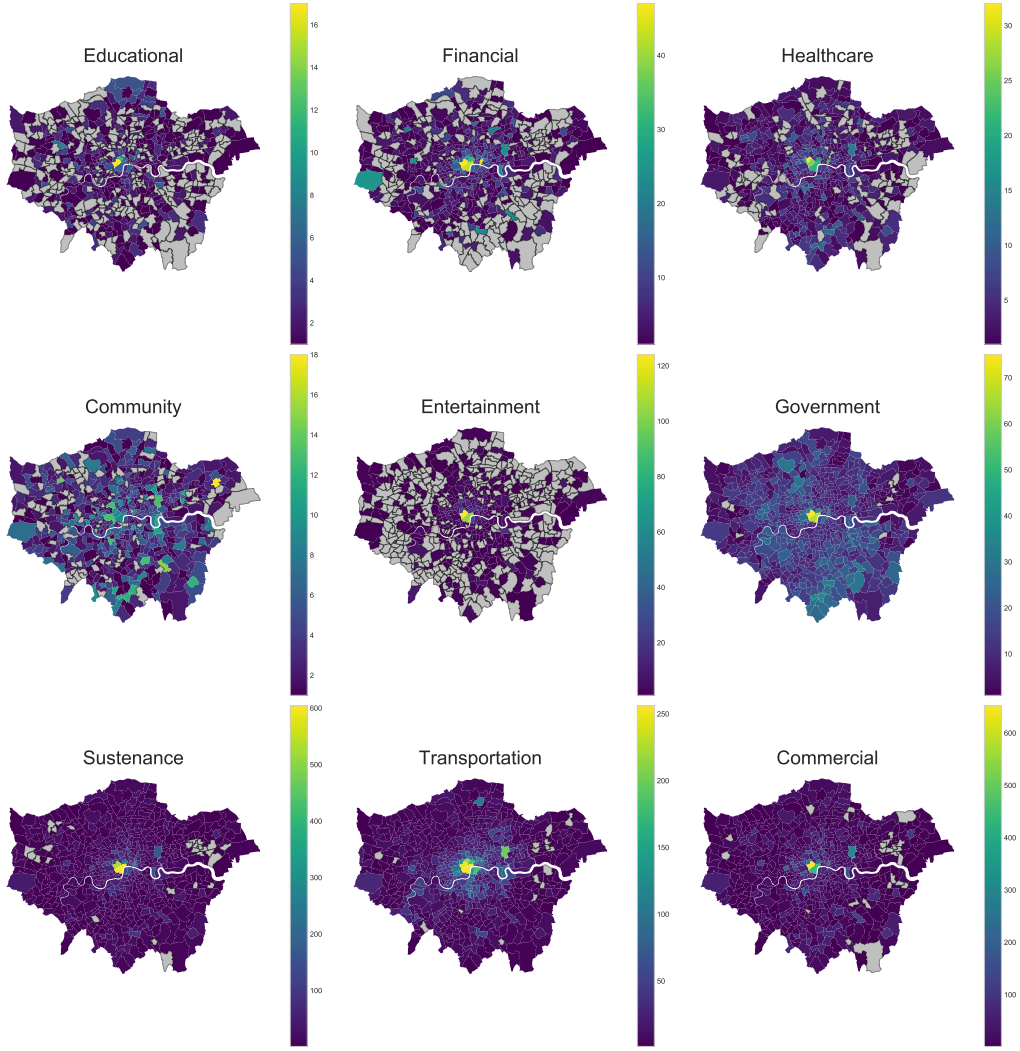

Figure 3.3: POI categories illustrated across wards of London

Modifiable Areal Unit Problem

It is evident that all the wards are outlined on the map; however, the wards are not on the same scale. They contain different area sizes and total population values. In this study, it is critical to get the right resolution or scale of the area to analyze properly. By using the official statistical values, it is ensured that values are empirically right. However, to analyze the summary of different variables such as total population per ward would be misleading as each ward is of different size and would lead to different values for the comparable metrics for the scale and shape of aggregation. This problem is quite evident for statisticians and researchers and it is deemed as Modifiable Areal Unit Problem (MAUP). MAUP affects results when point-based measures of spatial phenomena are aggregated into districts, for example, population density or death rates. The resulting summary values (e.g., totals, proportions, rates) are influenced by both the shape and scale of the aggregation unit [20]. To tackle this issue, the whole region can be divided across a custom grid of identically-sized units and the values distributed across the unit would be comparable. The units chosen are hexagon and the map is modified to a series of hexagons with identical shape and scale.

A map can be divided into regular shaped grids comprising of any shape such as square or any kind of polygon. Hexagon is chosen for the following reasons. When comparing polygons with equal areas, the more similar to a circle the polygon is, the closer to the centroid the points near the border are (especially points near the vertices). This means that any point inside a hexagon is closer to the centroid of the hexagon than any given point in an equal-area square or triangle would be (this is due to the more acute angles of the square and triangle versus the hexagon). Hexagons are also preferred

for analysis regarding connectivity and movement. For analysis regarding large areas, a hexagon grid will suffer less distortion due to the curvature of the earth than the shape of a fishnet grid[44] [53]. For the reasons above, hexagons are determined to be the best fit for solving the Modifiable Areal Unit Problem(MAUP) in this research.

The total population is adjusted to the hexagon based on the following method. The population of a hexagon is the summation of the proportion of wards' area intersecting in the hexagon times the ward population. For example, if the hexagon is comprising 10% area of a ward, then the population of the hexagon would constitute 10% of the population of the selected ward. In this way, the area and population of the hexagons can be determined.

As the hexagon is a polygon by itself, it contains a spatial boundary, and the POIs comprising in the hexagon can be obtained. From this, we can obtain the following necessary information for a hexagon: Population and Counts of POIs for the various categories. This is illustrated and discussed in the next section.

Exploratory Analysis

Figure 3.4 below represents the juxtaposition of distribution of the population of London across hexagons and wards. We can observe that the maximum population value for the hex-grid is 25000 and the ward is 30000. This shows that the hexagons are at a smaller level than the wards. Though that might be the case, the transformed hex-grid looks similar to the ward distribution in terms of the distribution of the population.

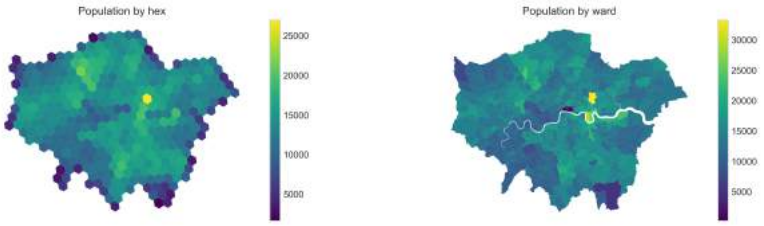

Figure 3.4: Population distributed across London comparison - Hexagons vs Wards

As the hexagons are spatial geometric bodies, the POIs depending on the location can be allocated to the hexagons using the geopandas package of Python. To illustrate all the categories on a similar scale, the data has been scaled using a QuantileTransformer function from the sklearn package in the preprocessing module. This method transforms the features to follow a uniform or a normal distribution. Therefore, for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of (marginal) outliers: this is therefore a robust preprocessing module [1]. This transformer is chosen over the other transformers as it would represent the POI categories in the best way possible given the outliers and representation of the frequent values over a large region. This is represented in the figure 3.5 below. We can observe that Transportation, Sustenance, Educational, Commercial, and Healthcare are widely distributed across all over London with Financial being the least widespread. For Financial, this seems to be right given the prominent amenities are ATM and bank only. The key observation identified is that the heart (center) of London seems to be the most prominent boasting the highest POI proportion compared to the rest of the city of London for all the POI categories.

For the rest of the categories, the outer regions seem deprived of the amenities available in the hexagon. Though given the purpose of the POI, the POI need not be available everywhere. For example, a government office or an amusement park (Entertainment) are not needed daily and can be available at a reachable distance within the city. This indicates that the city of London, in terms of accessibility, is bountiful of available POIs for the citizens.

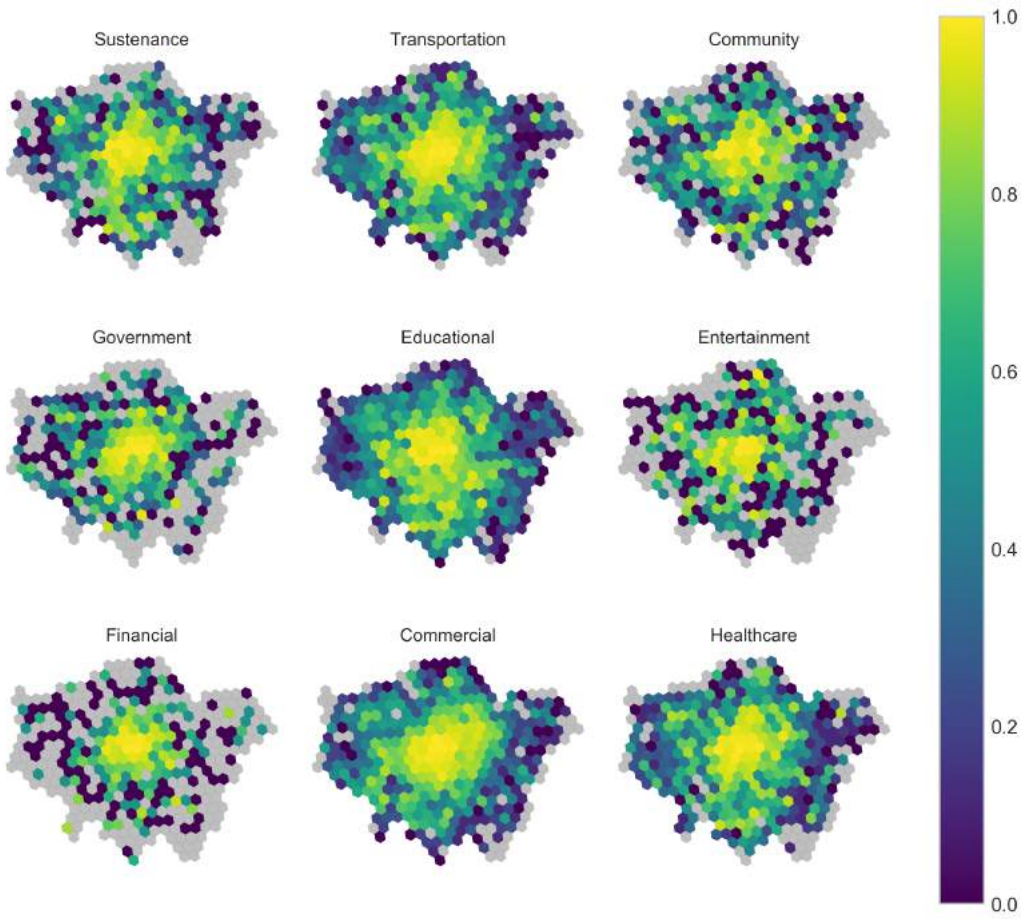

Figure 3.5: POI categories illustrated across hexagons of London

The scatter plots 3.6 below help us understand the distribution of amenities per hexagon. The scatter plots with linear regression is available in the appendix 8. Linear regression is fit to the plots to measure the goodness-of-fit metric. The plot in the appendix indicates that linear regression is not a good fit and multivariate regression is of absolute necessity to obtain good results. The number of hexagons the corresponding POI categories constitute is also available above each POI category of the scatter plot. This gives a closer look and an overview understanding of POIs across the hexagons categorically. The following POI categories are available in over 300 hexagons: Transportation, Sustenance, Government, and Commercial. In the scatter plot, the categories Sustenance and Commercial seem to follow a similar pattern, for the same range (1000) as well. Similarly, Community and Educational (both low range) categories exhibit a different pattern from the rest.

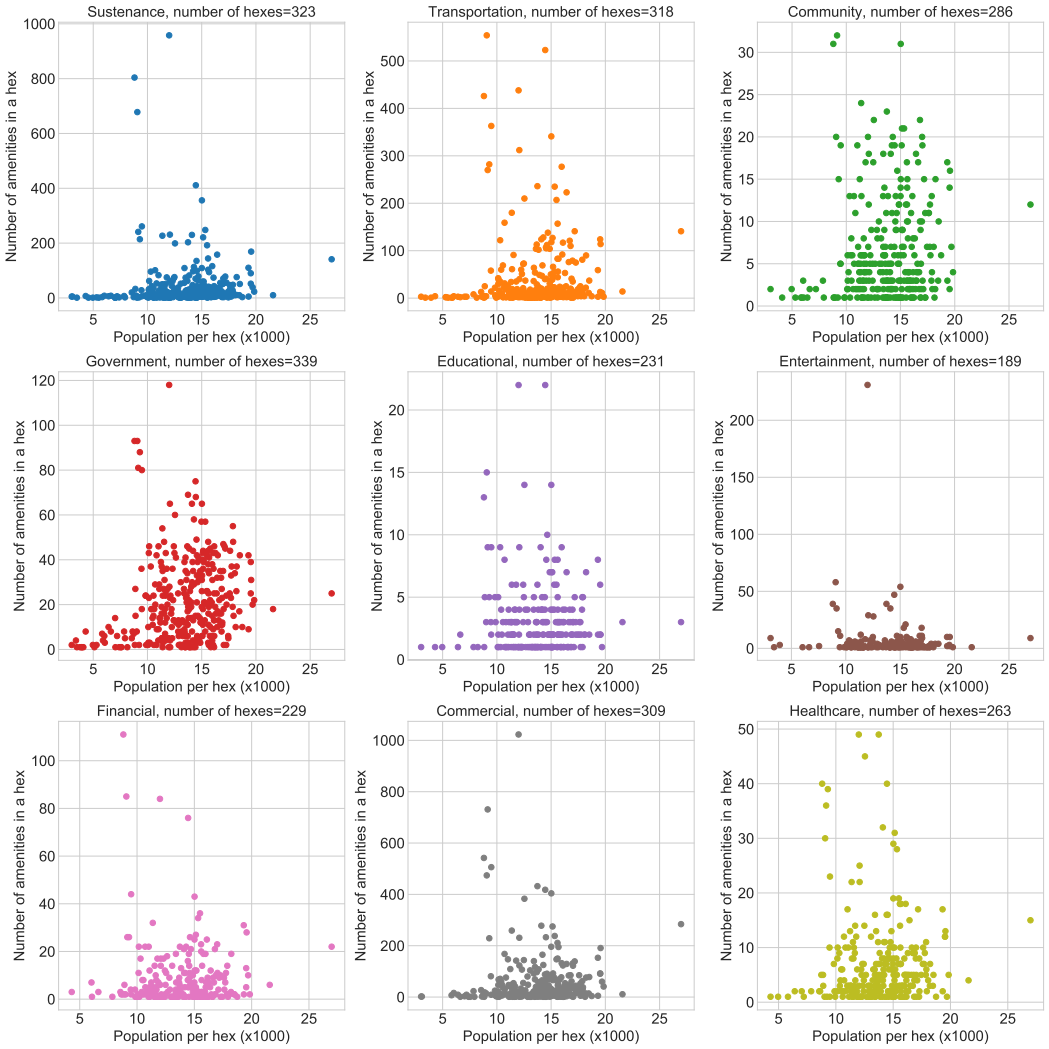

Figure 3.6: Scatterplots of POI categories across hexagons of London

From the heatmap below 3.7, we can observe the affinity of POI categories among themselves. Category Commercial has a high affinity towards Sustenance, Financial, Healthcare, and Transportation. This indicates that the POIs in Commercial (Shopping and Offices) are available in similar regions towards the POI categories of Sustenance, Financial, Healthcare, and Transportation. The highest values (greater than 0.9) observed are between the pairs (Commercial, Sustenance) and (Financial, Sustenance). This is also observed in the scatter plot ?? above. From a critical perspective, it can be said that shopping, offices, restaurants, eateries, and ATM exist within the same range prominently. This is observed in the real world. Entertainment has the highest affinity towards Sustenance and on a conceptual level, this makes perfect sense as people who perform Entertainment activities usually go to a bar/restaurant for food and drinks. Financial has a high affinity towards Sustenance indicating that ATMs' are usually around places offering food and drinks. It is also observed that Sustenance and Healthcare have high correlation values for all POI categories. Though Healthcare and Sustenance do not have an affinity, this indicates that these POI categories are much closer to POIs from other categories than the rest.

The lowest values (less than 0.6) observed are between the pairs (Entertainment, Community) and (Government, Entertainment). Community amenities exist usually in residential areas and Government buildings are public properties that exist in allocated locations. It would be safe to say Entertainment amenities would not exist near these locations and would require from those amenities locations to Entertainment locations. Comparing to the scatter plots ?? above, the patterns exhibited between these categories are not similar in any way. The Educational and Community categories though looked

similar in the scatter plots, the heat map value (0.65) does not affirm that hypothesis and tells us that the correlation between them is not high as it seems. By putting together the different perspective visualizations and analysis, we can understand the relationship between the POI categories in a deeper sense. Thus, this heatmap validates that the data collected is imitating real-life observations.

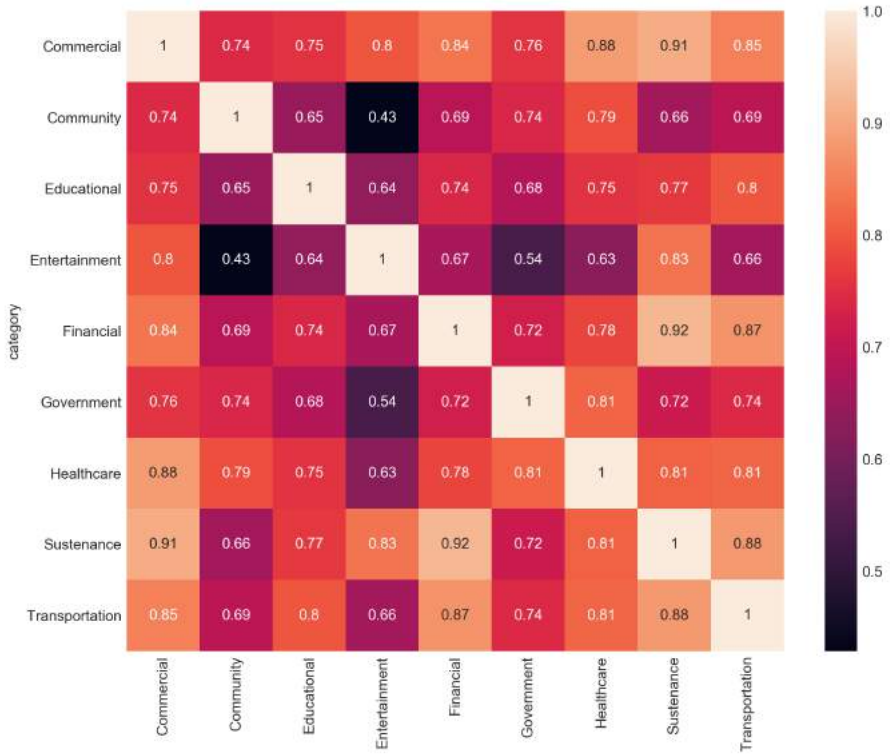

Figure 3.7: Heatmap of POI categories across hexagons of London

Travel data

The OD data files which contain the days of the week: MTT (Monday to Thursday), FRI (Friday), SAT (Saturday), and SUN (Sunday) for the different modes of rail network: London Underground (LU), London Overground (LO) and Docklands Light Railway (DLR). All these files contain similar data structure - The mode, Origin station codes and name, Destination station codes and name, count of people traveling between origin and destination across eight-time frames apart from the Total value. The time frames are as following: Morning (0500-0700), AM Peak (0700-1000), Inter Peak (1000-1600), PM Peak (1600-1900), Evening (1900-2200), Late (2200-0030), Night (0030-0300), and Early (0300-0500). All of the OD files are compiled together to form a single OD dataset. However, this dataset does not contain the spatial locations of the stations.

By using the Station Coordinates (SC) file, the locations of the stations in the OD dataset can be determined. There are a few mismatches given the names can differ in the datasets. For example, in the SC file, the name might be 'West Hampstead LO' for mode LO but in the OD file, the name is 'West Hampstead'. Such mismatches are observed and handled via renaming the station names. After this, there were a few observations where the station coordinates were not available for the stations in the OD dataset. This is handled by manually entering the coordinates observed from the Doogal website [19]. This website contains the spatial coordinates for all the stations in the city of London. Finally, by merging both the datasets we can obtain the OD compiled information along with the origin and destination spatial coordinates.

Exploratory Analysis

The figure below 3.8 showcases the Origin-Destination Flows across different days: MTT, FRI, SAT, and SUN for the modes: LU, LO, DLR. It is evident that most of the travel volume is for the mode London Underground (LU) by a huge margin and then followed by DLR and LO. MTT and FRI follow a similar pattern and this is the case of SAT and SUN as well; this proves to us that the travel flows differ between the weekdays and the weekend. For MTT and FRI, the first peak occurs at the AM Peak (0700-1000) and the second peak occurs at PM Peak (1600-1900). This corresponds to the travel volume people go to work in the morning and the evening, they either get back home or travel elsewhere after finishing their office duties. For SAT and SUN, there is only one peak at Inter Peak (1000-1600) and this shows us that on the weekends, people relax in the morning and go out to explore other activities at noon. There is one critical difference to be noted. On Sunday, the maximum flow observed is almost 1.2 million people while for all other days it is around 1.4 million people. This indicates that 200,000 people do not travel or do not use the rail on Sunday. For all days, after the peaks, the travel volumes fall drastically as the day turns to night.

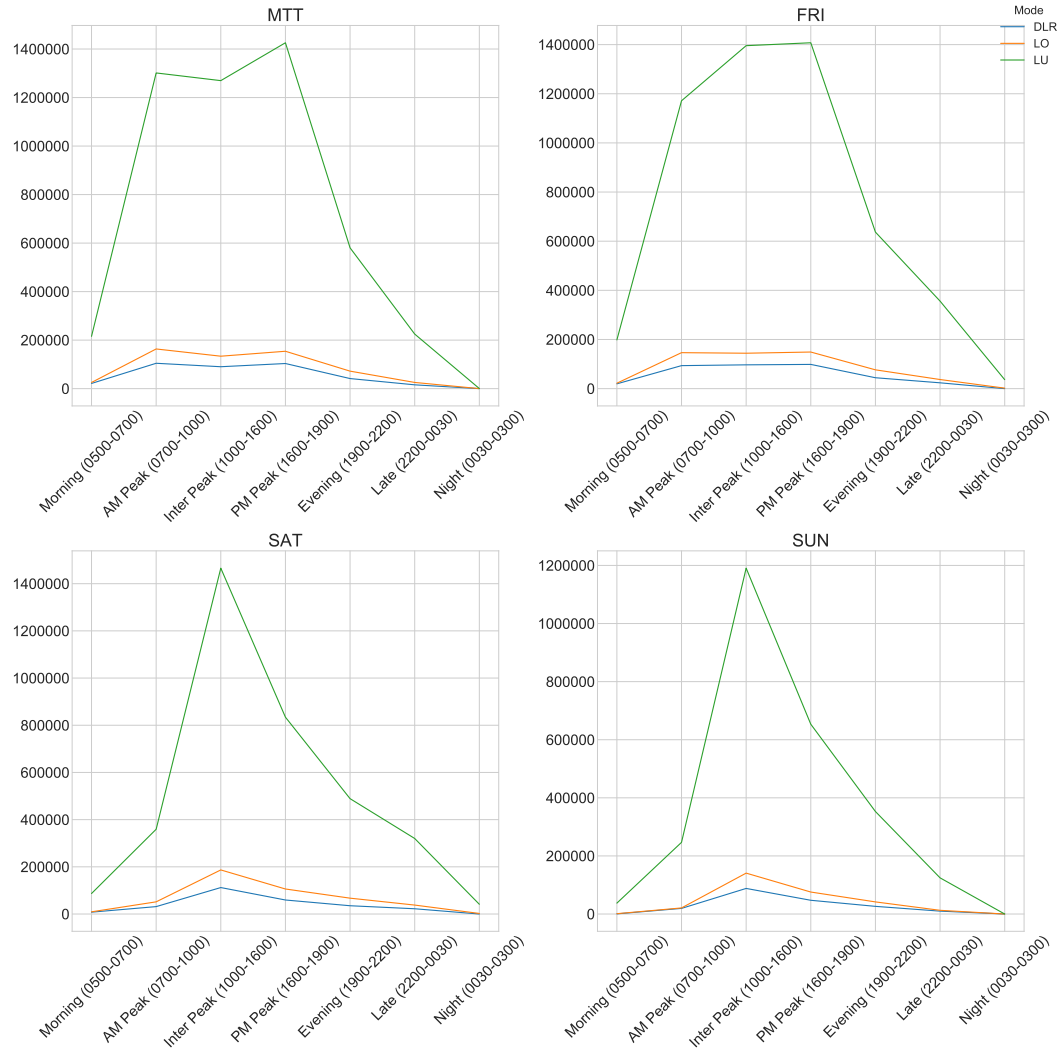

Figure 3.8: Origin-Destination Flows across different days: MTT, FRI, SAT and SUN for the modes: LU, LO, DLR

The below figure 3.9 highlights all the hexagons in which the rail stations of London, for the modes selected, are available. It is observed that the rail network is not spread across all of London and the heart (center) of London has a well-connected system. There might be other transportation networks for the rest of the non-rail hexagons of London which seems to be an extension of the main London region. For this research, the hexagons containing rail stations are considered as the travel volume is necessary for the study.

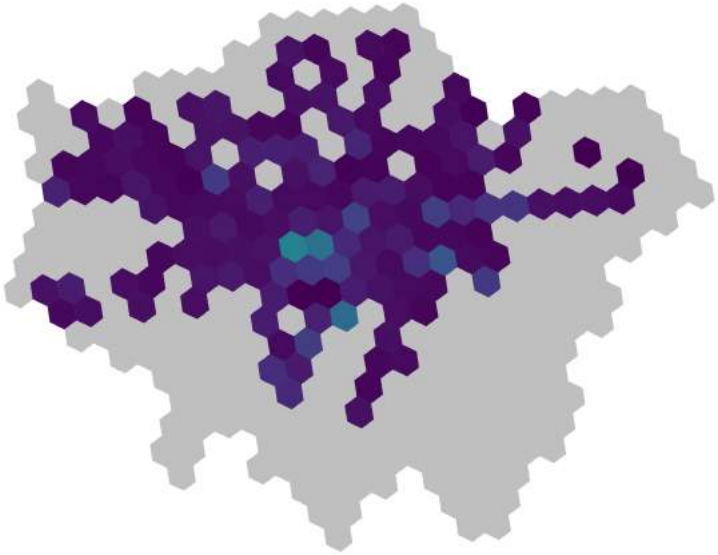

Figure 3.9: Hexagons representing the rail stations of London. The highlighted hexagons contains rail stations within them.

Final OD dataset

The two important datasets for the process are extracted, analyzed, and modified. By merging the POI and Travel datasets, we can obtain the final dataset. The station locations can be pinpointed in the hexagons and we can obtain the flow between the stations for the same. It is observed that there are multiple stations in a hexagon and this leads to multiple flows arriving in and out of the hexagons. It is important to have the flow at a hexagon as the research objective is about the influence of POI on mobility. Thus, the flow between hexagons can be obtained by aggregating the flows between stations residing across the hexagons in London. By summing up all flows at a hexagon level, we can finally obtain the total flow between hexagons for the research.

At this stage, we have the information at a hexagon level comprising of Total Population, Counts of POIs for the various categories, and Flow between them. The only missing variable is the distance between the hexagons. The critical assumption undertaken in terms of the calculating distance between hexagons chosen for the research is the distance of the centroid of the origin hexagon to the centroid of the destination hexagon. This is calculated using the geopandas package in python. After the calculations, we have the complete final OD dataset necessary.

It is observed that the different days of the week represent different travel flows and especially, the difference is between the weekdays and the weekend. By combining them and summarising them to a single day would not help us understand the above difference. This phenomenon is known as Simpson's paradox [78]. It is observed in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. Hence, the final OD dataset is divided across for all days: MTT, FRI, SAT, and SUN, and the multivariate regression models would be built for the same instead of having only one model. Thus, for this research four models will be developed for each of the days.

Exploratory Analysis

First, the day MTT is chosen for exploratory analysis. The POI distributions below are the same for all days as the difference between the days is only observed for the travel flows. Figure 3.10 below shows the frequency of hexagons for the POIs available per POI category For Commercial category, we can observe that there are around 110 hexagons with up to 100 commercial POIs in the hexagon. There are a few hexagons with high POIs count and can be seen for all POI categories. This can be indicated as the center or most popular area in London. For categories - Commercial, Sustenance, Transportation, and Entertainment it is observed that there is a large number of hexagons with low POIs count. For the categories Community and Government, there is an adequate number of POIs for the range of hexagons

available. However, they also contain a lower maximum range of count of hexagons compared to the rest of the categories. This is similar to the observations presented in the previous section.

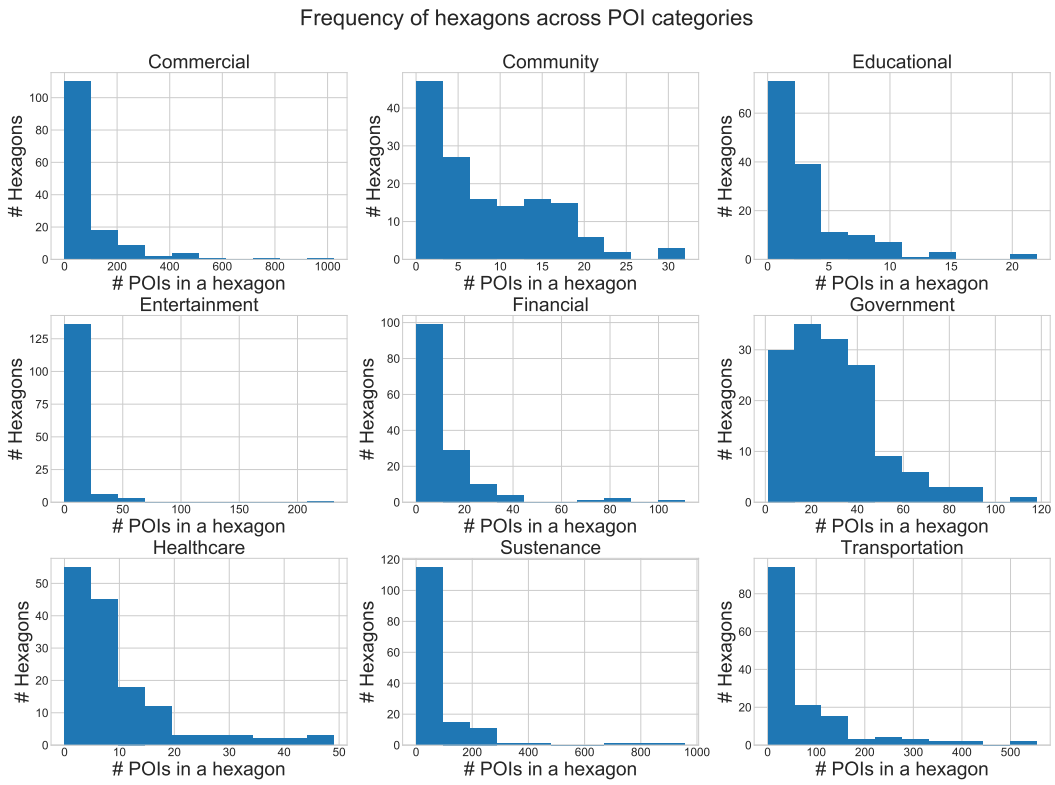

Figure 3.10: Frequency of Hexagons for POIs available across POI categories

Next, the figure 3.11 illustrates the POI categories distribution across hexagons of London. This is similar to figure 3.5 but in this case, the hexagons are highlighted if they contain rail stations within them and rescaled to fit the data. The MinMaxscaler function from the Preprocessing module of sklearn is used to get the categories on a similar scale. The transformer is chosen as transforms features by scaling each feature to a given range. It is observed there a few hexagons at the center with high POI distribution value while the outer regions contain the minimum value. This is similar to the figure 3.5 in the terms of the distribution of POIs. The hexagons in which the rail network exists has similar coverage of POIs for the city of London. Hence, the research study can be easily extended by adding different travel models.

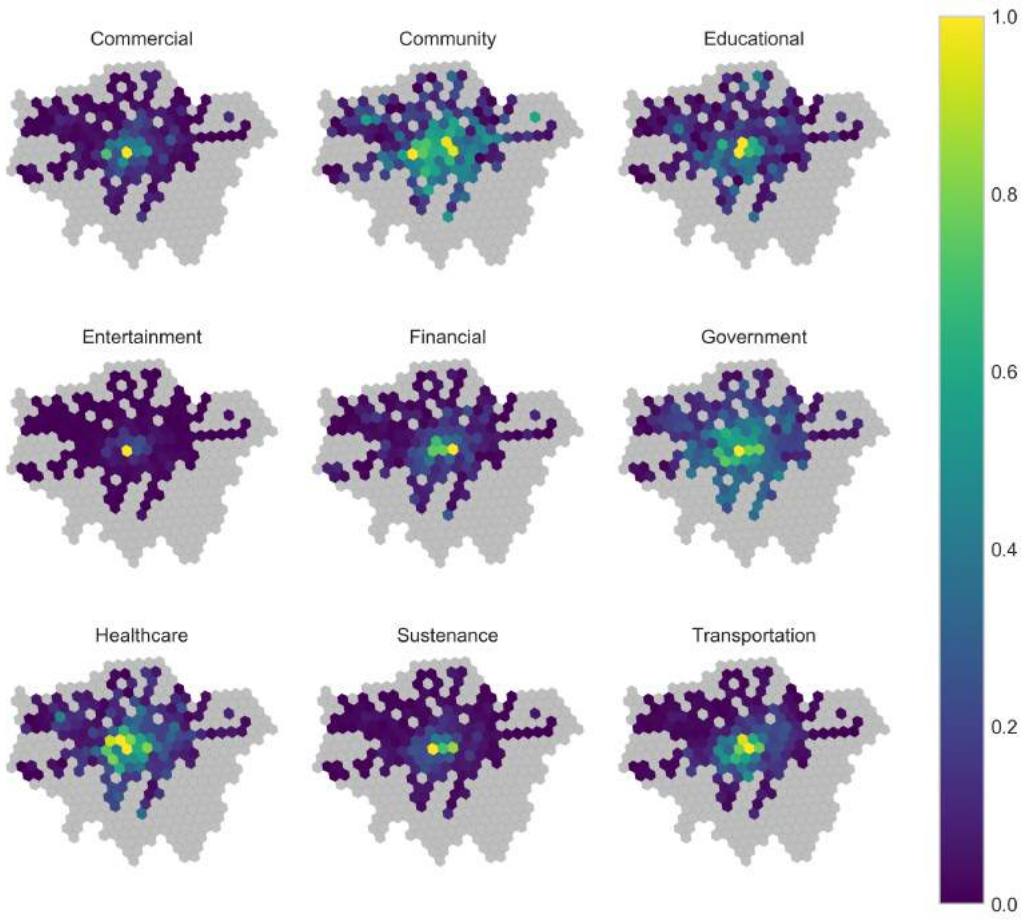

Figure 3.11: POI categories illustrated across hexagons of London

The POIs available per POI category across distance quantiles are examined in figure 3.12. This helps us understand the distance traveled by a person towards a destination hexagon comprising the POIs. For instance, a person has more than 4000 Commercial POIs within 16 km distance and has more choices within the Commercial category. As the distance increases, the counts of POIs available decreases drastically across almost all categories. This affirms that the data collected is accurate and a good fit for the estimation. It is interesting to note that Education and Sustenance POIs are distributed quite evenly irrespective of distance, implying that these POIs are essential for everybody.

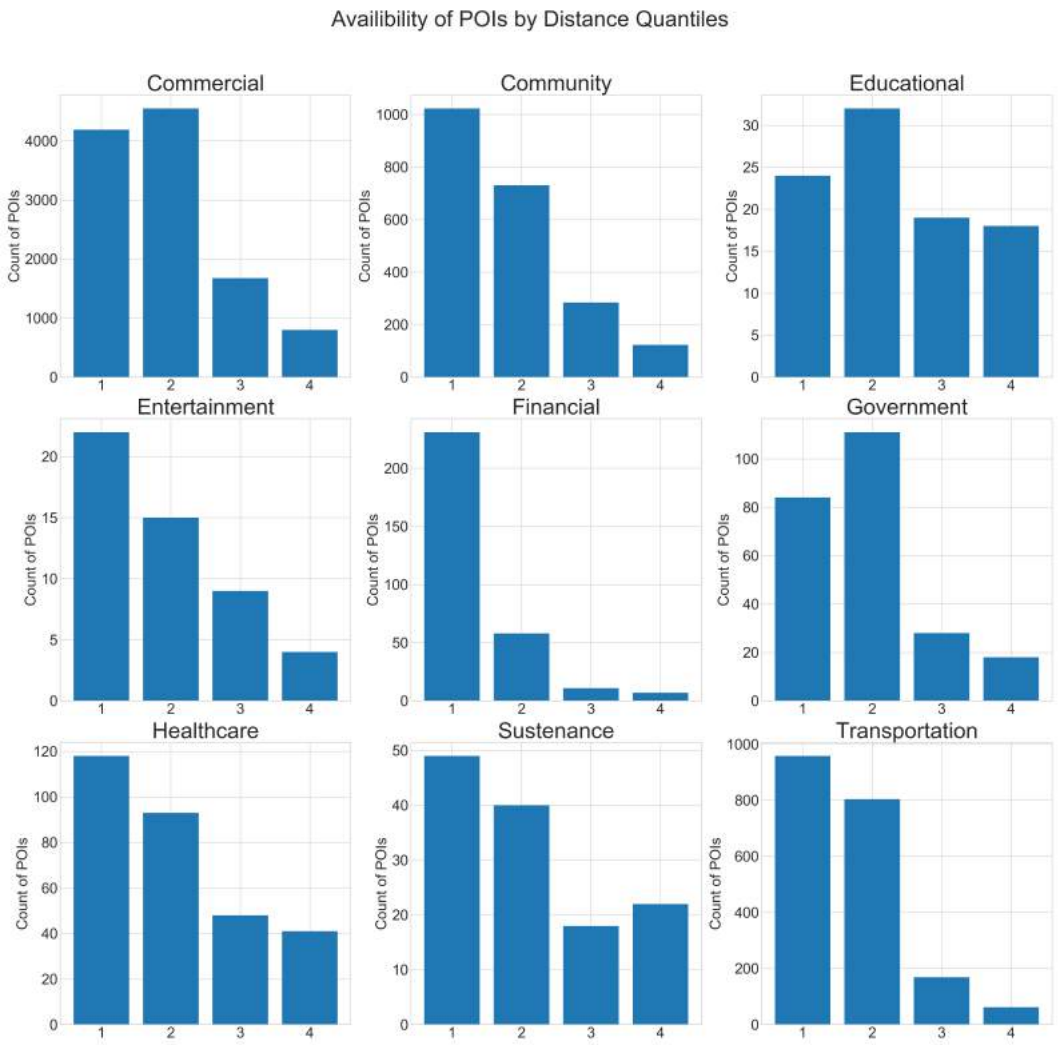

Figure 3.12: POIs available per category for a destination hexagon by distance quantiles

The travel data for all days are available for 8-time frames divided across a day with various periods. The time frame is described in the table below 3.5.

Table 3.5: Time frame distribution in a day - OD dataset

No.	Time frame	Time
1	Morning	0500-0700
2	AM Peak	0700-1000
3	Inter Peak	1000-1600
4	PM Peak	1600-1900
5	Evening	1900-2200
6	Late	2200-0030
7	Night	0030-0330
8	Early	0330-0500

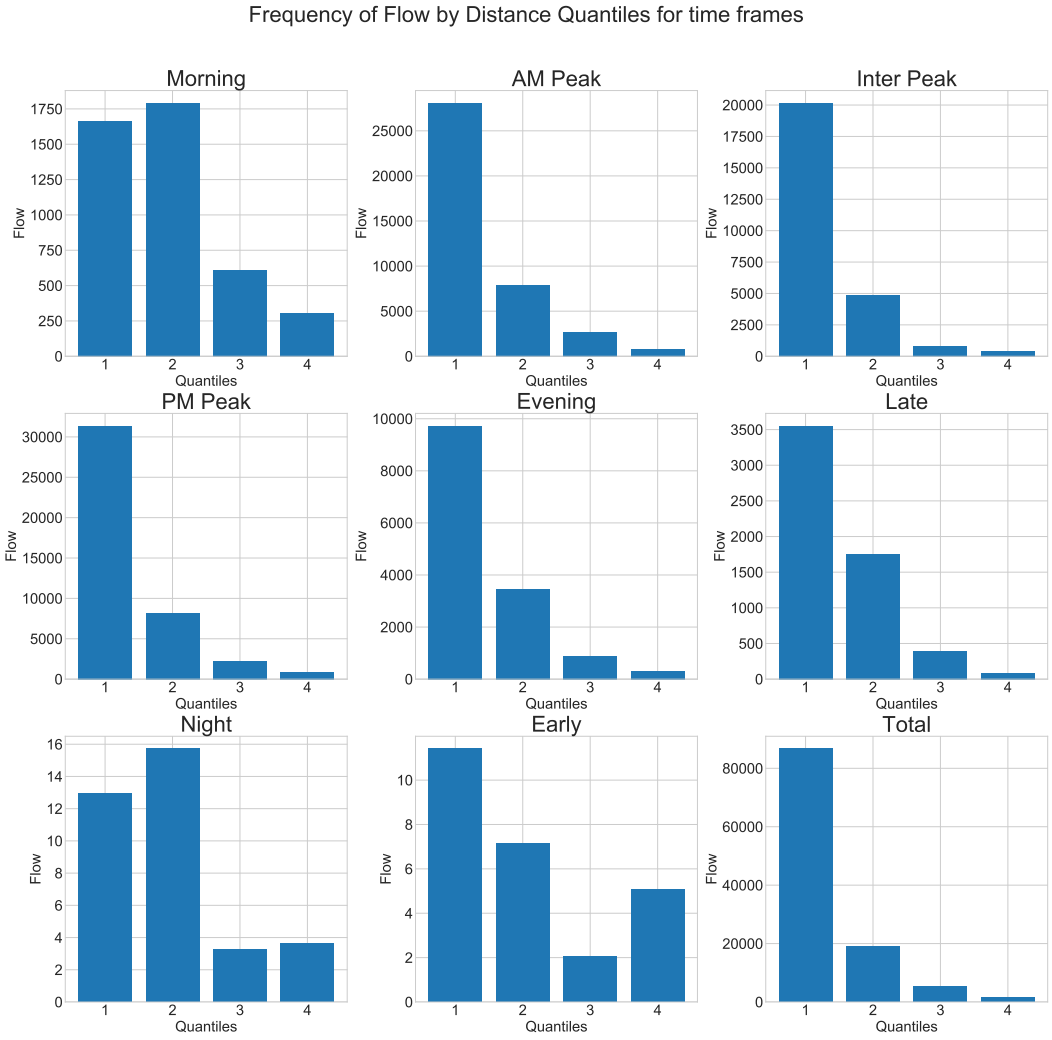

Figure 3.13: Distribution of flow for time frames of the day across distance quantiles - MTT

Figure 3.13 above shows the distribution of the flow of people traveling in different time frames of the day MTT across distance quantiles, the total distance is 80 km and each quantile valuing at 20 km. On examination, we find that, for Morning and Night, Quantile 2 is more than Quantile 1 which indicates that people are traveling up to 40 km in these time frames. In the rest of the time frames, the flow is the highest for quantile 1 (20 km) by a large margin and deteriorating as the distance increases.

To get a closer look, the same distribution is replicated for deciles where each decile consists of 8 km in figure 3.14. The Peak periods - AM Peak, Inter Peak, PM Peak along with Evening and Total seem to exhibit a similar pattern; The decile 1 being the largest flow and gradually decreasing for every decile. An interesting aspect to note is that the flow for decile 5 (32-40 km) is more than that of decile 4 (24 -32 km) for all time frames. This could indicate that people with closer proximity chose not to use the rail network and prefer other modes of transport (including private).

This graph helps us understand the usage of the rail network across London for different time frames. It shows the necessity of the rail network for people to travel to their respective destinations depending on the time of the day. The distributions if studied in detail, might help in predicting similar patterns for future projections and allocate travel demand volumes appropriately.

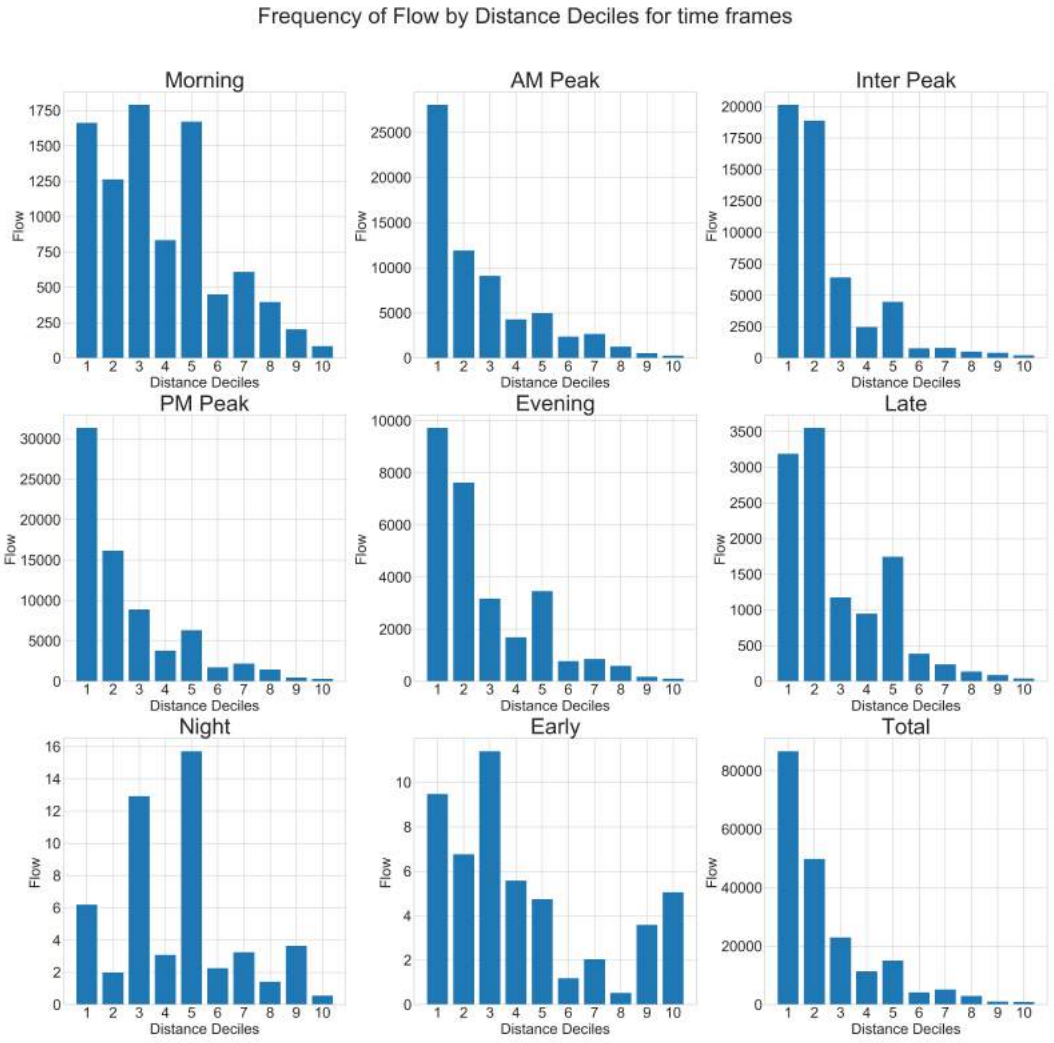

Figure 3.14: Distribution of flow for time frames of the day across distance deciles - MTT

For the different time frames of the day, the frequency of hexagons for corresponding flow values is illustrated in figure 3.15. It is observed that over most of the hexagons have had similar flow values per destination and a few hexagons have high flow values. This confirms the above observation that most of the hexagons follow similar characteristics. Similarly, the flow is illustrated on a similar scale for all time frames across the hexagons of London in figure 3.16. This helps us understand the location of the hexagons with high flow distributions, which seems to be the center. The center is observed to contain high counts of POIs as well as flow value compared to the rest of the city.

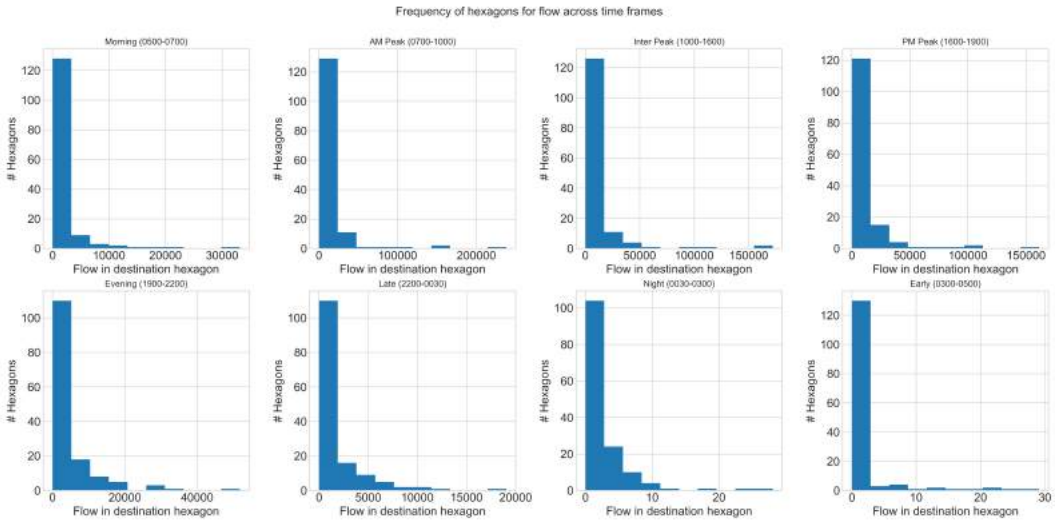

Figure 3.15: Frequency of Hexagons for travel flow across time frames of the day - MTT

The distributions seem to be power-law with extremes on both tail ends for both the axes. The rest of the days follow similar characteristics and hence are not showcased in the main section. They will be available in the appendix chapter 8. As the hexagon is at a bigger level than the station radius, the travel volumes are aggregated and representing similar characteristics for any time frame. If the level of a hexagon is much smaller, then different distributions might have been observed. For this research, it is assumed that all the time frames follow similar patterns in terms of distribution and would result in similar estimated values. However, experiments would be conducted to observe any key differences between estimating among the time frames.

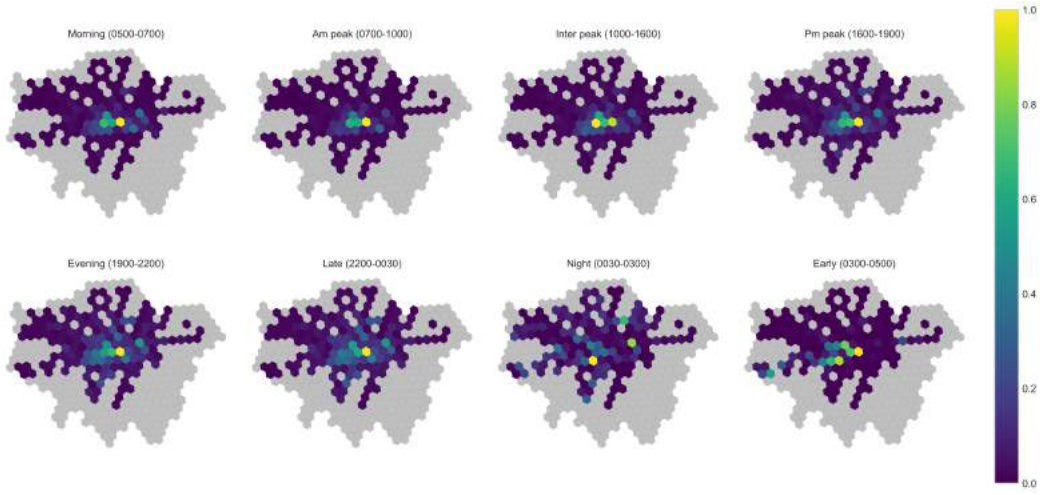

Figure 3.16: Flow distribution illustrated for time frames of the day across hexagons of London - MTT

The below figure 3.17 showcases the average distance traveled to the destination hexagon. On average, people travel 20 km to around 50 hexagons in an MTT day. After that, the frequency drastically

diminishes as distance increases. This indicates that 20 km is the optimal distance a person chooses to travel in the city of London using the rail network.

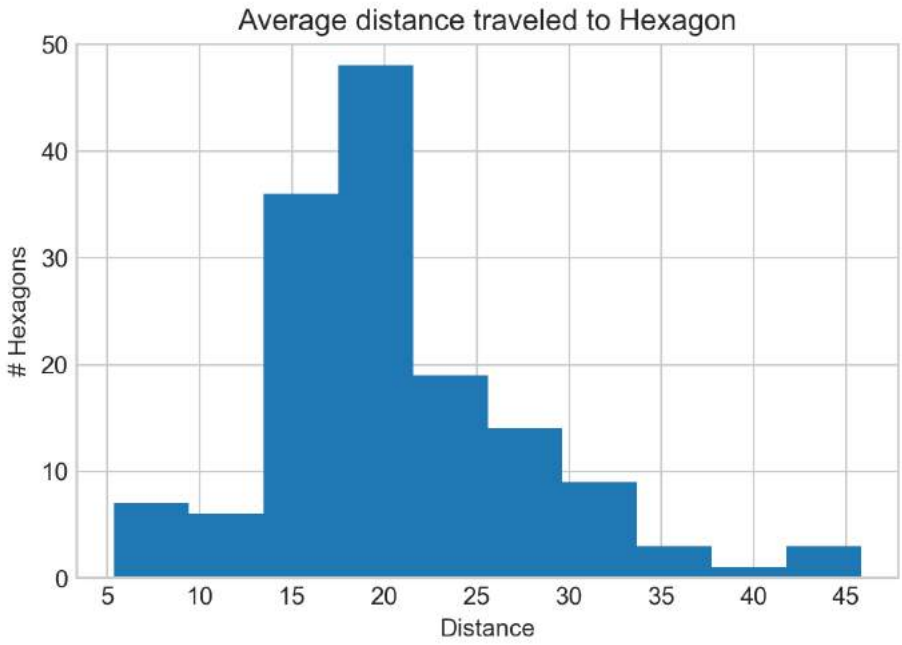

Figure 3.17: Average distance traveled to a destination hexagon - MTT

The distribution upon testing generated the following results represented in 3.18. Cullen and Frey's graph is used to determining the best possible fit among the distribution families [30] [24]. It could follow Weibull, lognormal, or gamma distributions depending on the observation and corresponding bootstrapped values. Depending on the distribution, data can tell us intricate information at a hexagon level and could be used to obtain accurate estimations. This can be explored for future work. The rest of the days - FRI, SAT, and SUN follow similar characteristics and are available in the appendix 8 for reference.

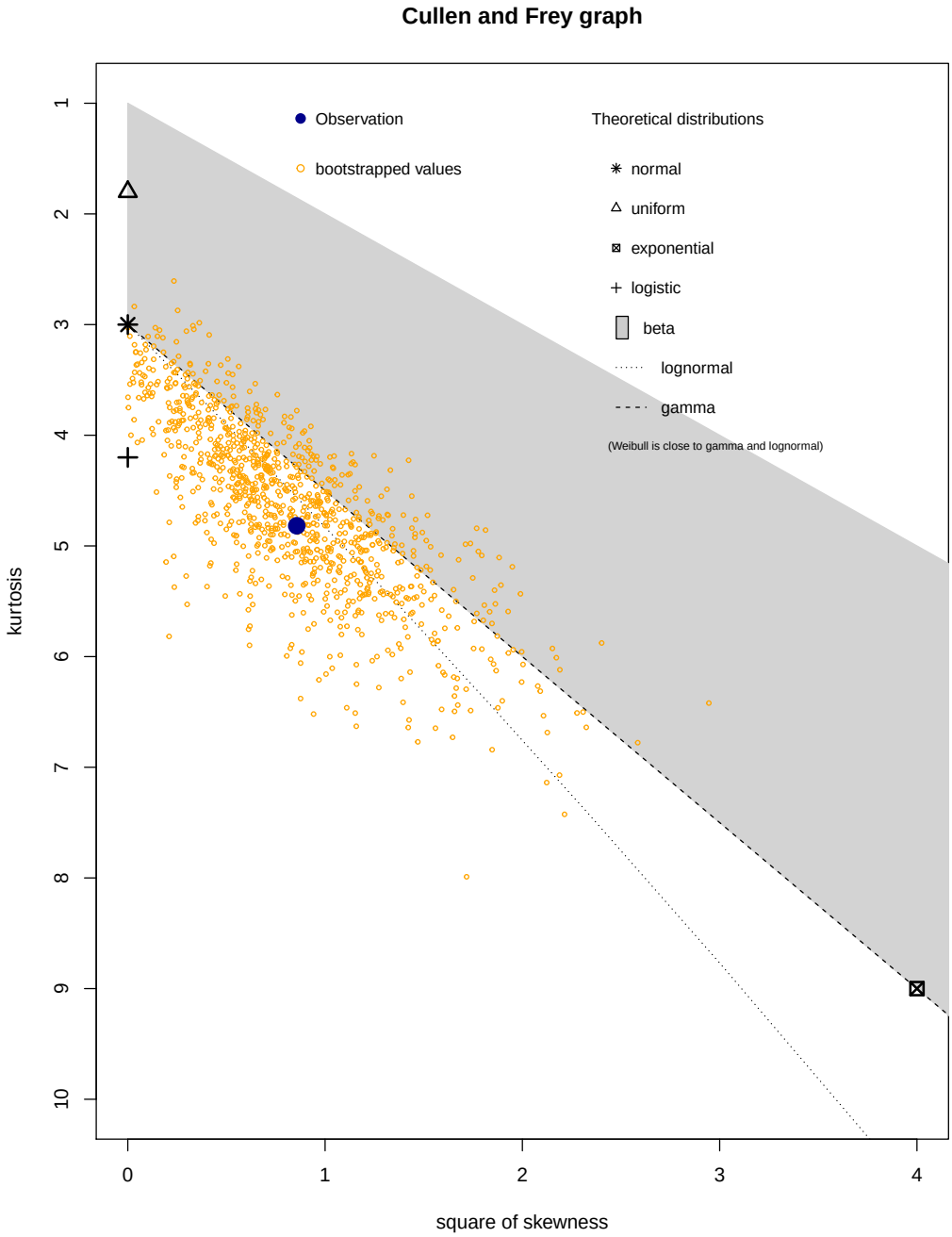

Figure 3.18: Cullen and Frey graph for obtaining the type of distribution represented in figure 3.17

3.5. Discussion

This chapter explored the individual datasets obtained from multiple sources and established the final OD dataset. The POI data were categorized and the relationship between the different POI categories was explored. It was determined that the data follows real-life characteristics and is ideal for the research given the limitations and assumptions. The travel data was briefly explored on a day (MTT, FRI, SAT, and SUN) basis and it was observed that there are no critical differences in the behavior patterns between them. The exploratory analysis of the final OD dataset was further conducted. The key insight generated is the travel flow follows the notion that the travel flow decreases with an increase in OD distance. This affirms that the dataset obtained is following the gravity law, as discussed in chapter 2. Further exploration tells us that the center is the heart of the city and it attracts high travel volumes as well as a high distribution of POIs compared to the rest of the city. The time frames visualizations showcase that travel volumes differ greatly across the different time frames of the day and distance plays a critical factor in determining the travel flow for the specific times.

It is also determined that linear regression (available in appendix 8)is not adequate and multivariate regression is necessary to solve our problem. In chapter 2, it was determined that multivariate regression is the model technique selected and POI categories will be input variables to the model. The POI counts are obtained for each POI category and they are categorical variables. Thus, the models are to be built for all the different days, and the models that input POI counts as variables are to be determined. This concludes this chapter and the methods for model development are discussed in the next chapter 4.

4

Methodology

In chapter 4, initially, the findings from the previous chapter 3 discussed and summarized. This is followed by Model development: the section where the Poisson and Negative Binomial regression methods are introduced. The methods to determine the final models from the experimentation between Poisson and Negative Binomial regression models are discussed in the Model selection section. The model selection methods combined form an advanced holistic model section system. Finally, the method 'Sorensex Similarity Index (SSI)' [82] used to validate the model results is discussed.

4.1. Introduction

In the previous chapter 3, the data preparation framework was designed and the final OD dataset was obtained. In section 3.4, the exploratory analysis was presented in detail. There are two important points to consider here. The primary division of the dataset is by the days and the regression models have to be selected. There are four days of travel flows available and if time can be added as a variable, then a singular model can be obtained for all days. From the exploratory analysis, it is discovered that all of the days follow similar behavior patterns, and hence, a model for one day can be developed and replicated for the other days. Next, the models that are yet to be built are to be decided. The POI categories are categorical variables and the regression models should have the capability to process them. This is elaborated in the model development section.

In a regression model, the dependent and independent variables exist. The dependent variable of the regression model would be the travel flows between origin and destination. The independent variables would be the origin and destination populations along with the distance between them as well as the POI categories. Multiple experiments need to be conducted to determine the changes in the dependent variable by manipulating the independent variables. An independent variable is a variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable. A dependent variable is a variable being tested and measured in a scientific experiment [105].

4.2. Model Development

In this phase, the methods used to develop the model are discussed. The model results are discussed in the next chapter. From the literature review, it is decided that multivariate regression analysis is selected for model development. However, there are many variations of multivariate regression analysis such as Logistic regression, Binomial regression, Poisson regression, and many more. Depending on the nature of the data, the regression analysis to be performed is selected. Given the data is count data and non-negative, Poisson regression and Negative Binomial regression can be performed. Regarding the gravity model, researchers have performed various tests and some of them argue that Poisson distribution is the better fit while others argue that Negative Binomial is the apt solution [63] [61] [64] [62]. From a statistician's perspective, it can be said depending on the distributing tests and model fitness tests, the appropriate model is can be determined. These models are known as generalized linear models and it is discussed in the following section.

Generalized linear models (GLM)

Generalized linear models were formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models, including linear regression, logistic regression, and Poisson regression [96]. They proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters. Maximum-likelihood estimation remains popular and is the default method on many statistical computing packages. Other approaches, including Bayesian approaches and least-squares, fit variance stabilized responses, have been developed.

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable through a link function (estimates) and by allowing the magnitude of the variance of every variable to be a function of its predicted value.

The model development phase is completely performed in the software R. R is an open-source free tool and is extremely popular for statistical analysis. It offers a wide range of libraries for data structuring to statistical tests. In R, you can perform all kinds of GLM variations and check if the data is appropriate for the model as well as to measure the estimates needed for the study. The glm package in R is built for the same purpose [33].

The dependent variables of the model are the origin and destination populations and the distance between them. These variables are essential to the model as they are the foundation for the traditional gravity model and this research utilizes the gravity model as mentioned in section 2.4. The remaining variables such as POI categories and any additional variables are independent variables for the model. The independent variables will go through a series of experiments to determine which of them are essential and non-essential to the model results.

Poisson regression

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables [66].

A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate. A common reason is the omission of relevant explanatory variables or dependent observations. Under some circumstances, the problem of overdispersion can be solved by using quasi-likelihood estimation or a negative binomial distribution instead of [52].

Negative binomial regression

Negative binomial regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This formulation is popular because it allows the modeling of Poisson heterogeneity using a gamma distribution [95]. The negative binomial regression indicated in Michael L. Zwillings work [122] follows the following equation.

$$ln(\mu) = \beta_0 + (\beta_1 x_1) + (\beta_2 x_2) + (\beta_3 x_3) + ... \tag{4.1}$$

Poisson and Negative binomial regression methods are discussed and have been utilized in researching mobility. Both models are going to be built in a series of experiments. To determine the best models between them, the model results have to be evaluated. To evaluate the model results, a few methods are discussed in the next section and a holistic system of model selection methods are determined.

4.3. Model Selection

Depending on the nature of the data and goodness-of-fit, we can determine the appropriate model for our research. Apart from that, the model results provide the following metrics - AIC, p-value, and Deviance residuals. These metrics determine if the data fits well with the model. The lower the AIC

value, the better the data fits the model. The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby the relative quality of statistical models for a given set of data [50] [49]. Given a compilation of models, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC could provide further insight for model selection.

In statistical significance testing, the p-value is the largest probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is correct [116]. A very small p-value means that such an extreme observed outcome is very unlikely under the null hypothesis. Reporting p-values of statistical tests is common practice in academic publications of many quantitative fields [94]. Though p-value is not the best metric to rely upon, it serves as a good indicator to differentiate the results [74].

Deviance is a goodness-of-fit statistic for a statistical model and it is often used for statistical hypothesis testing. It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood. It plays an important role in exponential dispersion models and generalized linear models [47]. The deviance residuals can be used to check the model fit at each observation for generalized linear models [7]. If the deviance residuals are relatively high, it indicates that the data did not fit the model efficiently and other methods might have to be used.

However, this does not tell us the effectiveness or any properties regarding the model. To understand the model's effectiveness and select the final models, the Root Mean Square Error (RMSE) and R squared (R2) are key measurements chosen for this process. In statistics, the coefficient of determination denoted R2 or r2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). When evaluating the goodness-of-fit of simulated (predicted) vs. measured (estimated) values, it is not appropriate to base this on the R2 of the linear regression. The R2 quantifies the degree of any linear correlation between empirical and predicted data, while for the goodness-of-fit evaluation only one specific linear correlation should be taken into consideration. Thus, for GLM models, being not linear regression, R2 can be used as a pseudo indicator [32] [81] [101].

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results [34].

The following equation would be a good representation of the model to be developed. The assumption is considered in the study that people travel from one location to another due to the attraction of the amenities available at the destination. The key objective of this research is to estimate the travel flow between the origin (O) and destination (D) with the aid of POIs. The gravity model along with the multivariate regression approach was decided in the literature review chapter 2. Given there are multiple variables at play, the multivariate regression is the guiding arrow for developing the model. Improving on the gravity model 2.1 and other equations 2.2, 4.1 presented so far, the following equation 4.2 is developed.

$$f(TravelFlow)_{OD} = Population_O + Population_D + Distance_{OD} + POI_1 + POI_2 + POI_3 \ldots. \quad (4.2)$$

where the dependent variable Travel Flow between the origin (O) and destination (D) is a function consisting of the following independent variables: Origin Population (Population_O), Destination Population (Population_O), the distance between OD (Distance_OD) and the remaining variables are the POI categories determined. The traditional gravity model does not contain the POI categories and by adding the POI categorical variables and modifying it, it is estimated that the effectiveness of the model developed will be superior. The novel idea presented in this research is evolving to fruition now.

The methods to select for model validation are discussed. The p-value would determine the significance of the dependent variable in the model. The AIC value helps us understand the model performance to a certain extent. The RMSE helps us understand the error difference between the estimated and empirical values. R2 aids us in evaluating the goodness-of-fit of the estimated data. The methods together encompass a holistic system for selecting the final models. The data obtained drives the model selection criteria and it is important to note that this is a data-driven approach. Thus we have determined that

both negative binomial and Poisson regression models are to be developed for experimentation and depending on the model selection methods, the final models will be decided.

4.4. Model Validation

In this phase, the methods for the model validation are discussed. The accuracy of the model is determined by validating the estimated values against real data. The Sørensen similarity index (SSI) is a similarity measure that evaluates the amount of closeness between two sample data sets [82]. It can be used to assess the performance of different models using the traditional goodness-of-fit measures for human mobility models.

$$SSI = \frac{2\Sigma_{i,j} min(T_{ij}^m, T_{ij}^d)}{\Sigma_{i,j} T_{ij}^m + \Sigma_{i,j} T_{ij}^d} \tag{4.3}$$

where T_{ij}^d and T_{ij}^m are the actual and predicted trip flows, respectively, from location i to location j. The value of SSI is between zero and one, with zero indicating complete disagreement and one indicating equality.

To analyze if the models are performing more efficiently than the existing gravity models, the traditional gravity models for the different days: MTT (Monday to Thursday), FRI (Friday), SAT (Saturday), and SUN (Sunday) are also developed. Thus in conclusion, for each day selected two models are developed and the estimated flows are compared to see the efficiency and accuracy of the gravity models. From these results, policy recommendations can be developed.

Thus, we have determined the complete methodology framework starting from data preparation to model validation. The methods for data preparation are discussed and the data is explored to determine the final OD dataset. The model development section will contain eight models in total; two for each of the days of the week - MTT (Monday to Thursday), FRI (Friday), SAT (Saturday), and SUN (Sunday). The methods for model development and model validation are also discussed. The next chapters will focus on the discussion of the results from the implementation of these methods.

5

Model Results

In this chapter, multiple models are created and among them, a few effective models are selected for the results. A few sets of experiments are conducted to analyze the POI categorical variables' significance in the estimation of flow between two selected locations for both Poisson and Negative Binomial regression models. Finally, from the various sets of experiments, the final effective models are selected.

5.1. Experimentation

In the Experimentation section, a series of experiments are designed to determine the right model(s) to solve our main research question - "How to estimate human mobility by using points of interest?". The purpose of this experimentation is to figure out the dependency of the variables in a model.

There are six sets of experiments conducted in this process. In all the experiments, both Negative Binomial and Poisson regressions are developed. The experiments contain combinations of the dependent variables. The gravity models, which compromises the origin-destination population and the distance between them, are also included in the experiments. After a series of iterations and experimentation, the models in each set of experiments are chosen. The iterations consisted of removing the POI categorical variables from a model consisting of all variables and adding variables to the gravity model. The main methods used to determine the design of experiments and the models within the experiments are p-value and AIC.

The experiments are conducted in a series. The first set of experiments is the foundation for the next set of experiments. Depending on the results obtained in Set I, the other set of experiments is actualized. The experiment overview is provided in the table below 5.1.

Table 5.1: Experiment overview

Set	Sample size	NB models	Poisson models	Description
Set I	14772	6	7	Complete flow range (Max value - 86640)
Set II	14762	5	6	The flow is limited to 30000
Set III	12579	6	4	The flow is limited to 380
Set IV	14697	7	4	The flow is limited to 10000
Set V	14772	7	4	The models are experimented for different time frames
Set VI	14772	4	4	The difference between origin destination POIs counts is chosen as the variables

First, the experiments are conducted for MTT and then the rest of the days will follow. It is important to note that the rest of the days are crucial to the analysis as the results for the other days will validate the model results for MTT. Apart from this, we can also learn any new insights from the difference between days. All the models contain origin and destination population along with the distance between them. Additionally, all the data points with distance '0' have been excluded from the analysis and they are less than 1% of the sample sizes for all the days; the assumption being these people are not traveling from an origin to a destination with a measured distance. They are the OD travel flows between stations in the same hexagon. A new variable 'Differentiator' is created to improve the effectiveness of the model. The design of the Differentiator variable is outlined in table 5.2 below.

Apart from this, three new groups (variables) are created for the experiments. The groups are a combination of POI categories and are showcased in the table 5.3 below. The categories are created

Table 5.2: Experiment overview

Differentiator level	Flow range
0	Less than 381 (Mean of flow)
1	381 (Mean) - 1000
2	1000 - 10000
3	Greater than 10000

based on the observations in Set I. Commercial and Community contain the least significance in p values when all the variables are present in the model. The most significant variables are group 3 comprising of Educational, Transportation, Financial, Healthcare. The rest of the variables are under group 2. The groups all together comprise all the POI categories and each of them holds significance in the model experiments.

Table 5.3: Experiment overview

Group	Categories
1	Entertainment, Sustenance, Government
2	Commercial, Community
3	Educational, Transportation, Financial, Healthcare

Set I and II experiments are going to be discussed in detail first and then the design process of the rest of the sets will follow. First, the sets are discussed for the day MTT and then the rest of the days is summarised.

Set I experiments

As mentioned above, in this set, the complete dataset is used to build negative binomial and Poisson regression models. It is important to note that all the models contain the origin and destination population along with the distance between them and the Differentiator variable. The key difference between the experiments is the combination of POI categories. First, the negative binomial models are described in the table below.

Table 5.4: Set I - Negative Binomial experiments

No.	Name	Description
1	nb1	All variables
2	nb2	All variables excluding Commercial, Community, Government and Healthcare
3	nb3	POI categories included are Educational, Transportation, Financial and Healthcare
4	nb4	POI categories Commercial, Community, Government, Entertainment and Sustenance are combined to one group and added to model, so it has all variables
5	nb5	Group 1 and 2 are added to individual Group 3 variables
6	nb7	All 3 Groups implying all variables

The negative binomial regression allows for certain conditions to converge and generate the model. The gravity model and any model without the Differentiator variable do not converge and thus, all the models contain the Differentiator variable. The plots for the flow predicted versus empirical are generated at a log scale and showcased in figure 5.1 below.

It is observed that all the models follow similar patterns and there's a clear division between levels of the data. This is caused by the Differentiator variable as it provides a clear distinction in the flow causing the predictions to follow similar level patterns. It is observed that there is a dense section of points on the left corner within 0-10000 and a few points are dispersed for the later ranges. To get a closer look, the plot scale is limited to 20000 on both axes and presented in figure 5.2. We see a similar nature in this scoped down plot as well. There is a dense section of points below 2500 for both axes and the rest is dispersing as the flow increases.

To understand the data properly, the same is represented in the log scale in figure 5.3. By having a log scale, we can observe the nature of data at a wider scale. In the log scaled plot, the nature of data is visually clearly represented with the levels defined by the Differentiator variable. As there four levels in the Differentiator, the output is also produced in four levels.

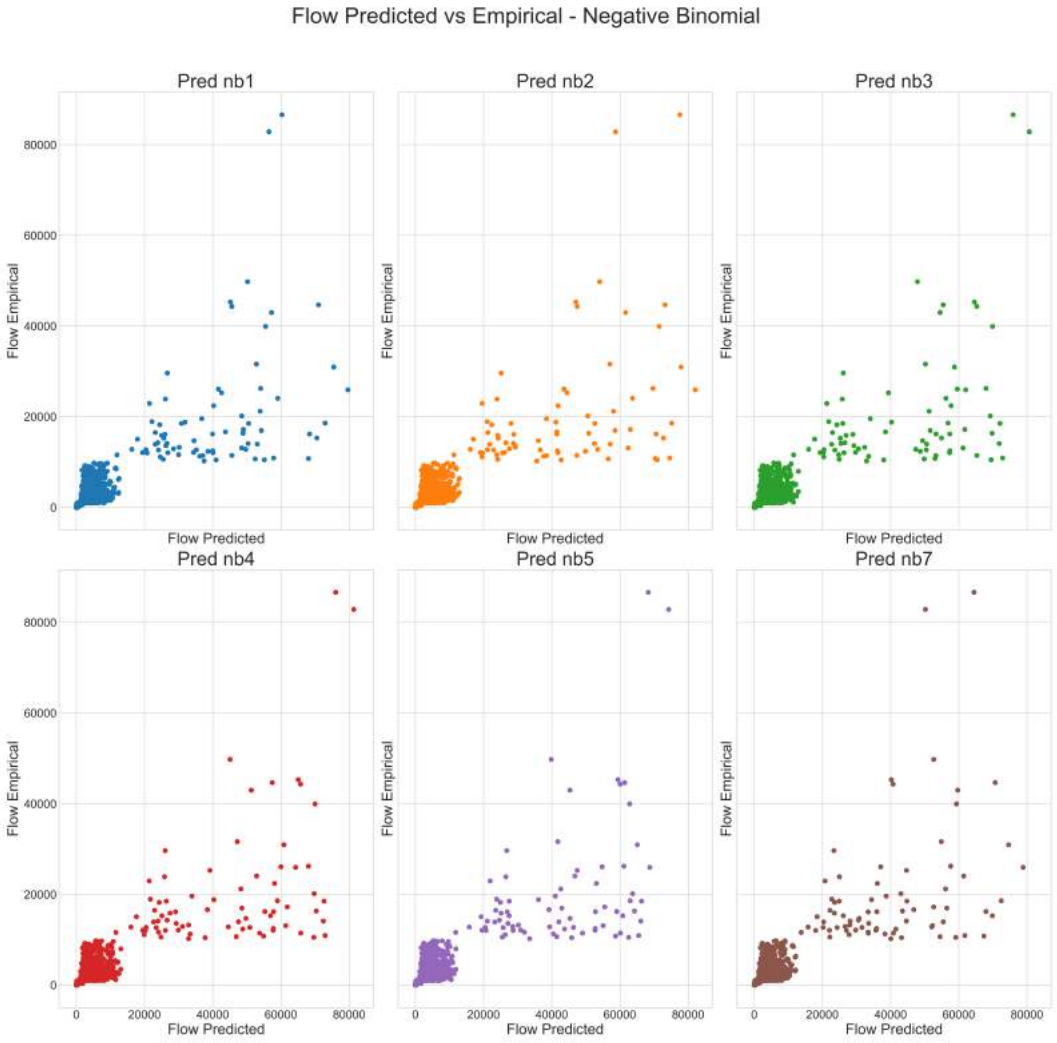

Figure 5.1: Negative Binomial experiments - Flow predicted vs empirical - MTT

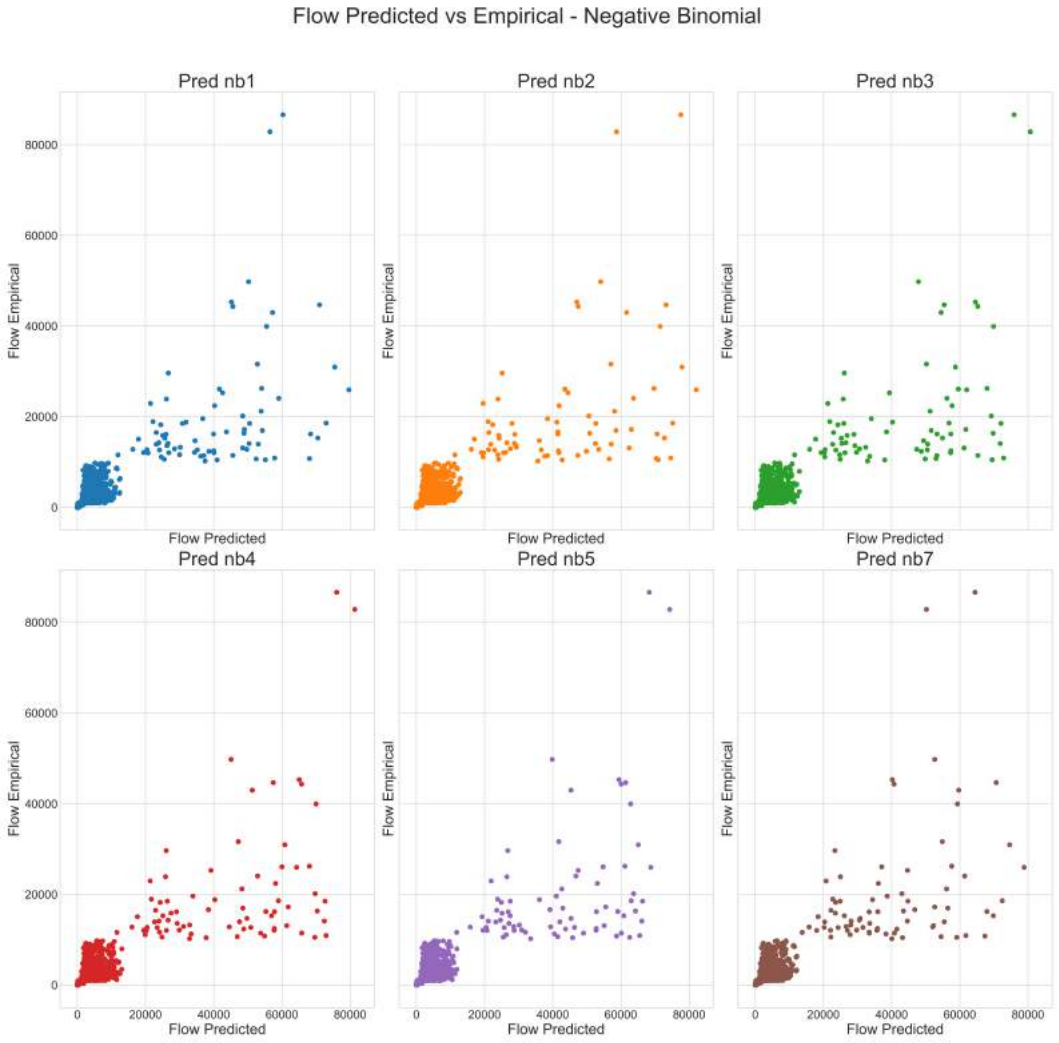

Figure 5.2: Negative Binomial experiments - Flow predicted vs empirical - MTT

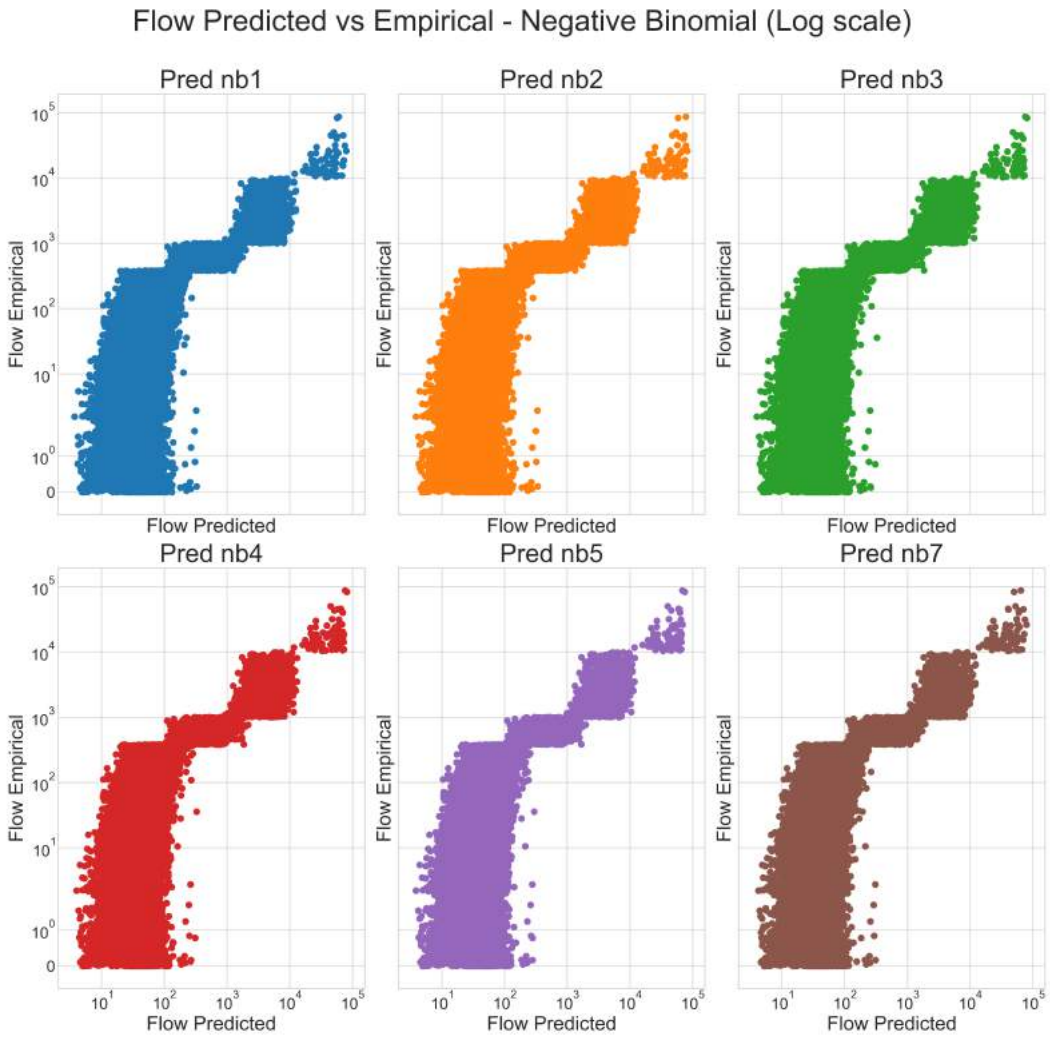

Figure 5.3: Negative Binomial experiments - Flow predicted vs empirical - MTT

Similarly, the Poisson experiments are conducted and the experiments are described in table 5.5 below. The gravity model can be developed with Poisson regression and helps us understand the model performance of traditional methods compared against the new methods.

Table 5.5: Set I - Negative Binomial experiments

No.	Name	Description
1	p1	All variables
2	p2	All variables excluding Commercial, Community, Government and Healthcare
3	pg	Gravity model
4	pg2	Gravity model with Differentiator
5	p3	POI categories included are Educational, Transportation, Financial and Healthcare
6	p5	Group 1 and 2 are added to Educational, Transportation and Financial (Healthcare excluded)
7	p6	All 3 Groups implying all variables

The negative binomial regression allows for certain conditions to converge and generate the model. The gravity model and any model without the Differentiator variable do not converge and thus, all the models contain the Differentiator variable. The plots for the flow predicted versus empirical flows are generated at a log scale and showcased in figure 5.1 below.

The Poisson regression converges better than the Negative Binomial regression models, given the gravity model can be generated for this method. The plots for the flow predicted against the empirical flow are presented in 5.4. The Pg plot, being the gravity model, showcases the dispersion of the data without any Differentiator variable. The remaining models include the Differentiator variable and follow a similar pattern. To get a closer look, the scale for the axes is converted to logarithmic and showcased in 5.5. We can observe that the plot looks similar to the Negative Binomial regression plots developed above.

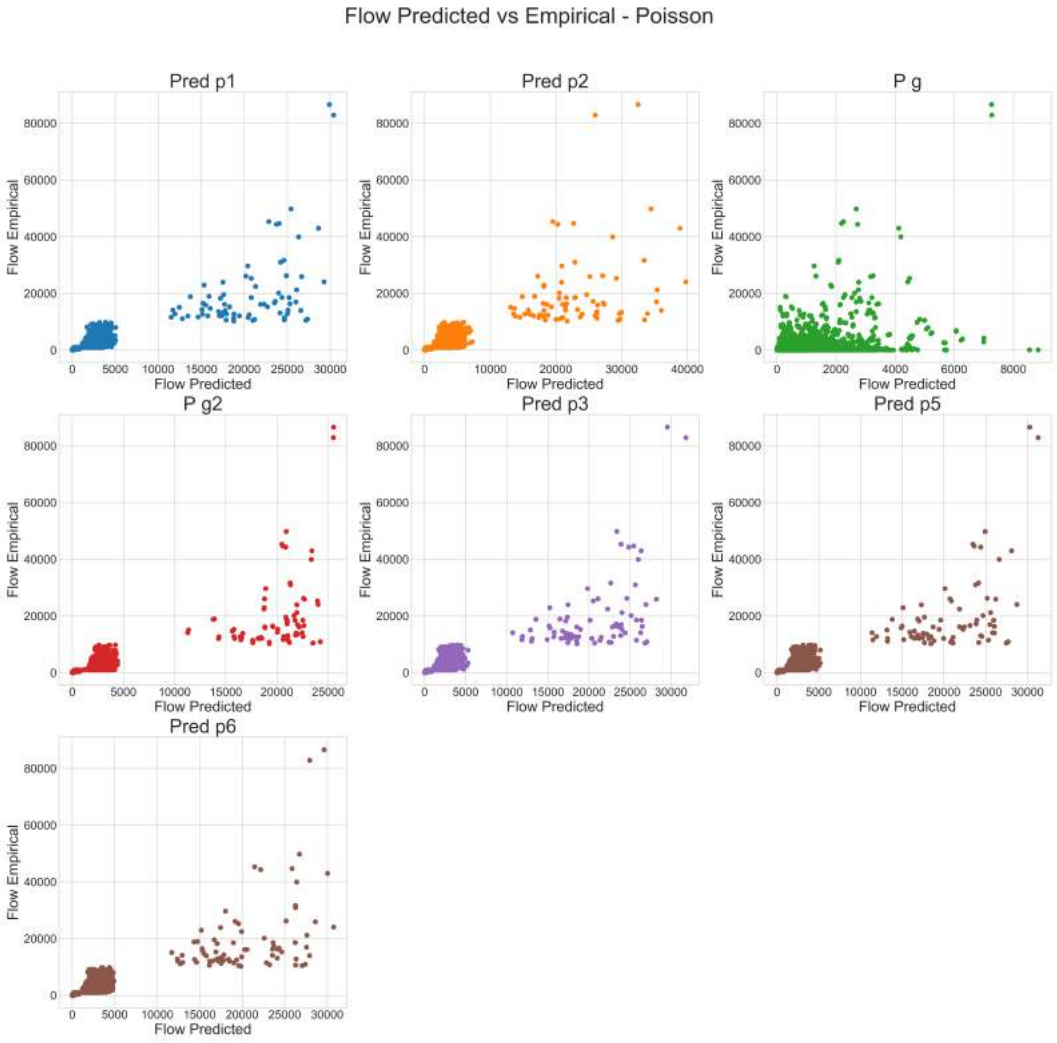

Figure 5.4: Poisson experiments - Flow predicted vs empirical - MTT

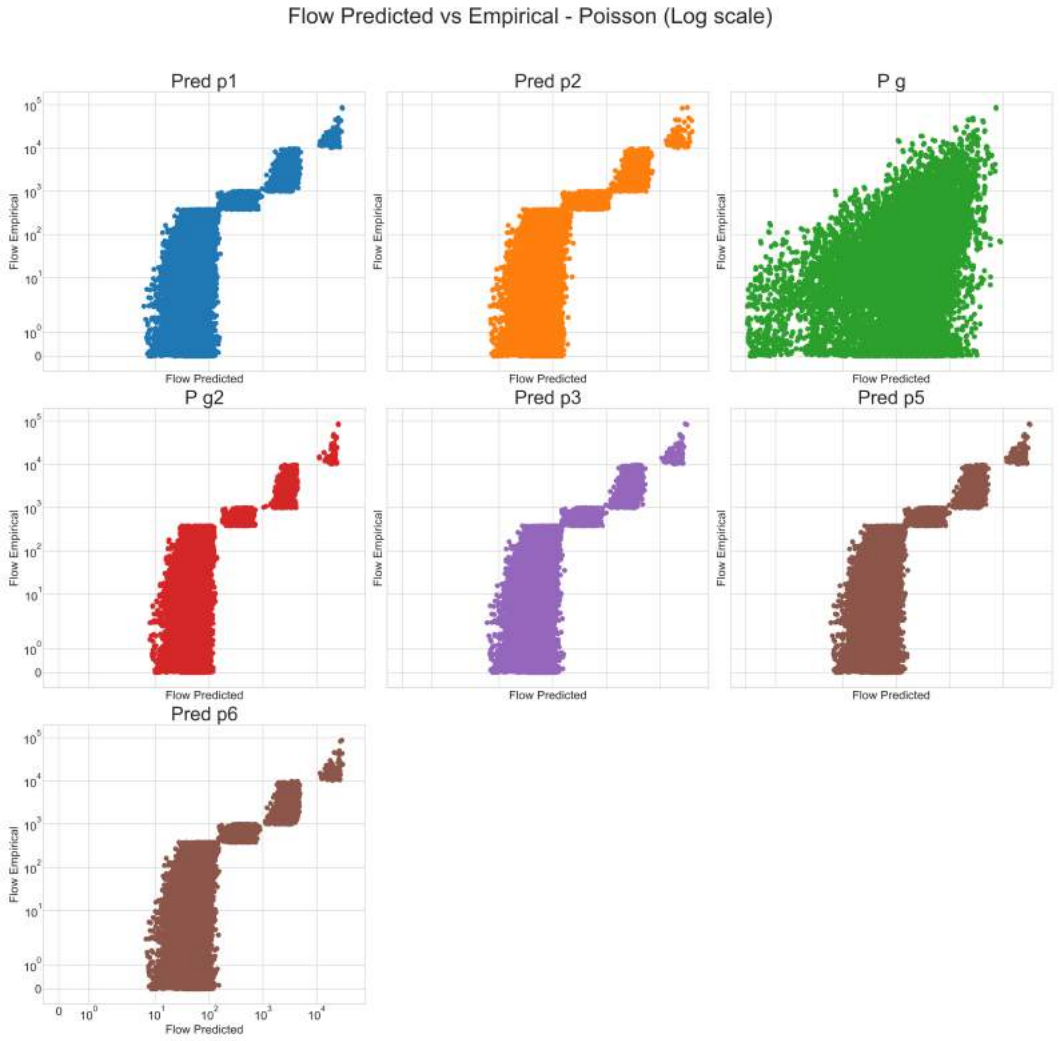

Figure 5.5: Poisson experiments - Flow predicted vs empirical - MTT

To look at the models' results from another perspective, the boxplot of the flow predicted and empirical is presented in figure 5.6. We can observe that the Negative binomial flows are having tail end values and can keep up with the high flow values. The Poisson are denser and have a maximum value of 40000 and show different behavior compared to the Negative binomial estimations. We can also see that the pred_p_g (Gravity model) is on the lower side and not up to mark for the estimations.

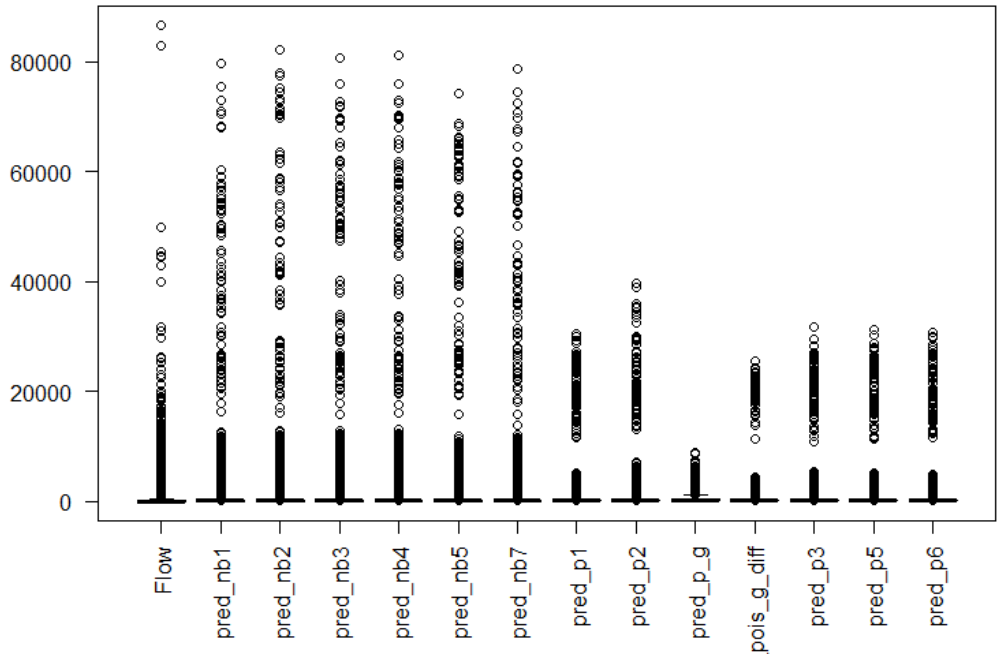

Figure 5.6: Boxplot of all flows of Set I - MTT

Set II experiments

In this set, the flow range is limited to 30000. Thus range is chosen as certain high range flow values might affect the model performance. There are ten high range flow values above 30000 and these are removed from the initial complete dataset. Similar to the above Set I experiments, it is important to note that all the models contain the origin and destination population along with the distance between them and the Differentiator variable. The Negative binomial and Poisson experiments conducted are described in table 5.6 below.

Table 5.6: Set II - Negative Binomial experiments

No.	Name	Regression Model	Description
1	nbx1	Negative binomial	All variables
2	nbx2	Negative binomial	POI categories included are Educational, Transportation, Financial and Healthcare
3	nbx3	Negative binomial	POI categories Commercial, Community, Government, Entertainment and Sustenance are combined to one group and added to model, so it has all variables
4	nbx4	Negative binomial	Group 1 and 2 are added to Educational, Transportation and Financial (Healthcare excluded)
5	nbx5	Negative binomial	All 3 Groups implying all variables
6	px1	Poisson	All variables
7	px2	Poisson	All variables excluding Healthcare
8	px3	Poisson	POI categories included are Educational, Transportation, Financial and Healthcare
9	px4	Poisson	POI categories Commercial, Community, Government, Entertainment and Sustenance are combined to one group and added to model, so it has all variables
10	px5	Poisson	Group 1 and 2 are added to Educational, Transportation and Financial (Healthcare excluded)
11	px6	Poisson	All 3 Groups implying all variables

Similar to Set I, plots for the empirical vs predicted flows are developed for all the experiments and showcased on a logarithmic scale in the figures below. The negative binomial results are available in figure 5.7 and the Poisson results are in figure 5.8. We can see that both of the plots are similar in nature to that developed in Set I and have the patterns differentiated in different levels. Though the large flow values were removed, the nature of the data is still the same.

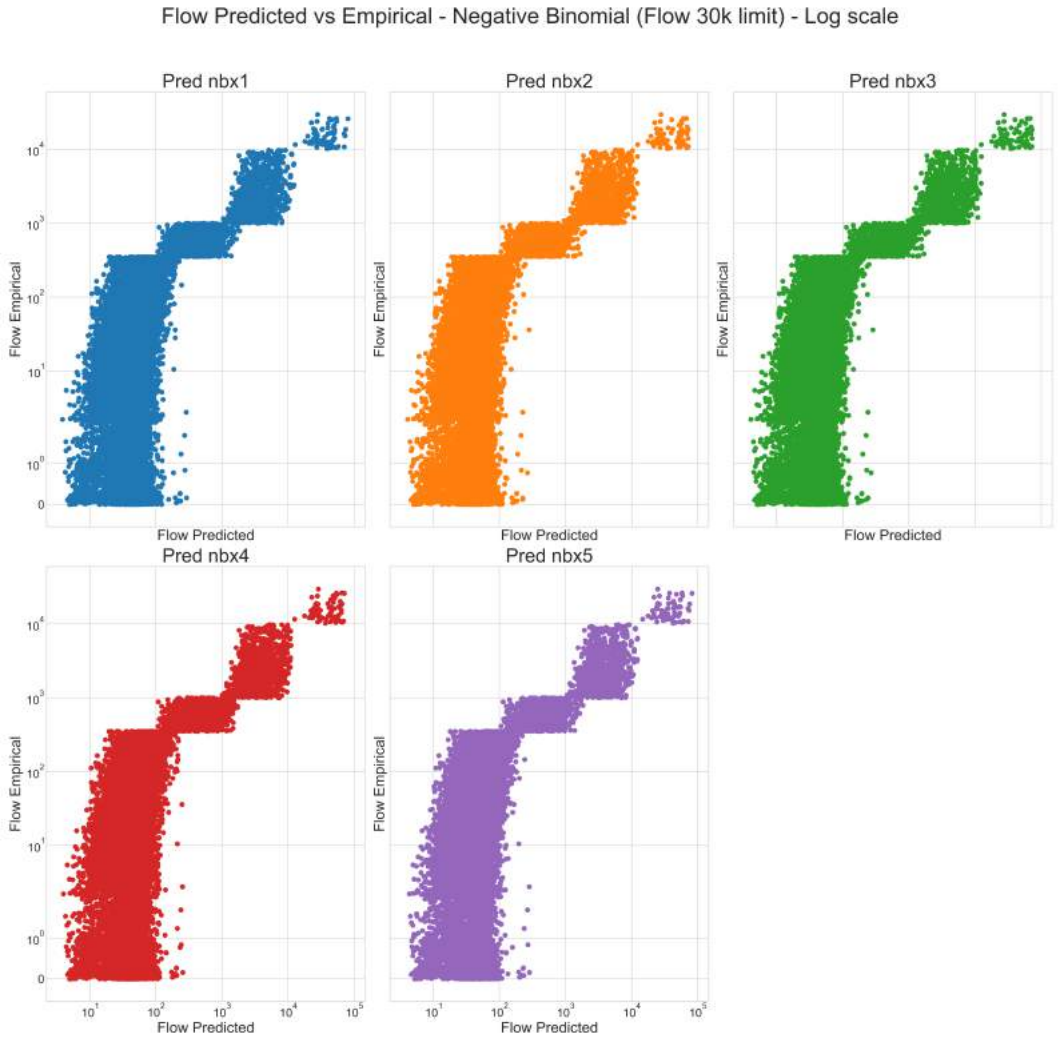

Figure 5.7: Negative Binomial experiments - Flow predicted vs empirical - MTT

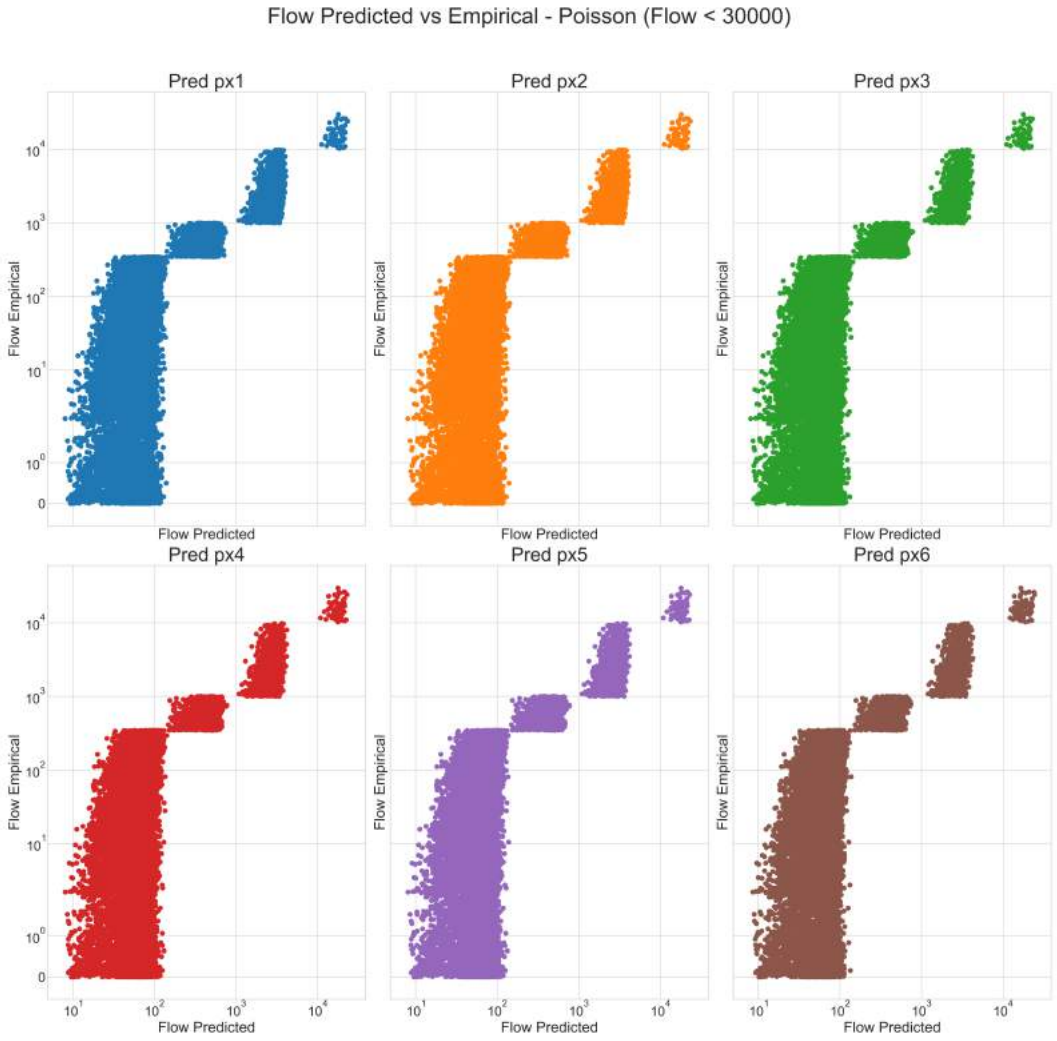

Figure 5.8: Negative Binomial experiments - Flow predicted vs empirical - MTT

To observe the results from another perspective, a boxplot of the flow predicted and empirical is presented in figure 5.9. We can observe that the Negative binomial flows are having tail end values and are exceeding the empirical flow values. The Poisson flows are denser and have a maximum value of less than the empirical data and show better flows comparable than the Negative binomial estimations. From this, we can deduce that Poisson models are performing better than the negative binomial models. However, this provides information regarding the range of the flows only and no further information.

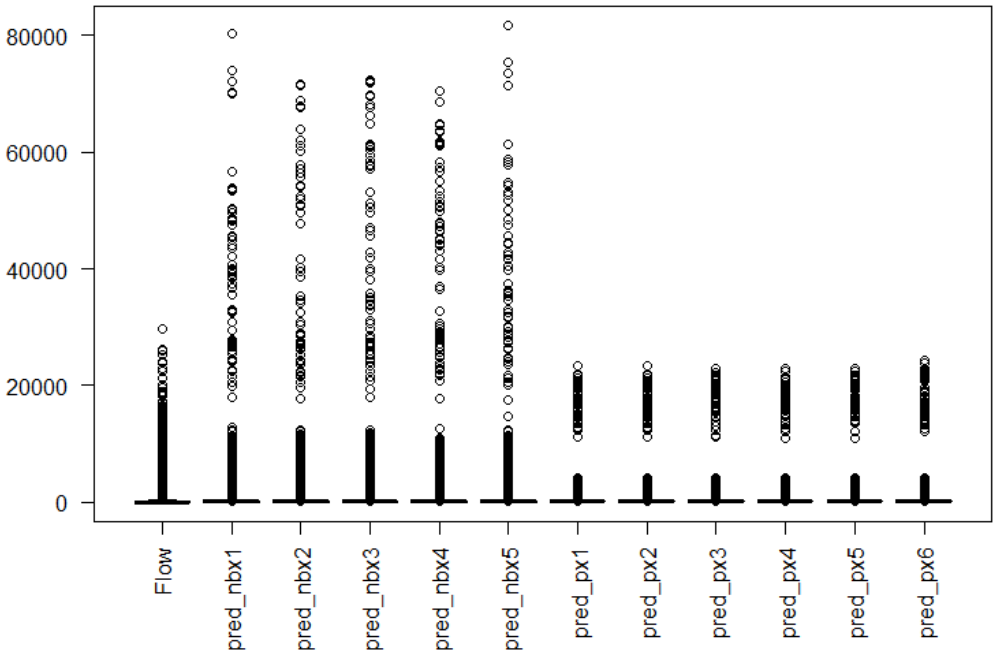

Figure 5.9: Boxplot of all flows of Set II - MTT

Experiments Set Design

The design of the rest of the sets of experiments is discussed in this section. The plots generated in the other sets follow similar characteristics as Sets I and II and thus, not showcased in the main section of the report. The predicted vs empirical plots for the other sets are showcased in the appendix.

Set III comprises of the dataset where the Differentiator variable level is equal to 0. So, it comprises of a smaller dataset where the maximum flow value is the mean value of the complete dataset. Though it contains half of the flow values, it contains 85% of the dataset used in Set I. This indicates that there are very few extremely high flow values, implying high popularity. This experiment is chosen to observe the estimations developed for small flow samples.

Set IV comprises of the dataset where the Differentiator variable level is not equal to 3. This indicates that the maximum flow value of this dataset is 10000 and it is only 75 values less than the complete dataset. The purpose of this experiment is to observe the estimations developed for a sample size between Set III and Set II.

Set V comprises of the models developed for the different time frames. In this case, the time frames for which the negative binomial regression converges are Morning, AM Peak, and Early. These time frames expect AM Peak to contain extremely low flow volumes and are not important time frames to consider. Regarding Poisson regression, the models are developed for all time frames and the time frames with high flow volumes are considered. The time frames with high flow volumes are AM Peak, Inter Peak, PM Peak, and Evening.

Set VI is a special set of experiments. Given people are traveling from origin to destination locations for the POIs available, the assumption considered here is that people are traveling for the amenities available at the destination and not available at the origin. So, the POI categories present in this set are the difference of POIs between origin and destination in the respective POI categories.

Experimentation Results Summary

In this section, the experiments are evaluated across 3 metrics - Root Mean Square Error (RMSE), R-squared (R2), and AIC. These metrics help us evaluate the models' performance and helps us in selecting the high performing models. The Set I and Set II experiments have good models with low RMSE and high R2 values. This indicates these models have the least errors and goodness of fit to the empirical data. The rest of the sets do not exhibit any effective models and are presented in the appendix section for reference. Set I and Set II results are presented in the figure 5.10 below.

	SET I				SET II			
	Experiment	RMSE	R2	AIC	Experiment	RMSE	R2	AIC
0	nb1	2023.781	0.682648	151423.9	nbx1	1939.8960	0.719870	150767.3
1	nb2	2245.439	0.677283	151450.0	nbx2	2208.0980	0.697461	150800.3
2	nb3	2246.352	0.693408	151456.8	nbx3	2215.5200	0.693650	150800.5
3	nb4	2251.621	0.690844	151458.3	nbx4	2119.4440	0.710242	150789.7
4	nb5	2108.923	0.691772	151444.0	nbx5	1993.2650	0.705145	150854.0
5	nb7	2039.789	0.673230	151513.4	px1	587.8919	0.820705	2290828.0
6	p1	1005.901	0.731441	2632252.0	px2	587.8919	0.820705	2290826.0
7	p2	1133.611	0.684695	2634573.0	px3	591.4747	0.819017	2308329.0
8	p_g	1809.205	0.135540	15599280.0	px4	591.6984	0.818984	2305873.0
9	pois_g_diff	1063.523	0.700108	2829046.0	px5	591.7522	0.819088	2305921.0
10	p3	1005.946	0.731408	2645384.0	px6	598.9442	0.814687	2323671.0
11	p5	1004.344	0.732259	2640507.0	NaN	NaN	NaN	NaN
12	p6	1020.273	0.723680	2670267.0	NaN	NaN	NaN	NaN

Figure 5.10: Set I and II experiments summary - MTT

Four models are chosen from these results. The Poisson models have lower RMSE values compared to the Negative binomial counterparts though, regarding the R2 values, the Poisson models are leading. This indicates that the Poisson models fit better and contain low errors in the estimated values. The AIC value is only used for comparison in case RMSE and R2 metrics are not sufficient and should not be compared across these types of regression. However, the case here looks like AIC value is not needed. It is also quite evident that the models are effective compared to the traditional gravity models.

The models chosen are p1, px1, nb1, and nbx1. There are the best models in the class of family models. Summarizing Set I and Set II together, there are four classes of models - Poisson and Negative Binomial models for Sets I and II respectively. In each of the sets, the first models performed better in the order of RMSE, R2, and AIC values. In their class of models, the first models have lower RMSE values, higher R2 values and finally, if these values are the same also, then the AIC value can be used to select a final model.

Initially, the idea was to choose the best models out of the sets. If that is the case, then the best models would be p1 and px1. Model p1 has RMSE value closer to 1000 and R2 value around 0.73, being the best in Set I (if AIC is compared to p5). model px1 has RMSE values as low as 590 while the R2 around 0.82, being the best in Set II (if AIC is compared to px5). The models are chosen depending on the combination of dependent variables. For instance, px2 is not chosen though it has exact RMSE and R2 values as px1 because there is a difference of only one dependent variable (Healthcare) and a minute difference in AIC value. The models for the day MTT are selected. The same is replicated for other days FRI, SAT, and SUN and the Differentiator variable is recreated depending on the mean of the flow for the day. It is observed that the rest of the days follow similar patterns as MTT. The POI categories Commercial, Community, Government, and Healthcare do not have significance in p-values across all days. The difference would be explicit if the counts of POIs available per category changes according to the days. Given the assumption that all the POIs are accessible to every day throughout any day, it is plausible that the models exhibit similar characteristics across all days.

The summary results for each of the days for Set I and II are showcased in the figures 5.11 and 5.12 respectively below. We can observe that for Set I, p1 and nb1 are effective models for all specified days and follow the same characteristics as they did for day MTT. The same is observed for Set II. The models px1 and nbx1 are performing the best for all the specified days.

	FRIDAY				SATURDAY				SUNDAY			
	Experiment	RMSE	R2	AIC	Experiment	RMSE	R2	AIC	Experiment	RMSE	R2	AIC
0	nb1	2025.113	0.685235	151677.6	nb1	1932.3450	0.741721	142416.6	nb1	2470.6530	0.691897	132518.6
1	nb2	2173.542	0.685572	151695.7	nb2	2187.7380	0.725973	142430.3	nb2	2749.1600	0.675845	132533.1
2	nb3	2215.617	0.695927	151707.0	nb3	2201.8220	0.732008	142447.4	nb3	2626.2540	0.703256	132543.4
3	nb4	2232.885	0.690555	151707.9	nb4	2201.6160	0.720941	142446.8	nb4	2635.2040	0.694434	132543.8
4	nb5	2107.784	0.689135	151693.9	nb5	2072.9160	0.723353	142434.0	nb5	2522.3900	0.692154	132530.2
5	nb7	2020.174	0.674077	151780.7	nb7	2055.3650	0.726808	142526.4	nb7	2595.0190	0.674760	132617.8
6	p1	1055.459	0.734764	2758947.0	p1	616.1005	0.793929	1922735.0	p1	421.0402	0.813225	1413262.0
7	p2	1188.212	0.686678	2763553.0	p2	704.1627	0.769541	1932830.0	p2	422.8220	0.809509	1424919.0
8	p_g	1910.279	0.134391	16217150.0	p_g	1258.1210	0.136774	11254610.0	p_g	873.5486	0.140592	7999277.0
9	pois_g_diff	1127.989	0.697550	2991405.0	pois_g_diff	650.3798	0.769365	2054161.0	pois_g_diff	434.5104	0.799887	1490077.0
10	p3	1058.660	0.733102	2780792.0	p3	618.9661	0.792260	1934510.0	p3	427.4344	0.808640	1425719.0
11	p5	1055.460	0.734724	2770060.0	p5	617.2784	0.793575	1930530.0	p5	425.2062	0.811245	1417876.0
12	p6	1071.175	0.726632	2803122.0	p6	627.5134	0.787326	1953202.0	p6	440.2926	0.801326	1443141.0

Figure 5.11: Set I experiments summary - FRI, SAT and SUN

	FRIDAY				SATURDAY				SUNDAY			
	Experiment	RMSE	R2	AIC	Experiment	RMSE	R2	AIC	Experiment	RMSE	R2	AIC
0	nbx1	1959.2230	0.716082	150865.2	nbx1	1711.5570	0.690399	143056.8	nbx1	2163.57200	0.536166	1.316972e+05
1	nbx2	2169.6740	0.701380	150903.0	nbx2	1901.6520	0.665435	143086.0	nbx2	2164.10500	0.546439	1.317271e+05
2	nbx3	2181.5930	0.696505	150903.7	nbx3	1896.9570	0.662122	143086.9	nbx3	2164.59988	0.539530	1.317271e+05
3	nbx4	2069.6110	0.713631	150885.4	nbx4	1813.0390	0.678829	143075.1	nbx4	2147.52600	0.543629	1.317144e+05
4	nbx5	1946.9820	0.709359	150963.9	nbx5	1841.4080	0.664326	143169.3	nbx5	2218.67300	0.523911	1.318032e+05
5	px1	609.9188	0.821434	2350871.0	px1	491.8387	0.793800	1974608.0	px1	404.71550	0.734351	1.370118e+06
6	px2	611.3411	0.820675	2352982.0	px2	492.0112	0.793658	1974759.0	px2	404.74100	0.734317	1.370136e+06
7	px3	617.0093	0.817794	2375496.0	px3	496.3604	0.790847	1988988.0	px3	412.69310	0.725994	1.382718e+06
8	px4	615.0468	0.819047	2367633.0	px4	497.1037	0.790467	1987327.0	px4	409.06850	0.731262	1.374106e+06
9	px5	614.7377	0.819213	2367033.0	px5	496.1847	0.791485	1986451.0	px5	409.59990	0.731323	1.374540e+06
10	px6	623.1818	0.814448	2388022.0	px6	506.0069	0.784519	2010854.0	px6	428.08020	0.714890	1.399865e+06

Figure 5.12: Set II experiments summary - FRI, SAT and SUN

One difference between the days observed is that the weekdays MTT and FRI have a maximum flow range surpassing 80000 while for the weekend days SAT and SUN have a maximum flow of 45576 and 28090 respectively. For Set II, the limit is reduced to 15000 for SUN while kept at 30000 for SAT (similar to MTT and FRI) as to observe the model performance for at a smaller scope while excluding the high range values. In the next section, model validation is discussed for the selected models.

All the first models contain all the variables i.e. all the POI categories including the origin and destination populations along with the distance between them and the Differentiator variable. Though the p-value shows that some categories have less significance, these categories (if included) produces the best results when compared across the other methods selected such as RMSE, R2, and AIC respectively. This indicates that all the variables are essential to the model and more variables or accurate datasets could enhance the model. Compared to the gravity model p_g, the rest of the models certainly performed better in all model selection methods excluding p-value. the p-value can be used as an indicator to check the significance but did not affect in determining the best model. It was utilized prominently in designing the sets of experiments. This affirms the fact "Though p-value is not the best metric to rely upon, it serves as a good indicator to differentiate the results" [74] presented in 4.3.

The methods were crucial in determining and understanding the models among the sets of experiments created. Among the numerous sets of experiments created, they helped me determine the best set of experiments and present in the experimentation section 5.1. Additionally, the methods helped me determine the best models in each of the sets of experiments. They show us where the model fails and where the models worked. R2 should be ideally 1 while RMSE and AIC should ideally be 0, if the model is ideal and estimates perfectly. This discrepancy in the model could be explored in the dataset. The distribution of amenities within POI categories and categorization play a critical role in the model fits. Even the flow values and selection of sample size (handling the outliers) could make a difference in the model results. The estimation of mobility using POIs is the objective of this study and these methods are crucial in evaluating the models and their results.

Thus, we can conclude that the models p1, nb1, px1, and nbx1 are the most effective models from the set of experiments and are selected for model validation. The model validation method Sorensex Similarity Index (SSI) is conducted for all the models in Set I and II to confirm that the model selected is indeed the best ones among the model family classes (Poisson and Negative binomial).

5.2. Model Validation

From the above section, we have identified the most effective models from the above experimentation section. The RMSE, R2, and AIC metrics are effective in determining the model selection. However, the Sorensex Similarity Index (SSI) is an effective method for calculating the similarity between the empirical and estimated data, as mentioned in 4.4. The SSI is calculated for all the models in Set I and II. The models p1, nb1, px1, and nbx1 were determined to be best among their respective family model classes. The gravity model is also included as a reference point at the end to affirm the fact that these new models perform better than the traditional methods. The calculated SSI values are presented in the below table 5.7 below for each of the class of models - Poisson Set I, Negative binomial Set I, Poisson Set II, and Negative binomial Set II.

Table 5.7: Comparison of performances of models based on the SSI value. The best models are highlighted in Green and the Gravity model is highlted in Gray.

Model	MTT	FRI	SAT	SUN
Set I experiments				
p1	0.456851	0.4571279	0.4597455	0.4487915
p2	0.4465971	0.4470796	0.4507769	0.4489118
p3	0.4564507	0.456495	0.4587912	0.4479785
p5	0.4565711	0.4568354	0.4591144	0.4483593
p6	0.4546081	0.4547847	0.4574559	0.4466884
nb1	0.4661721	0.4647947	0.4726917	0.4701298
nb2	0.4657472	0.4644129	0.4722659	0.4698681
nb3	0.4652745	0.464081	0.4723898	0.4695762
nb4	0.4652096	0.4639936	0.4721793	0.4694893
nb5	0.4655747	0.4643112	0.4722691	0.4696272
nb7	0.4643985	0.4628222	0.4712637	0.4686026
Set II experiments				
px1	0.4578144	0.4593119	0.4457552	0.4551868
px2	0.457814	0.4592551	0.4457608	0.4551759
px3	0.4571021	0.4585272	0.4448522	0.4544071
px4	0.4572318	0.458767	0.4450153	0.4547489
px5	0.4573258	0.4589	0.4450761	0.4548248
px6	0.456045	0.4575322	0.4433428	0.4531943
nbx1	0.4710257	0.4722695	0.4615867	0.4771408
nbx2	0.4702918	0.4715461	0.4611233	0.4765373
nbx3	0.4701676	0.471442	0.4609918	0.4764525
nbx4	0.4705494	0.4717109	0.4611796	0.4766391
nbx5	0.4693184	0.4703558	0.4600116	0.4757208
pg	0.3528249	0.3539256	0.3654026	0.3633367

The SSI value ranges from 0 to 1, with 1 indicating that the datasets are similar; the estimations are in perfect order. We can observe that for the Poisson models' SSI values are around 0.46 with p1 and px1 leading the board for all the days. The same is observed for Negative binomial models' - SSI values are around 0.47 with nb1 and nbx1 leading the board for this class. It is interesting to note that, according to the model selection methods such as RMSE and R2, the Poisson models performed better. In the case of SSI, the Negative binomial models are performing slightly better than the Poisson models. However, the difference is minimal. The SSI value deteriorates slightly with each consequent model indicating that removing the POI categorical variables is not the right step to execute.

Comparing across the days, SAT seems to be performing the best for Set I (p1 and nb1) with SUN and MTT performing the poorest for p1 and nb1 respectively. For Set II, FRI and SUN seem to be performing the best for models px1 and nbx1 respectively with SAT being the poorest. The differences are minimal, however. The upper limit of SAT flow for Set I is 45576 and Set II is 26511 with a difference of 8 data points. The model performance must have deteriorated slightly with the removal

of close-range high flow values.

The first models performed better using the model selection and model validation methods. All the first models contain all the variables i.e. all the POI categories including the origin and destination populations along with the distance between them and the Differentiator variable. However, the similarity between the datasets is only around 0.47. That means that there is still a difference of 0.53 to reach optimal estimations. There could be many reasons for this difference. The data quality can play a critical role in analyzing this issue. Starting with the data quality of OSM data and followed by the categorization of POIs and distribution of the POIs within these subjective categories. Alternatively, there is a possibility that the POI categories are not sufficient to estimate mobility to a complete extent and additional variables can be added to the model. This is explored further in the upcoming chapters.

In this chapter, first, the experimentation section was discussed. Multiple sets of experiments containing negative binomial and Poisson regression models were created for the day MTT. The variation of the sets and the difference between them were discussed. The results were reproduced for the rest of the days - FRI, SAT, and SUN. The model selection methods helped in determining the sets of experiments as well as the models within them. For each of the days, certain models were selected for model validation. In model validation, the Sorensex similarity index (SSI) was used to measure the similarity between the empirical and estimated data. The results are discussed in the next chapter.

6

Discussion

In this chapter, the different aspects of this research are discussed. First, the model results are discussed followed by the limitations observed throughout the study. Next, the academic value this model brings to the future of research regarding this study is discussed. Going on, the impact this research has on policy implications is elaborately discussed. Finally, future work is summarized in the end.

6.1. Model Results

In the chapter 5, first the experiments section 5.1 was discussed. The experiments Set I and Set II were discussed in detail while the rest of the set designs were elaborated. The main methods used to determine the design of experiments and the models within the experiments are p-value and AIC. the p-value can be used as an indicator to check the significance but did not affect in determining the best model. It was utilized prominently in designing the sets of experiments. This affirms the fact "Though p-value is not the best metric to rely upon, it serves as a good indicator to differentiate the results" [74] presented in 4.3. The AIC value should be close to zero and if the value of the model seemed to be increasing, then the models were not considered for the experiments. This helped in speed up the process as there were four days in total and given that MTT was completed first, the replication for the rest of the days was seamless throughout the study. There were not many differences observed between the results of the days throughout the study. The difference between weekday and weekend was slightly observed in the SSI values. The SAT day SSI seemed to perform well better than other days. This could be because the POI critical categories with huge amenities such as Sustenance, Commercial are available throughout the weekend especially SAT. The critical reason is that the POIs were not determined according to the timings of the weekday/weekend basis. The study assumed that all POIs were available at all times given the lack of information on POI timings.

The rest of the sets of experiments that did not perform well comparatively are presented in appendix 8. The Negative binomial estimations always appear to have long tail ends. If that issue is solved, the negative binomial regression models would probably be a good selection. However, the Poisson models seem to be performing better in many model selection methods. The model selection methods used are AIC, p-value, RMSE, and R2 as discussed in chapter 4 as well. The RMSE, R2, and AIC in the specified order helped in determining the best models among the lot. They also helped understand that each set (Set I and II) of experiments has two different classes (Poisson and Negative binomial) of family models and thus 4 family models in total. From each of these family models, the best models were determined.

The best models turned out to be the first models for each case. All the The first models contain all the variables i.e. all the POI categories including the origin and destination populations along with the distance between them and the Differentiator variable. Both the Poisson and Negative binomial regression models seemed to be performing well compared to the traditional models. Though according to the model selection methods such as RMSE and R2, the Poisson models performed better. However, in the case of SSI, the Negative binomial models are performing slightly better than the Poisson models. it is important to note that the differences are minimal. For future work, Poisson and Negative binomial regressions along with other new possible techniques could be explored to obtain optimal estimations.

After the model selection, the model validation was conducted using the 'Sorensex Similarity Index' (SSI) method to find the similarity between the estimated and empirical data. From the results, it is evident that the new models perform better than the traditional gravity models. However, the SSI indicator being 0.46 is quite low and there can improvements in many aspects to reach the optimal value. There could be multiple factors at play for this problem. The dataset contains flows from both sides origin acting as destination and vice versa. Usually, in all models, the destination is fixed and there are several origin locations. If a new framework with high server performance was facilitated then the SSI accuracy might improve. All the destinations estimations must be calculated separately for this to occur. Next, data quality can play a critical role in analyzing this issue. Starting with the data quality of OSM data and followed by the categorization of POIs and distribution of the POIs within these subjective categories. If the classification of the POIs is done using statistical methods such as clustering techniques or other classification methods, the results could be improved. In the study "Visualizing the relationship between human mobility and points of interest [119]", the POIs were obtained in clean categories however the coverage was limited and not extended to all the categories contributing to the city. In the study by Camargo [58], the OSM data was used to diagnose the problem and it was concluded that the distribution of POI was not accurate. For example, for a football match, 10000 people could travel by rail, however, for a cafe in that region, the expected audience is only 20. This density distribution should carefully be studied further. This is not explored in this paper as the objective of the paper was to estimate mobility using POI within a city and that has been achieved. Alternatively, there is a possibility that the POI categories are not sufficient to estimate mobility to a complete extent and additional variables can be added to the model. The SSI value deteriorates when the POI categorical variables are removed, affirming that variables could be added to the model. The additional variables can include housing/demographic categorical variables or other metrics which contribute towards mobility. The choice of variables can be inspired by traffic and transport models as well. For determining the right variables to include, mobility as a topic should further be explored.

6.2. Limitations

There are multiple limitations considered in this research for the study to be completed. Starting with the location, London. The city of London was chosen for its data availability primarily. Similar researchers were conducted in different countries (macro-level) and cities (micro-level) in the western world including England, France, Spain, Netherlands, Germany, etc, and cities such as San Francisco, New York, Paris. In the eastern hemisphere, cities in China and other metropolitan areas such as Hong Kong and Singapore have been utilized in research. There is no complete proper data coverage of mid-range cities and third world countries at all for at least research purposes and let alone commercial implications. The technologies to capture and maintain the information is developing rapidly and in the future, we can hopefully have real-time measurements to obtain the best policy implications for the improvement of the livability of human life. London also exhibits multicultural aspects and boasts one of the biggest and oldest transportation (rail network) systems in the world making it attractive for the case study of this problem.

The biggest challenge was putting together a final dataset from various sources and developing the framework. There is not a lot of open free source datasets regarding the points of interest or amenities available in a region. The commercial sources might exist but they must also be working towards achieving complete coverage such as Google or Uber as they haven't achieved it yet. If everything is available at a single source, then the research would be much faster and smoother.

Next, the limitations in OpenStreetMaps is discussed. The coverage of the points of interest (POI) is not complete for the city of London. The limitations were briefly discussed in section 3.4. The POIs available for offices are extremely low, for a city such as London where the working-class population (white or blue-collar) constitutes the major proportion of the population. They are the people who travel during the weekdays in the morning and evening, back and forth from home to work. The city of London contains multiple 30 storied buildings and many more with greater level buildings. These buildings, if commercial spaces contain multiple operating businesses that are also open during the weekend sometimes. The other limitations regarding OSM data are that operating times of the POIs are not available. For instance, if we find out a lot of people are traveling to a place with a lot of nightlife, it makes sense to open a restaurant/bar there as the target audience is readily available. This applies to all categories of POIs and is observed in many cities as well. By having the operating times

available, we can find out that the POIs available per hour given on any day. This information could be extremely useful for estimating the travel data from a deeper, closer, and accurate perspective.

The travel mode chosen is the rail network. The study is limited to one transportation mode as there are time limitations as well as rail transportation is quite complex with multiple modes within the city itself. Also, if the research proves to be effective it can be replicated for the other transportation modes. Another limitation is that the population who use the rail network need to be identified. It could be possible that the largest proportion of people who use the rail network belong to the low economic class (blue-collar people, students, families, etc). If this is identified, then the purpose of the amenities can be identified critically.

6.3. Academic Progression

The academic perspectives were discussed in the Literature Review chapter 2. The studies about mobility as well as mobility with points of interest or amenities are very limited (regarding techniques) or at a buzzing stage. The scope for urban development is immense and with a growing number of cities, urban sciences need to flourish and make an effect on the real world. This paper lays the foundation for several academic perspectives. From the technical literature, it affirms several studies confirm that Poisson regression is a better fit for gravity models (mobility and economic). It affirms several researchers' future work proposals that points of interest categorized together improves the model compared against the traditional models. But the models can be improved considering the Sorensex Similarity Index (SSI) is not above 0.5 where 1 is the maximum. If models can truly achieve 1, then in implementation they'll achieve close to perfection outcomes. the methods were discussed in the above section 6.1 and if explored, could reach the optimal value.

Additionally, the studies about mobility as well as mobility with points of interest or amenities are very limited (regarding techniques) or at a buzzing stage. There are mobility models such as radiation model, IO, PWO, etc which can be fitted with the POI data and measured against the traditional methods respectively. Due to time limitations, only the gravity model was chosen. However, now that it has been proven that gravity models perform better with POI data, the other models can be replicated with similar techniques. The radiation model was developed by Simini $<$ addresses the problems with the gravity model and is a parameter-free model for predicting human mobility. In this model, the total traffic going from A to B depends not only on the population in both locations, but also on the number of opportunities between A and B, and is measured as the total population within a circle of radius r, where r is the distance between locations A and B. The POIs could be a substitute for opportunities and this can be researched as the immediate step.

In addition to POIs, the model can include other variables such as land use, economic variables such as income levels at individual and household level or economic class, and demographic categorical counts based on age, gender, and much more. Additionally, happiness or life satisfaction variables can be of immense value to maintain harmony in the region. The key is to balance all the factors to create a perfect Utopia. Land-use data can be used to add additional information regarding the POI data but the data was not available in the format required. However, OSM data has the column landuse and is utilized in the categorization of POIs. Though the data points containing the landuse column is low, the data might get updated and can be utilized for future work.

There is research regarding people's choice or purpose of their destination and there are a lot of factors considered such as demographic, socio-economic, etc on why people chose to go a destination and the journey's purpose. That is explored in the transport and choice models. This study is to understand how existing amenities attract humans to go to certain destinations and fulfill their activities. This study can be explored to find a bridge between policy and choice models.

Regarding the categorization of POIs, it was done from a perspective of assigning each POI to one category depending on the result or the objective of the amenity. For example, all health-related amenities such as hospitals, pharmacies to swimming, working out, and yoga summing to be Health wellness and improvement. It can be argued that some amenities can belong to multiple categories, but categorization purpose has to discrete and engaging to the general audience as well. As long as that is clear, the POIs can belong to multiple categories. One should not make up the categories as well as try to increase the counts just to obtain the data. The data should have good volume as well as crisp content. After all, it's quality over quantity. If the classification of the POIs is done using statistical methods such as clustering techniques or other classification methods, the results could be drastically

improved as discussed in the above section 6.1.

Regarding model validation, the model performance can be explored by observing the POIs closely. There are high-density data points that might influence the model over a large portion of singular identities. These singular identities might not be balanced correctly and would not reflect in the model. The density is an important factor to consider. A new method needs to be developed or discovered to handle this issue of density between amenities functioning in a city.

The case study chosen can be done in wide ranges. The transportation mode can be extended towards other modes such as bus, car, bike, and maybe even, pedestrian. The city chosen can also differ. Instead of London, another city with the right data availability can be selected and replicated. The results might differ and this can draw some interesting insights.

6.4. Policy Impact

From a policy analyst perspective, urbanism is a sitting diamond mine. The world is rapidly growing and connecting seamlessly just like the ocean currents are one as a whole. The city can be interpreted as a living organism on its own with a heart that is the people who shape the culture in it. By understanding how people are and how they interact with the amenities available in the city, the city can be planned effectively. The locations which attract similar kind of people towards similar kinds of amenities can be planned effectively. This is quite evident in probably all countries subconsciously as well as inspired planning from history; all shopping amenities are in one location (shopping districts), government buildings are mostly close, and this can go on for every category of amenities when categorized effectively. However, there is another approach to this. Spatial planning with providing all kinds of amenities within the citizens' accessibility range is a unique and well-known design. This is prominent in the cities of Europe and Japan. The shape and geography of the city are also crucial to the design process. For instance, a city can be concentric circles or a grown from minor villages and relocated depending on the available land. A lot of factors are into play and history plays a major role in this context.

This model will help policymakers understand the importance of the impact of points of interest and help improve their policies to reduce congestion, improve infrastructure, and improve transit lines. For instance, the transit lines with high traffic volumes could be predicted and new policies such as shared vehicular usage or an increase in the width of roads could be implemented. A good example might be the frequency of transit could be determined depending on the flow between certain locations. The supply and demand for accessibility of transit can be understood forehand and implemented without causing any disruptions to people's lives. This paper can potentially show the importance of POI data and how the attraction of POI is affecting mobility. Urban planning can be improved by understanding the importance of activities' impact and develop policies for a smooth transition towards developing new functions within the city.

According to the Travel in London report generated by the Transport of London 2019 [109], the central aim is to achieve an 80 percent mode share for active, efficient, and sustainable modes by 2041, and follows important three themes - 1) Healthy Streets and healthy people, 2) A good public transport experience and 3) New homes and jobs. This is in alignment with the purpose and usage of this study. By understanding the travel demand from different perspectives, city planning and performance can be improved drastically across all metrics. The recommendations for the city of London cannot be determined now as the study is at the minuscule stage of urban planning. However, the distribution of amenities can be rearranged to control traffic flows. Especially, the center attracts heavy travel flows and this can be redirected by reallocating heavy attraction POIs (a single POI such as a concert might attract 2000 people). The outer regions can establish new POIs to divert traffic flow from surrounding regions. If the models obtain optimal estimations, then the quantified values associated with POI categories could be experimented with and played around with to determine the flow between two locations within a city. This would help greatly in determining the travel flows within a city if the amenities are modified in a region. This could drastically change the field of urban planning.

For policymakers, there are multiple ways this can be interpreted with and argued with. This can be related to the Smart City development program running across the world as well. As mentioned above, by including certain categories which are not just amenities but also in the context of people as well as the history of the city. This study can also be utilized for traffic models which are estimating and predicting traffic flows within a city. This would vastly decrease congestion in cities which in turn would improve the livability factor for people in multiple ways; improve health due to less pollution, decrease

noise/mental disturbance due to multiple vehicles, free space to facilitate new amenities (prominently arts), reduce all types of inequality, to name a few. As mentioned in 1 chapter, the last mile problem can be observed from this perspective. By making sure, people have necessary amenities within their vicinity in sufficient quantities (POIs include bikes or other transport modes) with the bare necessities of ranges of categories of any kind will ensure in solving the last mile problem. This will ensure in promoting a sustainable lifestyle and open mindset to the citizens.

6.5. Future Work

The possibilities for future work are enormous as discussed in all the sections above - In the fields of model results, academic research, policy/decision making as well as psychology and sustainability of the city. The model results section discusses the inefficiency in model results and methods to reach optimal estimations. it also acknowledges that the new model developed is effective than the traditional Gravity model, especially for estimating mobility within a city. In academic research, the academic progression section 6.3 explains the different mechanisms to progress the study further in a technical aspect. The policy impact section 6.4 discovers the possibilities for policymakers to design cities effectively, efficiently, and harmoniously. It also talks about the objective to design a Utopia and this could be one of the ways to look forward to it.

Regarding the usability and reproducible ability of the study, it can be easily replicated. The key variables such as Origin and Destination populations along with the distance between them and the points of interest available in the comprising region should be obtained for any given city and the work can be easily reproduced. The steps to further explore this work are discussed in chapter 8.

7

Conclusion

7.1. Revisiting Research Question

In this section, the research question "How to estimate human mobility by using points of interest?" and the sub-questions are revisited. Human mobility was estimated using the points of interest and the model developed is certainly better than the traditional gravity model. The research conducted for a city and that is a novel idea by itself. From an academic perspective, this research provided proof that POIs can be used to estimate mobility and the models incorporated with POIs are certainly performing better than the traditional models. it also outlines new methodologies and techniques to solve the problem. For London and policymakers, this research provides the foundation for a new thought process in urban science and elaborates in detail the ripple effect this study can provide. The research aimed to 1) Quantify the influence of POI on mobility and 2) To estimate mobility with the aid of POI. this research has satisfied and achieved both these objectives.

the problem was introduced in the Introduction chapter 1. The methods to explore and solve the problem was discussed in the literature review chapter 2 along with the limitations and assumptions. The methods to solve the problem was explained in chapter 3 and a new data preparation framework was laid in place. Also, the limitations and assumptions were discussed and justified accordingly to solve the problem in the best way possible given the time and technical capabilities. The visualizations are placed in a fashionable order to give the reader a clear understanding and the researcher's perspective. The model results 5 shows why the points of interest data help improve in estimating mobility. Quantifying the points of interest is also justified and the categorization of POIs though can be improved lays the foundation for a new thought process. This answers the final research sub-question - "Do modern techniques and models exist for estimating mobility using POI?". The model validation section 5.2 shows the similarity between the empirical data and estimated data. Though the models performed poorly overall, it is still better than the traditional gravity model and can be improved further by incorporating new ideologies presented in the Discussion chapter 6.

the sub-questions which guided this study are revisited and evaluated if the purpose of the study with their help. The first sub-question is "Do modern techniques and models exist for estimating mobility using points of interest?". The models and techniques for estimating mobility with POI were relatively low but there was enough evidence that POIs could improve the model performance in estimating mobility. The second sub-question is "What are the existing models to estimate mobility?". The existing models such as the Gravity model, radiation model, etc were explored and it was determined that the gravity model performed the best among all and is used as the inspiration for the new model development. The third sub-question is "What role do points of interest play in assessing the Origin Destination travel flows?". The points of interest certainly influence mobility. From this research, it was determined that within a city, the points of interest cover-up 46% variance (SSI) of the travel flow estimations. the methods to improve this are also discussed in section 6.1. The final sub-question is "How does the new model compare and contrast against other existing models?" and this is the right question to conclude this study. The new model performs better than the traditional gravity across all model selection (RMSE and R2) and model validation (SSI) methods. However as mentioned, there is room for so much improvement as the scope of this research is immense.

7.2. Reflection

Before coming to the Netherlands and attending the Engineering and Policy Analysis (EPA) Masters program, I was a problem solver at heart and earned for new challenges while looking to make the world a better place with the skill sets I possess. But I did not know how but the drive was always there and the feeling I should try was rooted in me. The thirst for knowledge and curiosity enabled me to explore and understand Policy analysis from various perspectives and always trying to achieve a global understanding. I was never satisfied with the program in many ways, however, I appreciated the vast amount of knowledge presented by the faculty of Technology, Policy and Management (TPM) faculty, and TU Delft. I realized it was not possible to understand everything and so, I had to take what is important and the methods and different ideas that are available in the best way possible. I had no idea about the scope that urban development had to offer until I stumbled upon this project and realized it in the end stages of the thesis. Imagine "You're watching a movie and you don't understand everything in the beginning, but in the climax everything makes sense and your mid explodes". That's the feeling I got at the end of this thesis and I felt like I finally understood the impact of urbanism.

However, there were challenges as the thesis is a project that had to be done with a new novel idea as well as should be within the EPA framework. Urban Science is a growing field and encompasses multiple Sustainable Development Goals (SDG) designed by the United Nations (UN). the core of EPA is the Sustainable Development Goals and urban science is a complex system with multiple holistic systems acting within it. this is a perfect fit for the program. the reporting methods and quality was learned during the program and it would not be possible to write this thesis without the numerous courses where reports had to be submitted. The literature review could not have been possible if I did not understand how to analyze scientific papers. The EPA program helped me understand how to utilize technical methods as tools to solve a problem rather than focus on the tools entirely and miss the big picture/policy aspect. The most critical takeaway from the master's program is the thinking capability. The different thought processes such as Discrete Simulation, System Dynamics, and Agent-based modeling courses have taught me how different systems operate, how to understand complex systems, and how to study and represent data. EPA has taught me that any problem however complex can be solved via multiple methods. Coming from a technical background, EPA has expanded my thinking and helps me understand how to use the right tools for a problem and how to understand the value of a complex problem. Apart from this, EPA has helped me understand the importance of scientific communication and the different mechanism to communicate with an international audience.

This thesis would not have happened with the committee team backing me and I am thankful for them guiding me in the right direction. It was tough to finish the thesis during the pandemic, especially without the proper workspace infrastructure. I hope that this research inspires someone the way it did to me, to find something you care about and work towards it. A city functions in all areas of life and without any component it would be lacking. For a city to prosper, there needs to be peace, harmony, and dedication from the people to lead better lives. The mindset of the people has to be primarily developed and might be possible for it to happen. However, this is idealistic and requires huge complex systems in design.

This inspires me to work towards a major project - "Design of Utopia - A land prosperous due to perfect design". Every city has problems however, the issue can be minimized and a sustainable and happy lifestyle can be achieved. That is the core idea. I leave this study with a quote written by yours truly,

"A city encompasses everything from art to politics, inter cultures to multi disciplinaries, inspired from history to it's geography and has a heart with the citizens bringing life to it that yearns to be explored"

References

[1] API Reference scikit-learn 0.23.2 documentation. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing.

[2] Auto & Mobility Trends In 2019 - CB Insights Research. URL https://www.cbinsights.com/research/report/auto-mobility-trends-2019/.

[3] Battery-powered pod taxi in Gothenburg, Sweden - Insider. URL https://www.insider.com/battery-powered-pod-taxi-bzzt-sweden-gothenburg-stockholm-congestion-problems-stockholm-2016-12.

[4] Census Output Area population estimates London, England (supporting information) - Office for National Statistics. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/censusoutputareaestimatesinthelondonregionofengland.

[5] Congestion Pricing: Examples Around the U.S. - One Page Brief - Electronic Tolling / Congestion Pricing. URL https://ops.fhwa.dot.gov/congestionpricing/resources/examples_us.htm.

[6] datetime Basic date and time types Python 3.8.5 documentation. URL https://docs.python.org/3/library/datetime.html.

[7] Deviance Residuals. URL https://v8doc.sas.com/sashtml/insight/chap39/sect55.htm.

[8] E-Bike Now Biggest Category in the Netherlands - Bike Europe. URL https://www.bike-eu.com/sales-trends/nieuws/2019/03/e-bike-now-biggest-category-in-the-netherlands-10135442.

[9] Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

[10] Europe :: United Kingdom The World Factbook - Central Intelligence Agency. URL https://www.cia.gov/library/publications/the-world-factbook/geos/uk.html.

[11] FHWA Office of Operations - Travel Demand Management. URL https://ops.fhwa.dot.gov/aboutus/one_pagers/demand_mgmt.htm.

[12] Freedom of information | London City Hall. URL https://www.london.gov.uk/about-us/governance-and-spending/sharing-our-information/freedom-information.

[13] GeoPandas 0.8.0+19.g0b804ea GeoPandas 0.8.0+19.g0b804ea documentation, . URL https://geopandas.readthedocs.io/en/latest/.

[14] Geofabrik Download Server, . URL http://download.geofabrik.de/.

[15] Global Power City Index (GPCI) - Institute for Urban Strategies. URL http://www.mori-m-foundation.or.jp/english/ius2/gpci2/.

[16] io Core tools for working with streams Python 3.8.5 documentation. URL https://docs.python.org/3/library/io.html.

[17] Key:amenity - OpenStreetMap Wiki. URL https://wiki.openstreetmap.org/wiki/Key:amenity.

[18] Last Mile RIPPL. URL http://www.rippl.bike/tag/last-mile/.

[19] London stations. URL https://www.doogal.co.uk/london_stations.php.

[20] M| Definition - Esri Support GIS Dictionary. URL https://support.esri.com/en/other-resources/gis-dictionary/search/.

[21] Matplotlib: Python plotting Matplotlib 3.3.0 documentation. URL https://matplotlib.org/.

[22] National Statistics Online. URL https://web.archive.org/web/20090108101256/http://www.statistics.gov.uk/cci/nugget.asp?id=384.

[23] No. 1: London. URL https://www.forbes.com/pictures/edgl45ghmd/no-1-london/#7a4f4c9434fd.

[24] (No Title). URL https://www.r-project.org/conferences/useR-2009/slides/Delignette-Muller+Pouillot+Denis.pdf.

[25] Numpy and Scipy Documentation Numpy and Scipy documentation. URL https://docs.scipy.org/doc/.

[26] Online TDM Encyclopedia - Why Manage Transportation Demand. URL https://www.vtpi.org/tdm/tdm51.htm.

[27] OpenStreetBrowser/Category list - OpenStreetMap Wiki. URL https://wiki.openstreetmap.org/wiki/OpenStreetBrowser/Category_list.

[28] os Miscellaneous operating system interfaces Python 3.8.5 documentation. URL https://docs.python.org/3/library/os.html.

[29] pandas · PyPI. URL https://pypi.org/project/pandas/.

[30] Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with ... - Alison C. Cullen, H. Christopher Frey, Christopher H. Frey - Google Boeken. URL https://books.google.nl/books/about/Probabilistic_Techniques_in_Exposure_Ass.html?id=P9l5q4RjcwcC&redir_esc=y.

[31] Pyrosm pyrosm 0.5.2 documentation. URL https://pyrosm.readthedocs.io/en/latest/.

[32] R-squared measures for generalized linear models | modTools. URL https://modtools.wordpress.com/2014/10/30/rsqglm/.

[33] R: Fitting Generalized Linear Models. URL https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html.

[34] RMSE: Root Mean Square Error - Statistics How To. URL https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.

[35] requests · PyPI. URL https://pypi.org/project/requests/.

[36] Spatial algorithms and data structures (scipy.spatial) SciPy v1.5.2 Reference Guide. URL https://docs.scipy.org/doc/scipy/reference/spatial.html.

[37] Statistical GIS Boundary Files for London - London Datastore, . URL https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

[38] Statistical functions (scipy.stats) SciPy v1.5.2 Reference Guide, . URL https://docs.scipy.org/doc/scipy/reference/stats.html.

[39] Sustainable Last Mile Deliveries with New Babboe e-Cargo Pro Line - Bike Europe. URL https://www.bike-eu.com/sales-trends/nieuws/2020/02/sustainable-last-mile-deliveries-with-new-babboe-e-cargo-pro-line-10137318?_ga=2.84690585.1608128389.1590591517-1467874816.1590591517.

[40] The Subterranean Railway: How the London Underground was Built and How it ... - Christian Wolmar - Google Books, . URL https://books.google.nl/books/about/The_Subterranean_Railway.html?id=wVs9BQAAQBAJ&redir_esc=y.

[41] The Shapely User Manual Shapely 1.8dev documentation, . URL https://shapely.readthedocs.io/en/latest/manual.html.

[42] The Last Mile the term, the problem and the odd solutions, . URL https://medium.com/the-stigo-blog/the-last-mile-the-term-the-problem-and-the-odd-solutions-28b6969d5af8.

[43] Transport for London API. URL https://api-portal.tfl.gov.uk/docs.

[44] Why hexagons?ArcGIS Pro | Documentation. URL https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-whyhexagons.htm.

[45] World's biggest bike parking garage opens in Utrecht but Dutch dream of more | World news | The Guardian. URL https://www.theguardian.com/world/2017/aug/07/worlds-biggest-bike-parking-garage-utrecht-netherlands.

[46] zipfile Work with ZIP archives Python 3.8.5 documentation. URL https://docs.python.org/3/library/zipfile.html.

[47] Correlated Data Analysis: Modeling, Analytics, and Applications. Springer New York, 2007. doi: 10.1007/978-0-387-71393-9.

[48] Omid Reza Abbasi and Ali Asghar Alesheikh. Exploring the potential of location-based social networks data as proxy variables in collective human mobility prediction models. Arabian Journal of Geosciences, 11(8):1–14, 4 2018. ISSN 18667538. doi: 10.1007/s12517-018-3496-4.

[49] Ken Aho, Dewayne Derryberry, and Teri Peterson. Model selection for ecologists: The worldviews of AIC and BIC, 2014. ISSN 00129658.

[50] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. pages 199–213. Springer, New York, NY, 1998. doi: 10.1007/978-1-4612-1694-0{_}15. URL https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

[51] Mohamed Bakillah, Steve Liang, Amin Mobasheri, Jamal Jokar Arsanjani, and Alexander Zipf. Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science, 28(9):1940–1963, 9 2014. ISSN 1365-8816. doi: 10.1080/13658816.2014.909045. URL http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.909045.

[52] Richard Berk and John M. MacDonald. Overdispersion and poisson regression. Journal of Quantitative Criminology, 24(3):269–284, 9 2008. ISSN 07484518. doi: 10.1007/s10940-008-9048-4. URL https://link.springer.com/article/10.1007/s10940-008-9048-4.

[53] Colin P.D. Birch, Sander P. Oom, and Jonathan A. Beecham. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecological Modelling, 206(3-4):347–359, 8 2007. ISSN 03043800. doi: 10.1016/j.ecolmodel.2007.03.041.

[54] Geoff Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems, 65:126–139, 9 2017. ISSN 01989715. doi: 10.1016/j.compenvurbsys.2017.05.004.

[55] Christa Brelsford, Taylor Martin, Joe Hand, and Luís M.A. Bettencourt. Toward cities without slums: Topology and the spatial evolution of neighborhoods. Science Advances, 4(8):eaar4644, 8 2018. ISSN 23752548. doi: 10.1126/sciadv.aar4644.

[56] Dirk Brockmann, Vincent David, and Alejandro Morales Gallardo. Human Mobility and Spatial Disease Dynamics The Open-Access Journal for the Basic Principles of Diffusion Theory, Experiment and Application. Technical report, 2009.

[57] Roberto Camagni, Maria Cristina Gibelli, and Paolo Rigamonti. Urban mobility and urban form: The social and environmental costs of different patterns of urban expansion. Ecological Economics, 40(2):199–216, 2002. ISSN 09218009. doi: 10.1016/S0921-8009(01)00254-3.

[58] Chico Q. Camargo, Jonathan Bright, and Scott A. Hale. Diagnosing the performance of human mobility models at small spatial scales using volunteered geographical information. Royal Society Open Science, 6(11), 11 2019. ISSN 20545703. doi: 10.1098/rsos.191034.

[59] Brian Caulfield, Margaret O'Mahony, William Brazil, and Peter Weldon. Examining usage patterns of a bike-sharing scheme in a medium sized city. Transportation Research Part A: Policy and Practice, 100:152–161, 6 2017. ISSN 09658564. doi: 10.1016/j.tra.2017.04.023.

[60] Keith C. Clarke. Advances in Geographic Information Systems. Computers, Environment and Urban Systems, 10(3-4):175–184, 1 1986. ISSN 01989715. doi: 10.1016/0198-9715(86)90006-2.

[61] Robin Flowerdew. Fitting the Lognormal Gravity Model to Heteroscedastic Data. Geographical Analysis, 14(3):263–267, 1982. ISSN 15384632. doi: 10.1111/j.1538-4632.1982.tb00075.x.

[62] Robin Flowerdew. Modelling migration with poisson regression. In Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications, pages 261–279. IGI Global, 2010. ISBN 9781615207558. doi: 10.4018/978-1-61520-755-8.ch014.

[63] Robin Flowerdew and Murray Aitkin. A METHOD OF FITTING THE GRAVITY MODEL BASED ON THE POISSON DISTRIBUTION. Journal of Regional Science, 22(2):191–202, 5 1982. ISSN 14679787. doi: 10.1111/j.1467-9787.1982.tb00744.x. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9787.1982.tb00744.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9787.1982.tb00744.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1467-9787.1982.tb00744.x.

[64] Robin Flowerdew and Andrew Lovett. Fitting Constrained Poisson Regression Models to Interurban Migration Flows. Geographical Analysis, 20(4):297–307, 9 2010. ISSN 00167363. doi: 10.1111/j.1538-4632.1988.tb00184.x. URL http://doi.wiley.com/10.1111/j.1538-4632.1988.tb00184.x.

[65] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley. A theory of power-law distributions in financial market fluctuations. Nature, 423(6937):267–270, 5 2003. ISSN 00280836. doi: 10.1038/nature01624.

[66] William Gardner, Edward P. Mulvey, and Esther C. Shaw. Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models. Psychological Bulletin, 118(3):392–404, 1995. ISSN 00332909. doi: 10.1037/0033-2909.118.3.392.

[67] Jean François Girres and Guillaume Touya. Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS, 14(4):435–459, 8 2010. ISSN 13611682. doi: 10.1111/j.1467-9671.2010.01203.x. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9671.2010.01203.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2010.01203.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2010.01203.x.

[68] Michael F. Goodchild and Linna Li. Assuring the quality of volunteered geographic information. Spatial Statistics, 1:110–120, 5 2012. ISSN 22116753. doi: 10.1016/j.spasta.2012.03.002.

[69] Luigi Guiso, Paola Sapienza, and Luigi Zingales. Does culture affect economic outcomes?, 2006. ISSN 08953309.

[70] Rongxing Guo. Determinants of spatial (dis)integration. In China's Spatial (Dis)integration, pages 67–105. Elsevier, 1 2015. doi: 10.1016/b978-0-08-100387-9.00004-x.

[71] Mordechai Haklay. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design, 37(4):682–703, 8 2010. ISSN 0265-8135. doi: 10.1068/b35097. URL http://journals.sagepub.com/doi/10.1068/b35097.

[72] Mordechai Haklay and Patrick Weber. OpenStreet map: User-generated street maps. IEEE Pervasive Computing, 7(4):12–18, 10 2008. ISSN 15361268. doi: 10.1109/MPRV.2008.80.

[73] Robin Hickman and Giacomo Vecia. Discourses, travel behaviour and the 'Last Mile' in London. Built Environment, 42(4):539–553, 2016. ISSN 02637960. doi: 10.2148/benv.42.4.539.

[74] Raymond Hubbard and R. Murray Lindsay. Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. Theory & Psychology, 18(1):69–88, 2 2008. ISSN 0959-3543. doi: 10.1177/0959354307086923. URL http://journals.sagepub.com/doi/10.1177/0959354307086923.

[75] Thomas Jekel, Adrijana Car, Josef Strobl, Gerald Griesebner, and GI_Forum 2012 Salzburg. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. Geovizualisation, society and learning conference proceedings, 2012. ISBN 9783879075218.

[76] Shan Jiang, Joseph Ferreira, and Marta C. Gonzalez. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. IEEE Transactions on Big Data, 3(2):208–219, 11 2016. ISSN 2332-7790. doi: 10.1109/tbdata.2016.2631141.

[77] Nathan Keyfitz. Individual Mobility in a Stationary Population. Population Studies, 27(2):335, 7 1973. ISSN 00324728. doi: 10.2307/2173401.

[78] Ned Kock and Leebrian Gaskins. Simpson's paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration. Technical report.

[79] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban Gravity: a Model for Intercity Telecommunication Flows. Journal of Statistical Mechanics: Theory and Experiment, 5 2009. ISSN 17425468.

[80] Minjin Lee and Petter Holme. Relating land use and human intra-city mobility. PLoS ONE, 10 (10), 10 2015. ISSN 19326203. doi: 10.1371/journal.pone.0140152.

[81] David R. Legates and Gregory J. McCabe. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1):233–241, 1 1999. ISSN 00431397. doi: 10.1029/1998WR900018. URL https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/1998WR900018https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998WR900018https://agupubs.onlinelibrary.wiley.com/doi/10.1029/1998WR900018.

[82] Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, and Guillaume Deffuant. A Universal Model of Commuting Networks. PLoS ONE, 7(10):e45985, 10 2012. ISSN 19326203. doi: 10.1371/journal.pone.0045985.

[83] Maxime Lenormand, Aleix Bassolas, and José J. Ramasco. Systematic comparison of trip distribution laws and models. Journal of Transport Geography, 51:158–169, 2 2016. ISSN 09666923. doi: 10.1016/j.jtrangeo.2015.12.008.

[84] Er-Jian Liu and Xiao-Yong Yan. A universal opportunity model for human mobility. 1 2020. URL http://arxiv.org/abs/2001.03701.

[85] Kang Liu, Peiyuan Qiu, Song Gao, Feng Lu, Jincheng Jiang, and Ling Yin. Investigating urban metro stations as cognitive places in cities using points of interest. Cities, 97:102561, 2 2020. ISSN 02642751. doi: 10.1016/j.cities.2019.102561.

[86] Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. Cities, 99:102610, 4 2020. ISSN 02642751. doi: 10.1016/j.cities.2020.102610.

[87] Tianjun Lu, Jennifer Lansing, Wenwen Zhang, Matthew J. Bechle, and Steve Hankey. Land Use Regression models for 60 volatile organic compounds: Comparing Google Point of Interest (POI) and city permit data. Science of the Total Environment, 677:131–141, 8 2019. ISSN 18791026. doi: 10.1016/j.scitotenv.2019.04.285.

[88] Vida Maliene, Vytautas Grigonis, Vytautas Palevičius, and Sam Griffiths. Geographic information system: Old principles with new capabilities, 3 2011. ISSN 13575317. URL www.palgrave-journals.com/udi/.

[89] Marko Matulin. Influencing Travel Demand by the Means of Urban Mobility Management. Technical report. URL https://www.researchgate.net/publication/305303425.

[90] Grant McKenzie and Benjamin Adams. Juxtaposing thematic regions derived from spatial and platial user-generated content. In Leibniz International Proceedings in Informatics, LIPIcs, volume 86. Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 9 2017. ISBN 9783959770439. doi: 10.4230/LIPIcs.COSIT.2017.20.

[91] Graham McNeill, Jonathan Bright, and Scott A. Hale. Estimating local commuting patterns from geolocated Twitter data. EPJ Data Science, 6(1), 12 2017. ISSN 21931127. doi: 10.1140/epjds/s13688-017-0120-x.

[92] Paul Demaio Metrobike. Bike-sharing: History, Impacts, Models of Provision, and Future. Technical report.

[93] Amin Mobasheri, Yeran Sun, Lukas Loos, and Ahmed Ali. Are Crowdsourced Datasets Suitable for Specialized Routing Services? Case Study of OpenStreetMap for Routing of People with Limited Mobility. Sustainability, 9(6):997, 6 2017. ISSN 2071-1050. doi: 10.3390/su9060997. URL http://www.mdpi.com/2071-1050/9/6/997.

[94] Marcus R. Munafò, Brian A. Nosek, Dorothy V.M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. A manifesto for reproducible science, 1 2017. ISSN 23973374. URL www.nature.com/nathumbehav.

[95] NCSS and LLC. NCSS Statistical Software Negative Binomial Regression. Technical report.

[96] J A Nelder and R W M Wedderburn. Generalized Linear Models. Technical Report 3, 1972.

[97] Duccio Piovani, Elsa Arcaute, Gabriela Uchoa, Alan Wilson, and Michael Batty. Measuring accessibility using gravity and radiation models. Royal Society Open Science, 5(9), 9 2018. ISSN 20545703. doi: 10.1098/rsos.171668.

[98] Cheng Qian, Philipp Kats, Sergey Malinchik, Mark Hoffman, Brian Kettler, Constantine Kontokosta, and Stanislav Sobolevsky. Geo-tagged social media data as a proxy for urban mobility. In Advances in Intelligent Systems and Computing, volume 610, pages 29–40. Springer Verlag, 2018. ISBN 9783319607467. doi: 10.1007/978-3-319-60747-4{_}4.

[99] James E. Rauch and Vitor Trindade. Ethnic Chinese networks in international trade, 2 2002. ISSN 00346535. URL https://www.mitpressjournals.org/doix/abs/10.1162/003465302317331955.

[100] Hannah Ritchie and Max Roser. Urbanization, 2020. URL https://ourworldindata.org/urbanization.

[101] Axel Ritter and Rafael Muñoz-Carpena. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. Journal of Hydrology, 480:33–45, 2 2013. ISSN 00221694. doi: 10.1016/j.jhydrol.2012.12.004.

[102] Hernán D. Rozenfeld, Diego Rybski, José S. Andrade, Michael Batty, H. Eugene Stanley, and Hernán A. Makse. Laws of population growth. Proceedings of the National Academy of Sciences of the United States of America, 105(48):18702–18707, 12 2008. ISSN 00278424. doi: 10.1073/pnas.0807435105.

[103] Filippo Simini, Marta C. González, Amos Maritan, and Albert László Barabási. A universal model for mobility and migration patterns. Nature, 484(7392):96–100, 4 2012. ISSN 00280836. doi: 10.1038/nature10856.

[104] Sumeeta Srinivasan, Joseph Fepdfra, Operations Research, Thesis Supervisor, and Frank Levy. Linking Land Use and Transportation: Measuring the Impact of Neighborhood-scale Spatial Patterns on Travel Behavior. Technical report, 1994.

[105] James Stewart, Australia Brazil, Mexico, and Singapore. Calculus Early Transcendentals. Technical report, 2016. URL www.cengage.com/highered.

[106] Samuel A. Stouffer. Intervening Opportunities: A Theory Relating Mobility and Distance. American Sociological Review, 5(6):845, 12 1940. ISSN 00031224. doi: 10.2307/2084520.

[107] O Z Tamin and L G Willumsen. Transport demand model estimation from traffic counts. Technical report, 1989.

[108] Nebiyou Tilahun, Piyushimita Vonu Thakuriah, Moyin Li, and Yaye Keita. Transit use and the work commute: Analyzing the role of last mile issues. Journal of Transport Geography, 54: 359–368, 6 2016. ISSN 09666923. doi: 10.1016/j.jtrangeo.2016.06.021.

[109] Transport of London. Travel in London. Technical report, 2019.

[110] United States. Federal Highway Administration. and Institute of Transportation Engineers. Intelligent transportation primer. Institute of Transportation Engineers, 2000. ISBN 0935403450.

[111] Wouter van Heeswijk, Rune Larsen, and Allan Larsen. An urban consolidation center in the city of Copenhagen: A simulation study. International Journal of Sustainable Transportation, 13(9): 675–691, 10 2019. ISSN 15568334. doi: 10.1080/15568318.2018.1503380.

[112] Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento. Urban mobility study using taxi traces. In TDMA'11 - Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis, pages 23–30, New York, New York, USA, 2011. ACM Press. ISBN 9781450309332. doi: 10.1145/2030080.2030086. URL http://dl.acm.org/citation.cfm?doid=2030080.2030086.

[113] Jinzhong Wang, Xiangjie Kong, Feng Xia, and Lijun Sun. Urban Human Mobility: Data-Driven Modeling and Prediction. Technical report. URL http://www.abs.gov.au/ausstats/abs@.nsf/detailspage/200.

[114] Pu Wang, Marta C. González, César A. Hidalgo, and Albert László Barabasi. Understanding the Spreading Patterns of Mobile Phone Viruses. Science, 324(5930):1071–1076, 5 2009. ISSN 00368075. doi: 10.1126/science.1167053.

[115] Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. González. Understanding Road Usage Patterns in Urban Areas. Scientific Reports, 2, 12 2012. doi: 10.1038/srep01001. URL http://arxiv.org/abs/1212.5327http://dx.doi.org/10.1038/srep01001.

[116] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's Statement on p-Values: Context, Process, and Purpose, 4 2016. ISSN 15372731. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108.

[117] Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. Quantifying the impact of human mobility on malaria. Science, 338(6104):267–270, 10 2012. ISSN 10959203. doi: 10.1126/science.1223467.

[118] Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. International Journal of Geographical Information Science, 31(4):825–848, 4 2017. ISSN 1365-8816. doi: 10.1080/13658816.2016.1244608. URL https://www.tandfonline.com/doi/full/10.1080/13658816.2016.1244608.

[119] Wei Zeng, Chi Wing Fu, Stefan Müller Arisona, Simon Schubiger, Remo Burkhard, and Kwan Liu Ma. Visualizing the Relationship Between Human Mobility and Points of Interest. IEEE Transactions on Intelligent Transportation Systems, 18(8):2271–2284, 8 2017. ISSN 15249050. doi: 10.1109/TITS.2016.2639320.

[120] Liming Zhang and Dieter Pfoser. Using OpenStreetMap point-of-interest data to model urban changeA feasibility study. PLOS ONE, 14(2):e0212606, 2 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0212606. URL https://dx.plos.org/10.1371/journal.pone.0212606.

[121] George Kingsley Zipf. The P 1 P 2 D Hypothesis: On the Intercity Movement of Persons. American Sociological Review, 11(6):677, 12 1946. ISSN 00031224. doi: 10.2307/2087063.

[122] Michael Zwilling. Negative Binomial Regression. The Mathematica Journal, 15, 2013. ISSN 1097-1610. doi: 10.3888/tmj.15-6.

List of Figures

List of Tables

8

Appendix

8.1. Experiment Sets Predicted vs Empirical plots

Set III

The Predicted vs Empirical plots for Set III for MTT are presented below. The figure 8.1 (Negative binomial) shows that the ratio of empirical to estimated flows are dominated towards the left and this is a poor case of regression. For a closer look, the figure 8.2 is represented on a log scale. The same is conducted for Poisson regression models and they exhibit similar characteristics. This is represented in figures 8.3 and 8.4.

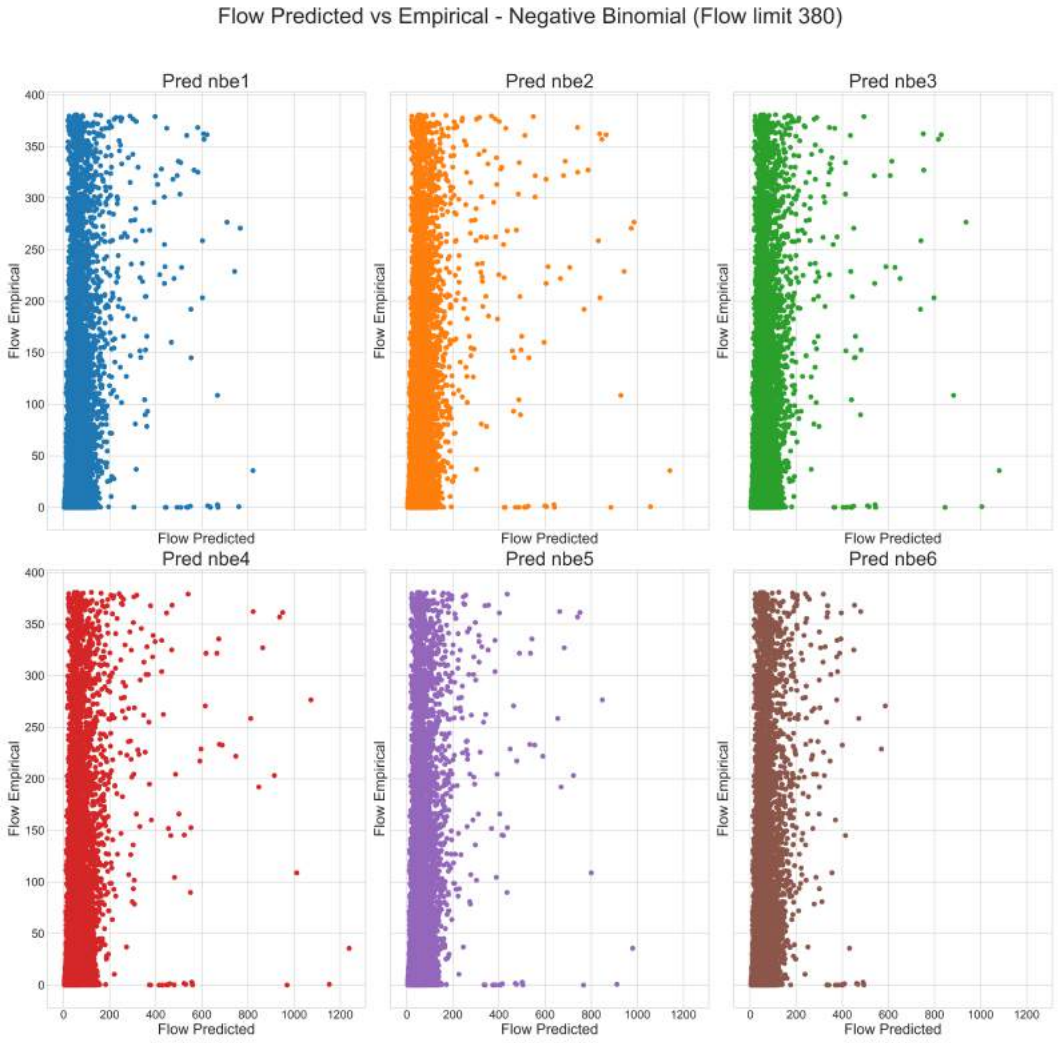

Figure 8.1: Negative Binomial experiments - Flow (380) predicted vs empirical - MTT

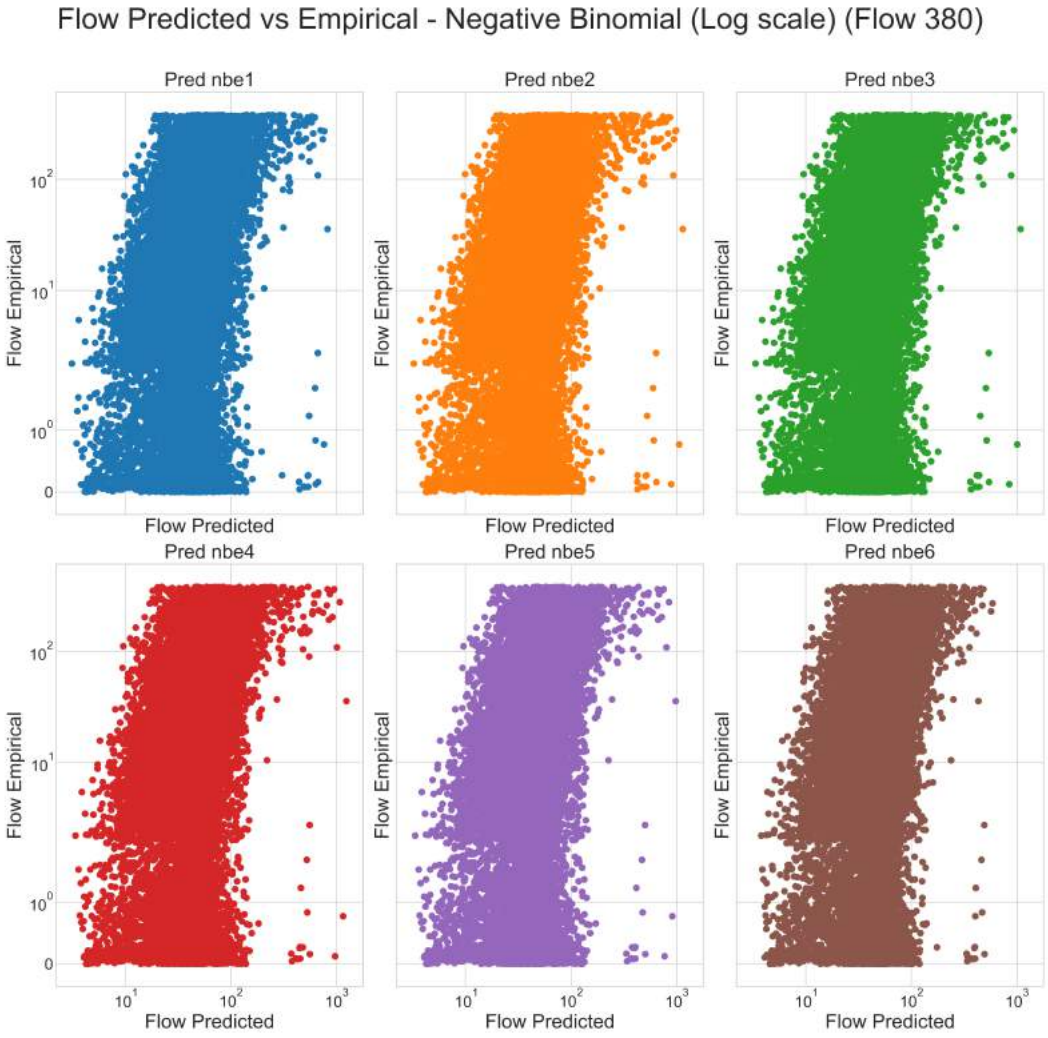

Figure 8.2: Negative Binomial experiments - Flow (380) predicted vs empirical [Log scale] - MTT

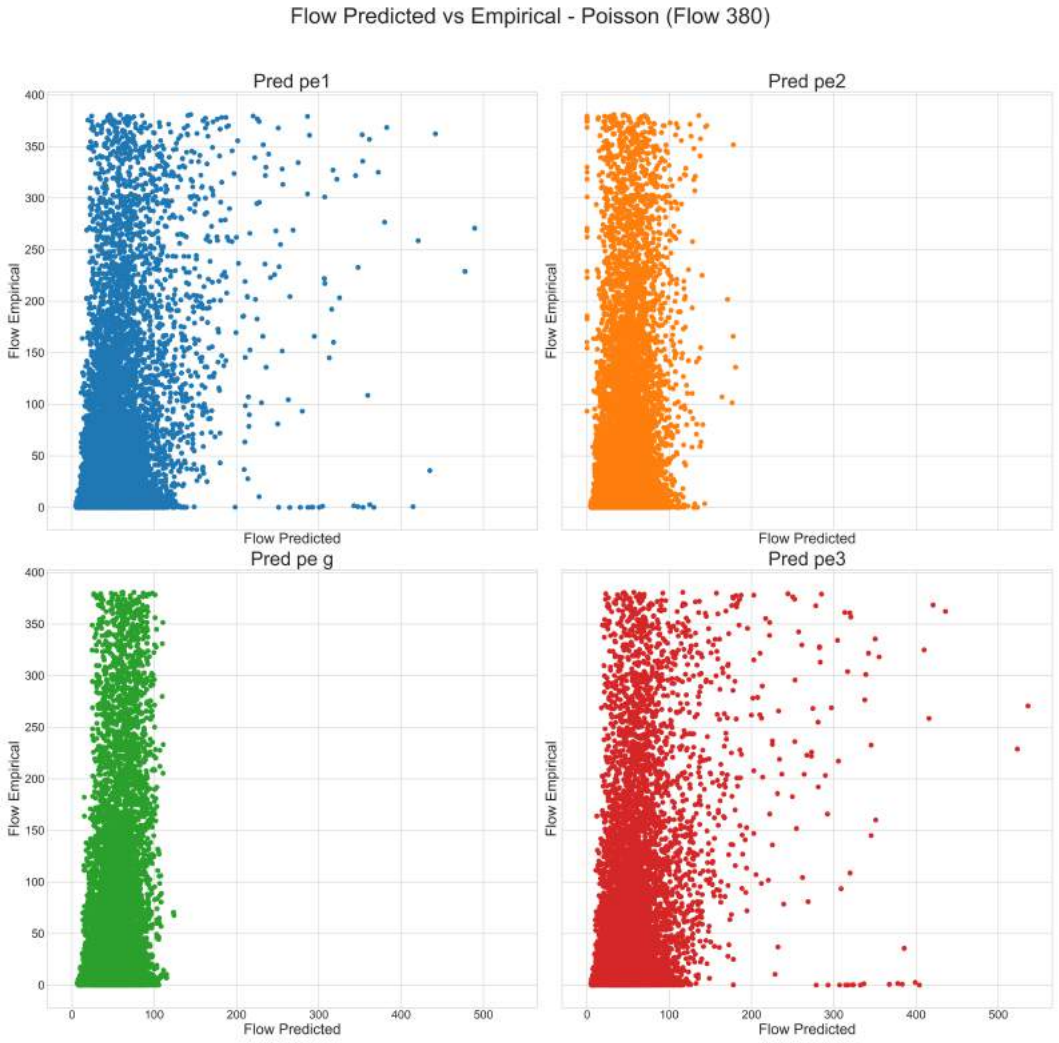

Figure 8.3: Poisson experiments - Flow (380) predicted vs empirical - MTT

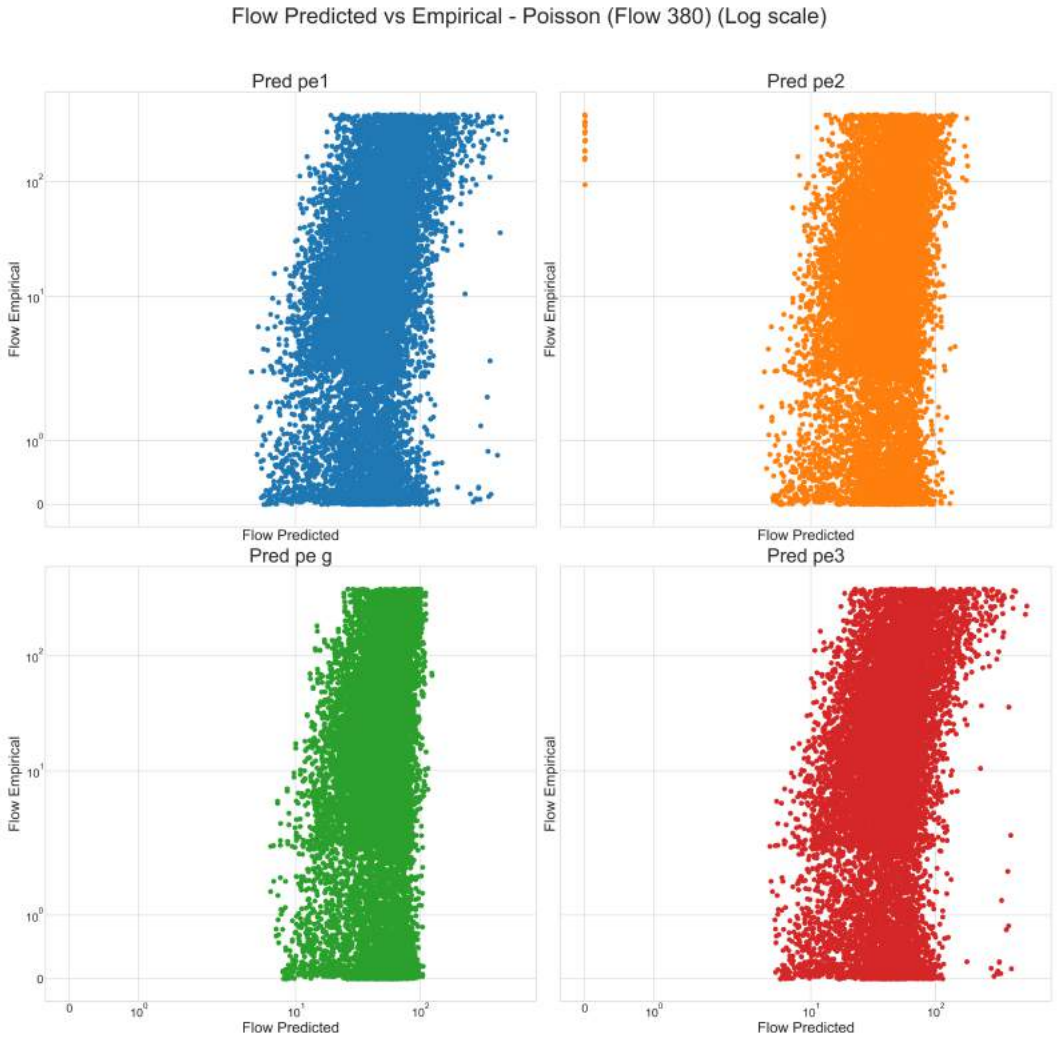

Figure 8.4: Poisson experiments - Flow (380) [Log scale] predicted vs empirical - MTT

Set IV

The Predicted vs Empirical plots for Set IV for MTT are presented below. The figure 8.5 (Negative binomial) shows that the ratio of empirical to estimated flows are dominated towards the left and this is a poor case of regression. For a closer look, the figure 8.6 is represented on a log scale. The same is conducted for Poisson regression models and they exhibit similar characteristics. This is represented in figures 8.7 and 8.8.

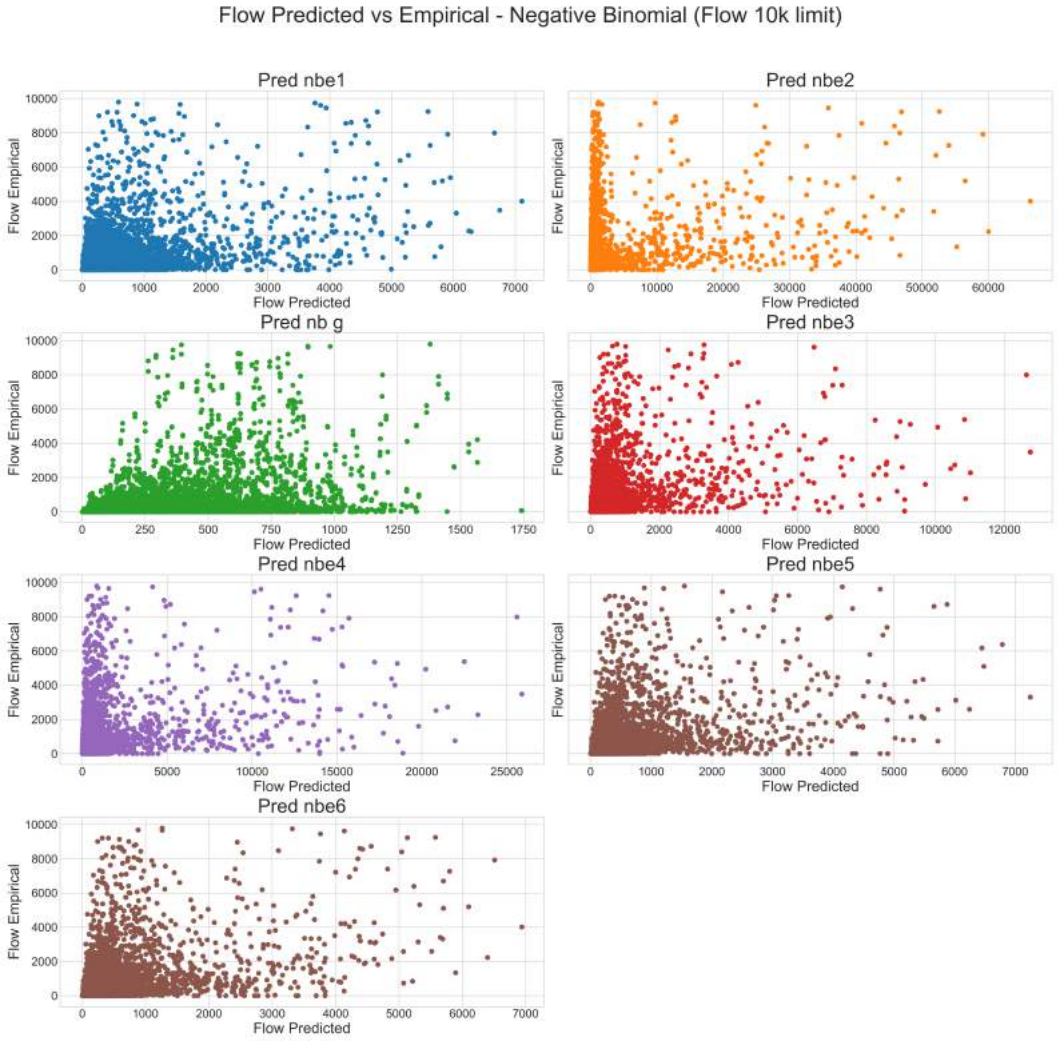

Figure 8.5: Negative Binomial experiments - Flow (10k limit) predicted vs empirical - MTT

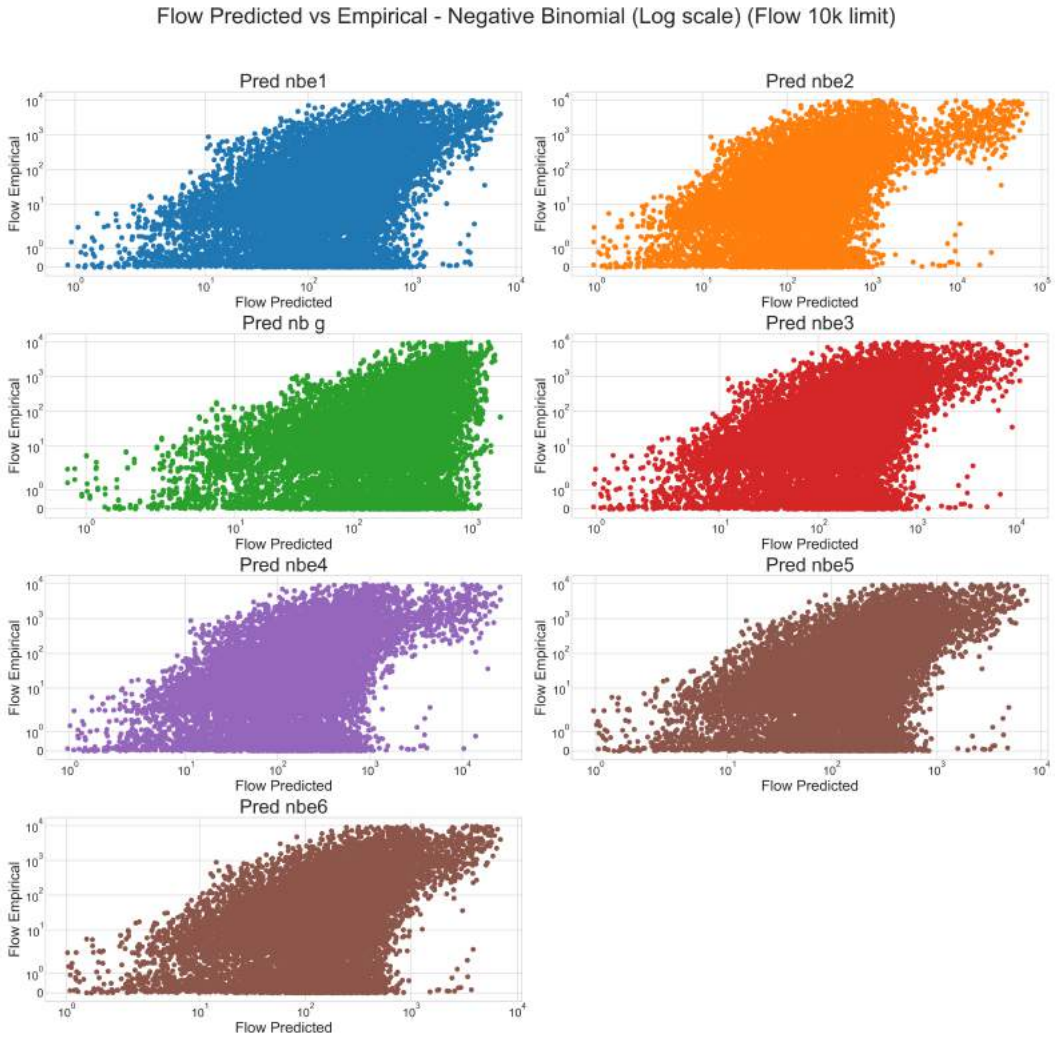

Figure 8.6: Negative Binomial experiments - Flow (10k limit) [Log scale] predicted vs empirical - MTT

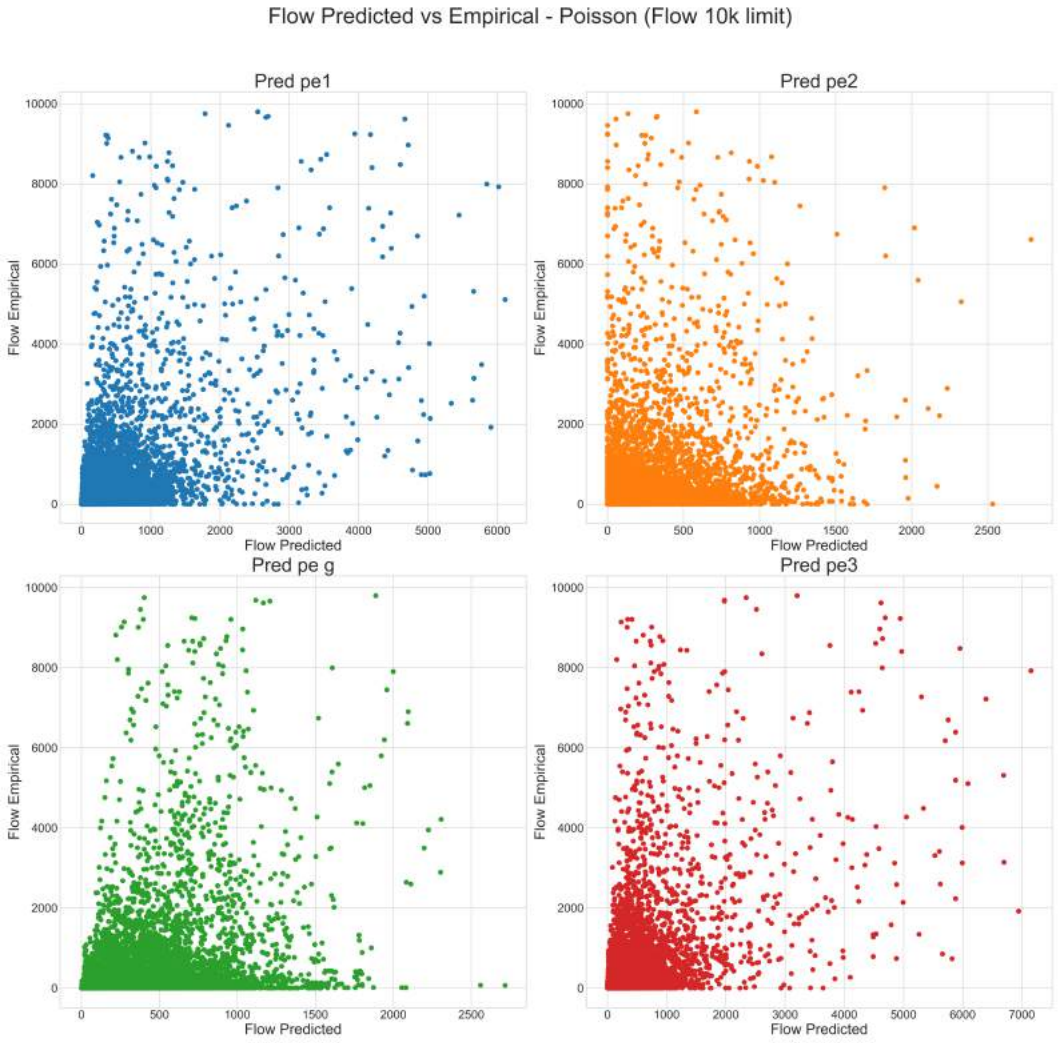

Figure 8.7: Poisson experiments - Flow (10k limit) predicted vs empirical - MTT

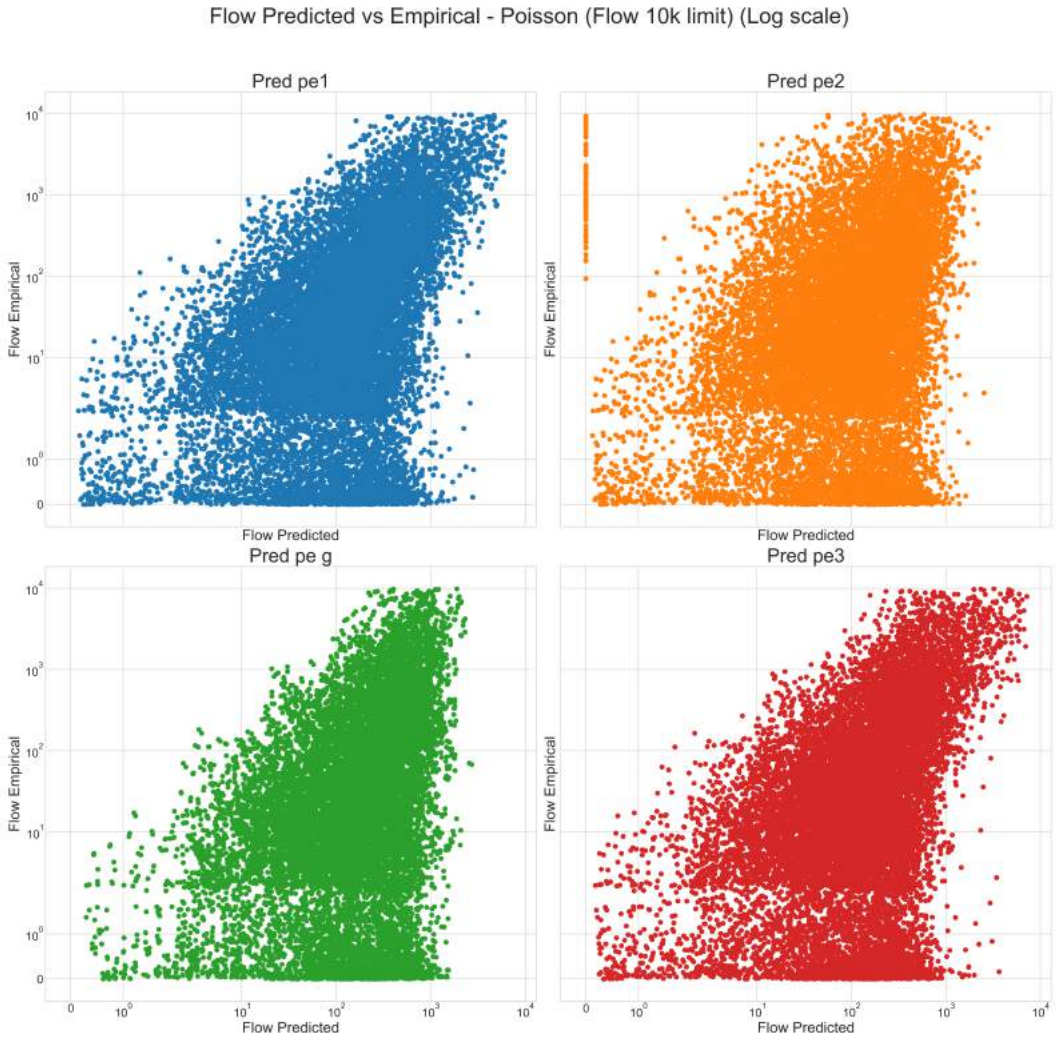

Figure 8.8: Poisson experiments - Flow (10k limit) [Log scale] predicted vs empirical - MTT

Set V

The Predicted vs Empirical plots for Set V for MTT are presented below. The figure 8.9 shows that the ratio of empirical to estimated flows are closely dense to the left corner and this does not give us much information. For a closer look, the figure 8.10 is represented on a log scale. In this set, only Poisson models are available for the following time frames with high flow volumes are AM Peak (t2), Inter Peak (t3), PM Peak (t4), and Evening (t5).

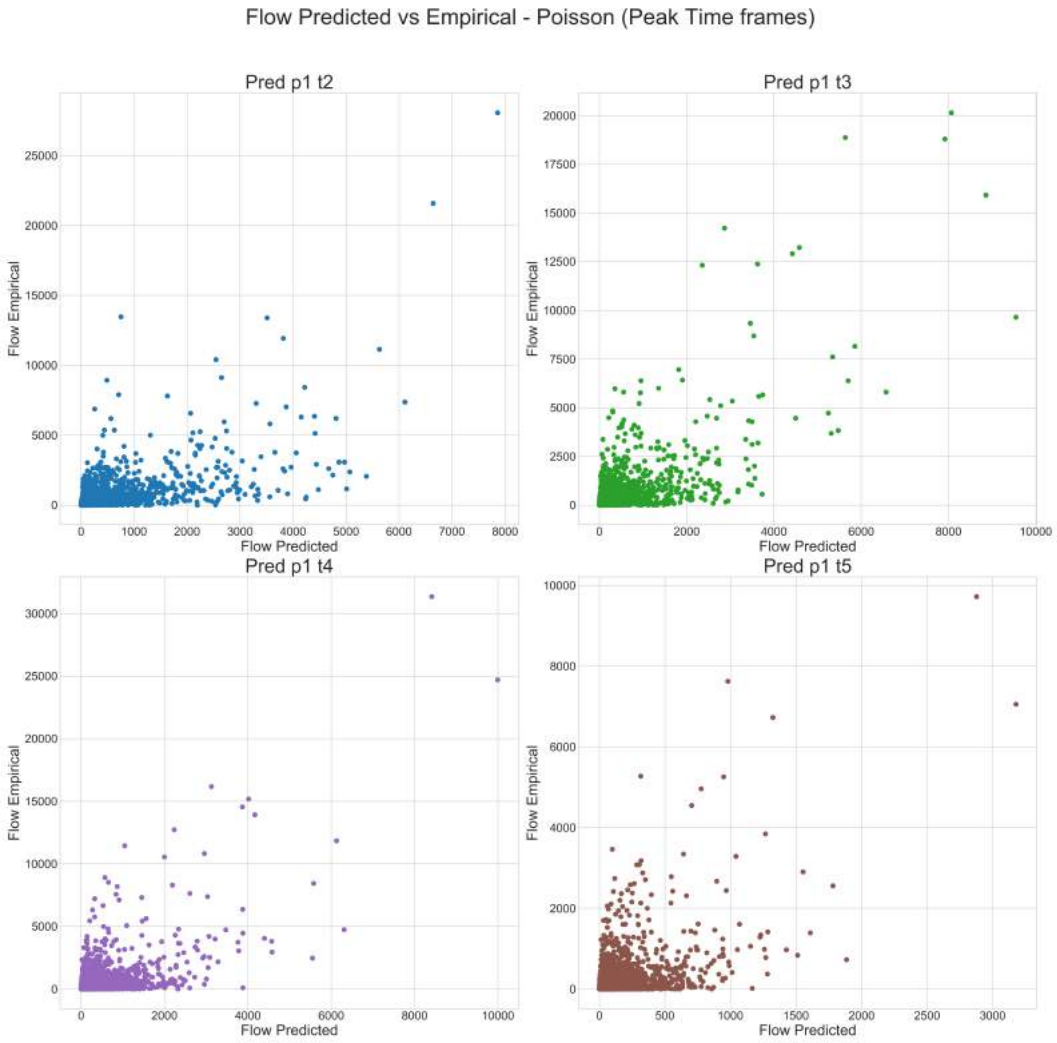

Figure 8.9: Poisson experiments - Time frames Flow predicted vs empirical - MTT

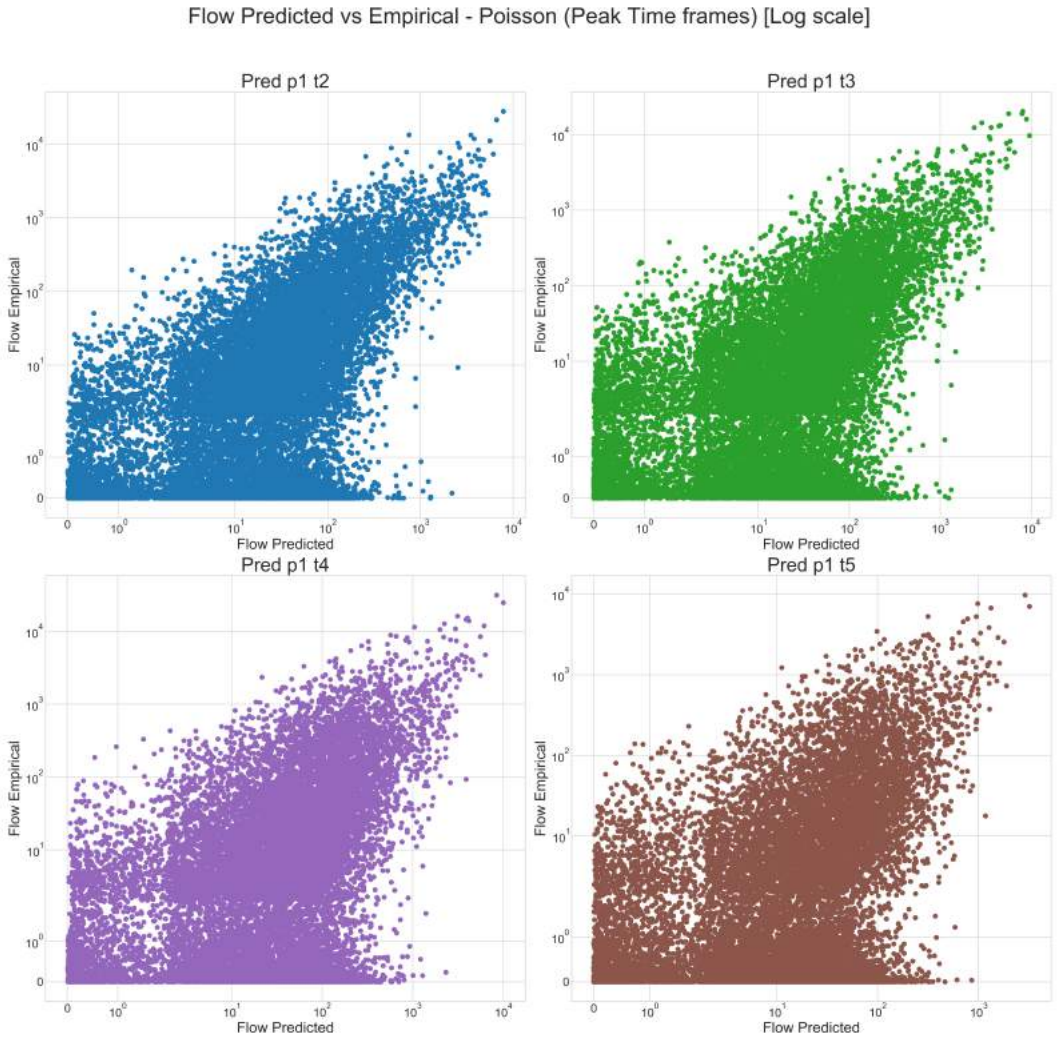

Figure 8.10: Poisson experiments - Time frames Flow predicted vs empirical [Log scale] - MTT

Set VI

The Predicted vs Empirical plots for Set VI for MTT are presented below. The figure 8.11 shows that the ratio of empirical to estimated flows on a log scale and do not represent ideal regression characteristics. They exhibit strange patterns, far from expected behaviours. In this set, two Negative binomial and two Poisson models are available.

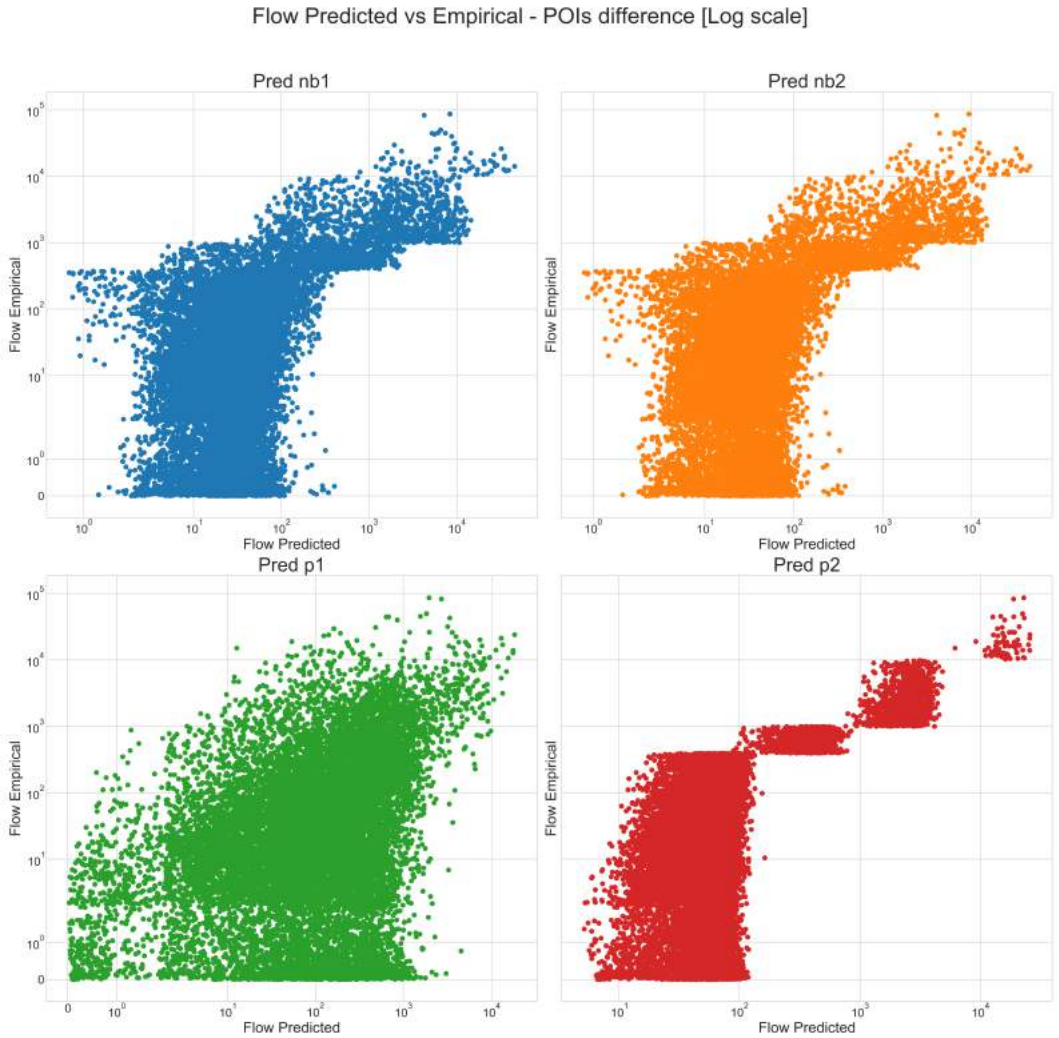

Figure 8.11: POIs difference experiments - Flow predicted vs empirical [Log scale] - MTT

8.2. OD Exploratory Analysis - FRI, SAT and SUN

Distribution of flow for time frames of the day across distance quantiles

The distribution of flow for time frames of the day across distance quantiles are represented in the figures 8.12, 8.13 and 8.14 for the days FRI, SAT and SUN respectively. They exhibit similar patterns between themselves and similar to the day MTT.

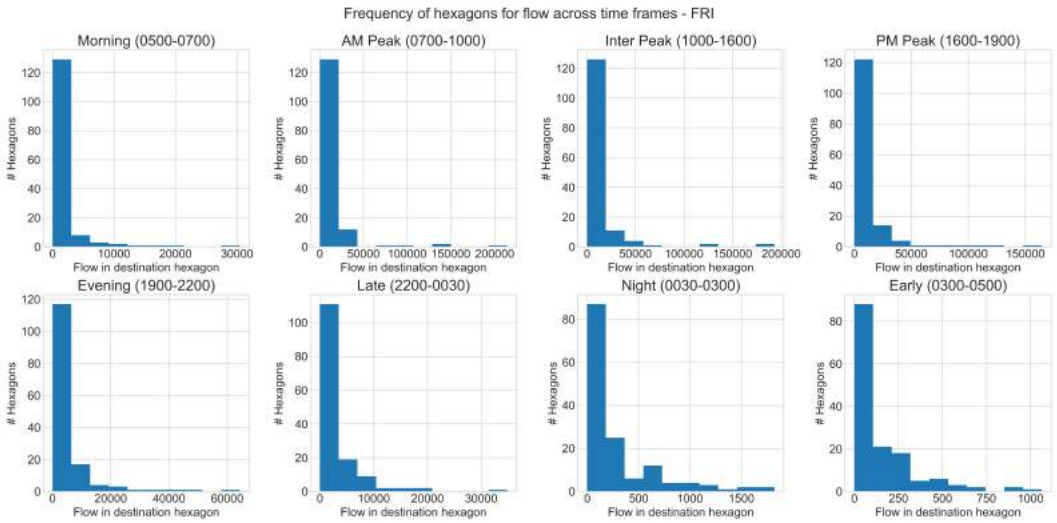

Figure 8.12: Distribution of flow for time frames of the day across distance quantiles - FRI

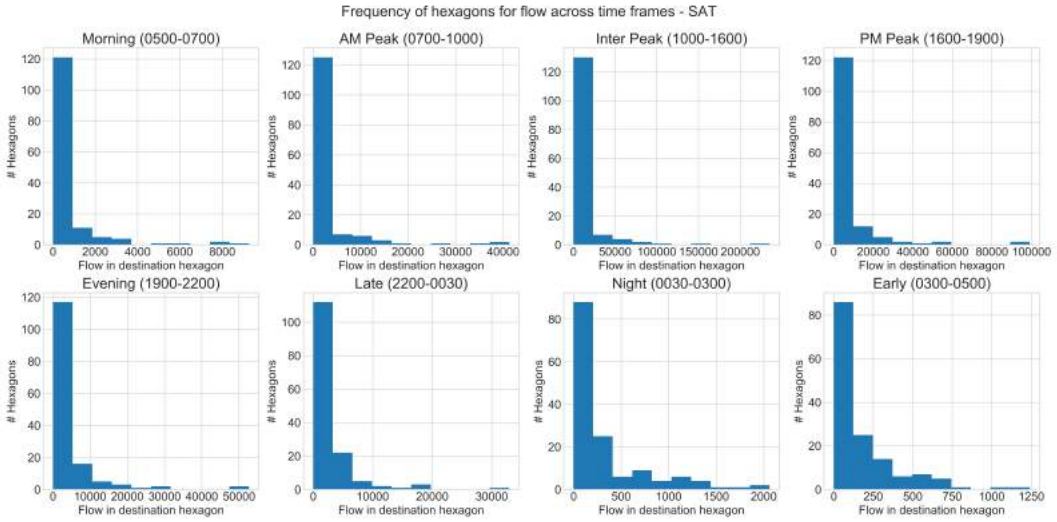

Figure 8.13: Distribution of flow for time frames of the day across distance quantiles - SAT

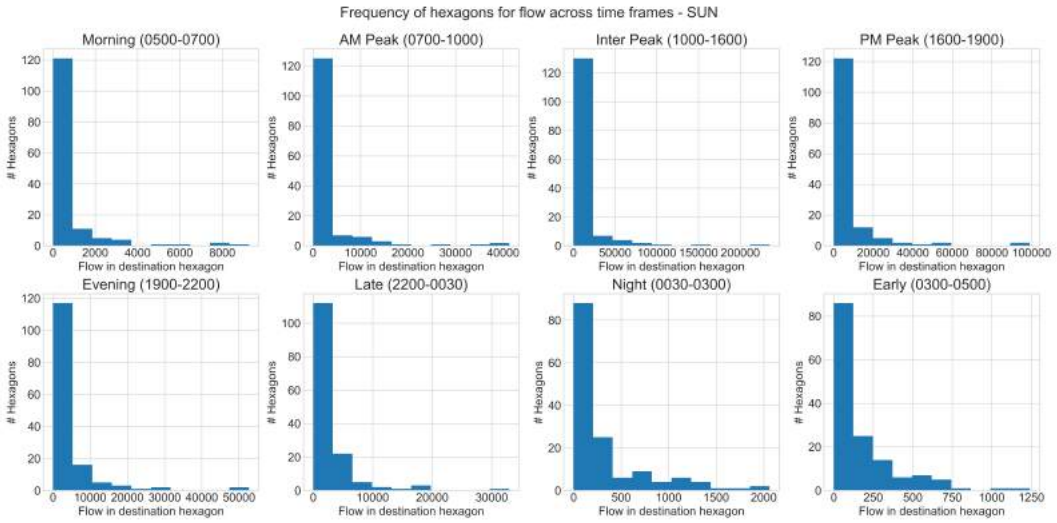

Figure 8.14: Distribution of flow for time frames of the day across distance quantiles - SUN

Distribution of flow for time frames of the day across distance deciles

The distribution of flow for time frames of the day across distance deciles are represented in the figures 8.15, 8.16 and 8.17 for the days FRI, SAT and SUN respectively. They exhibit similar patterns when compared between themselves and similar to the day MTT.

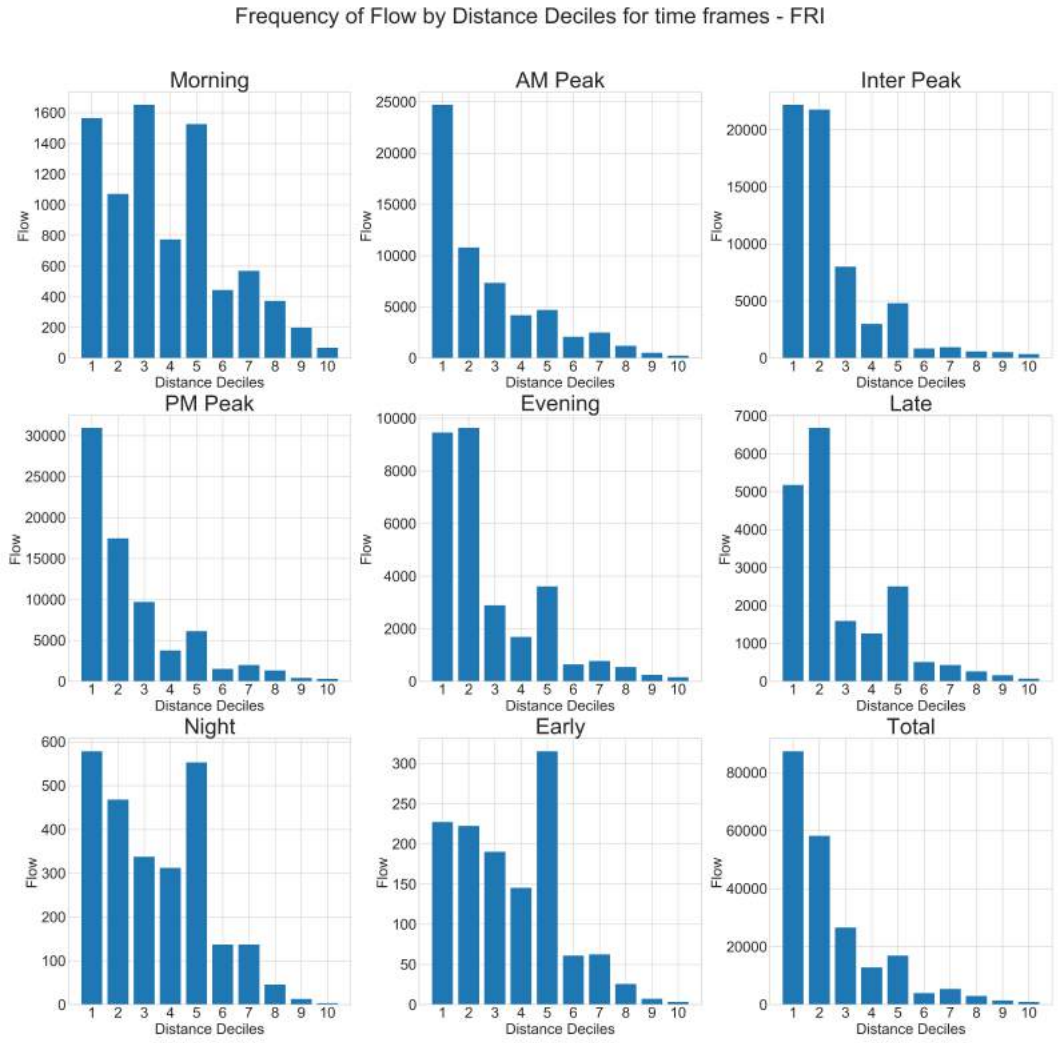

Figure 8.15: Distribution of flow for time frames of the day across distance deciles - FRI

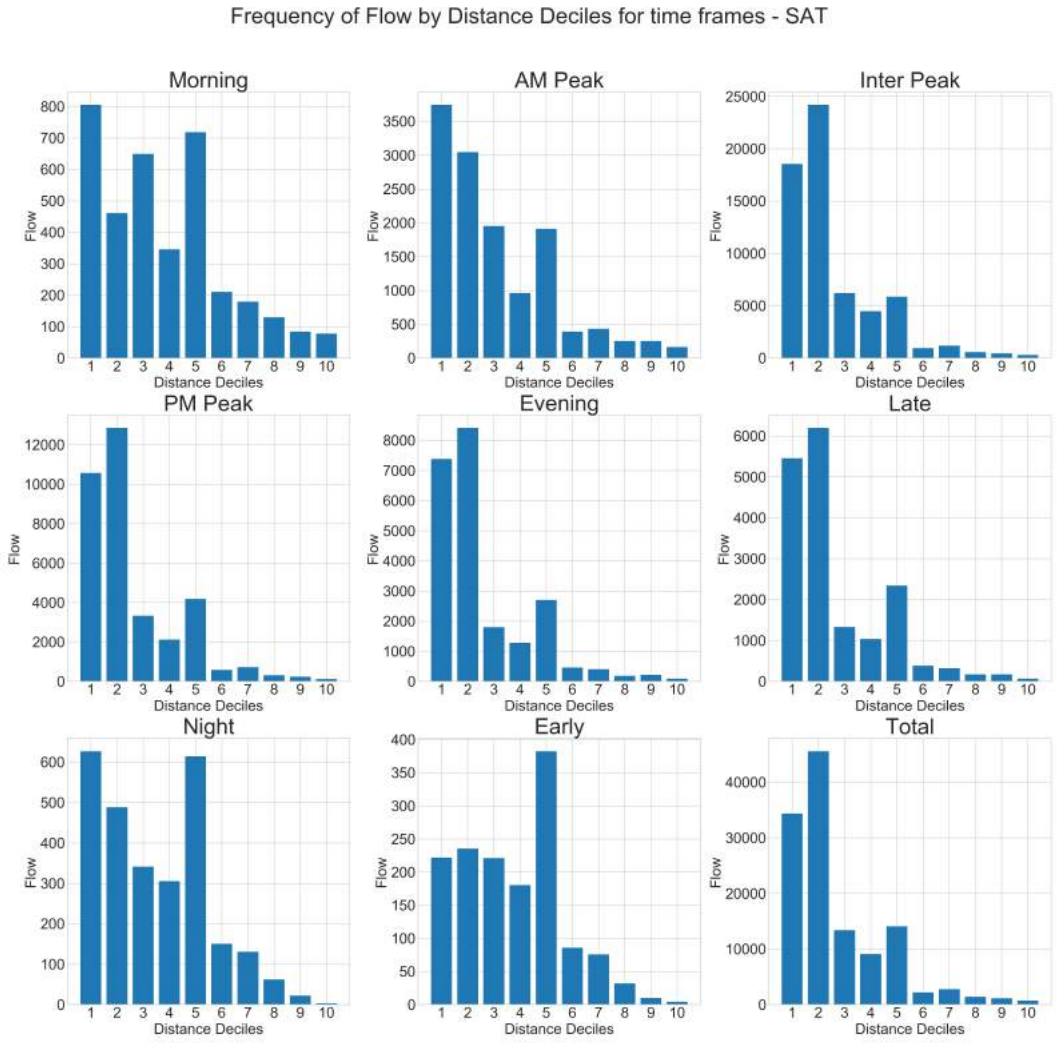

Figure 8.16: Distribution of flow for time frames of the day across distance deciles - SAT

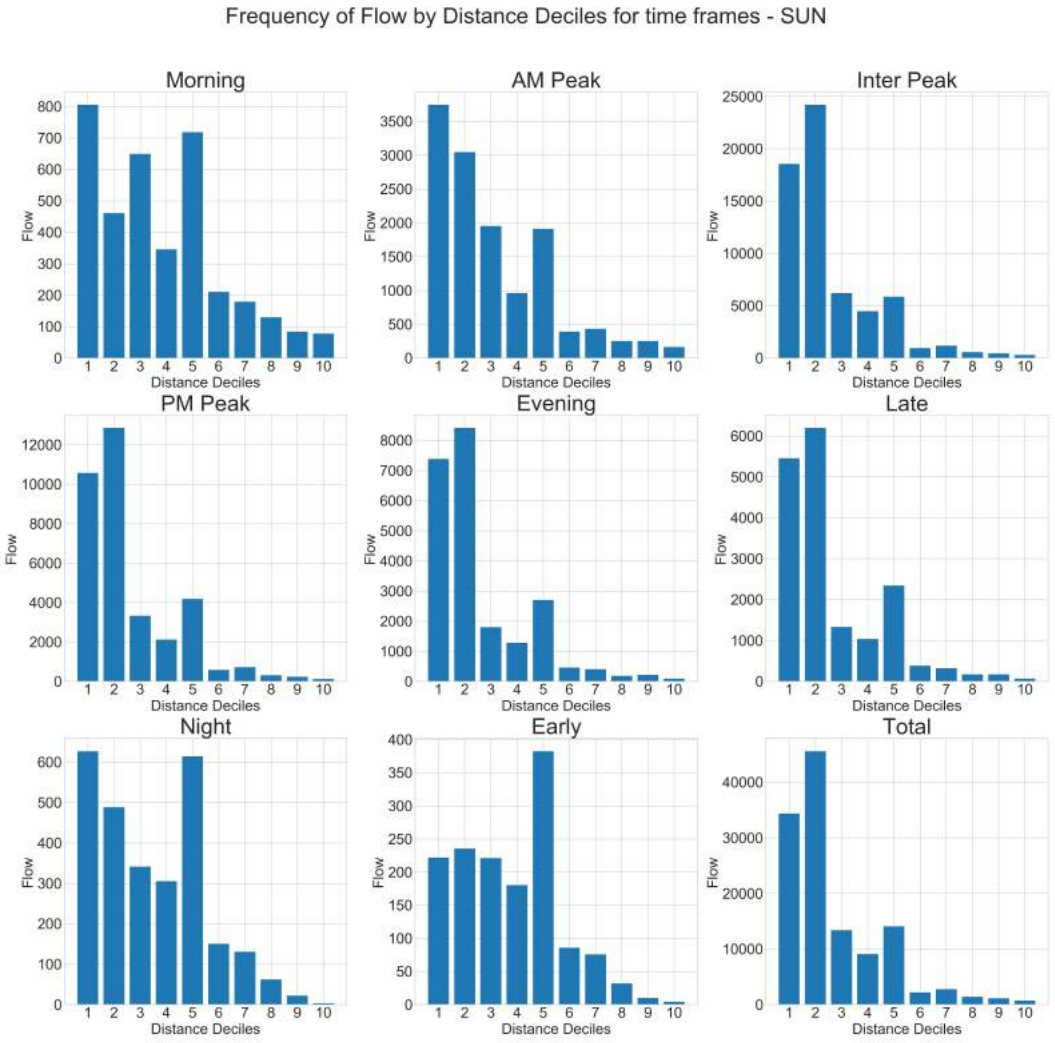

Figure 8.17: Distribution of flow for time frames of the day across distance deciles - SUN

Frequency of Hexagons for travel flow across time frames of the day

The Frequency of Hexagons for travel flow across time frames of the day are represented in the figures 8.18, 8.19 and 8.20 for the days FRI, SAT and SUN respectively. They exhibit similar patterns when compared between themselves and similar to the day MTT.

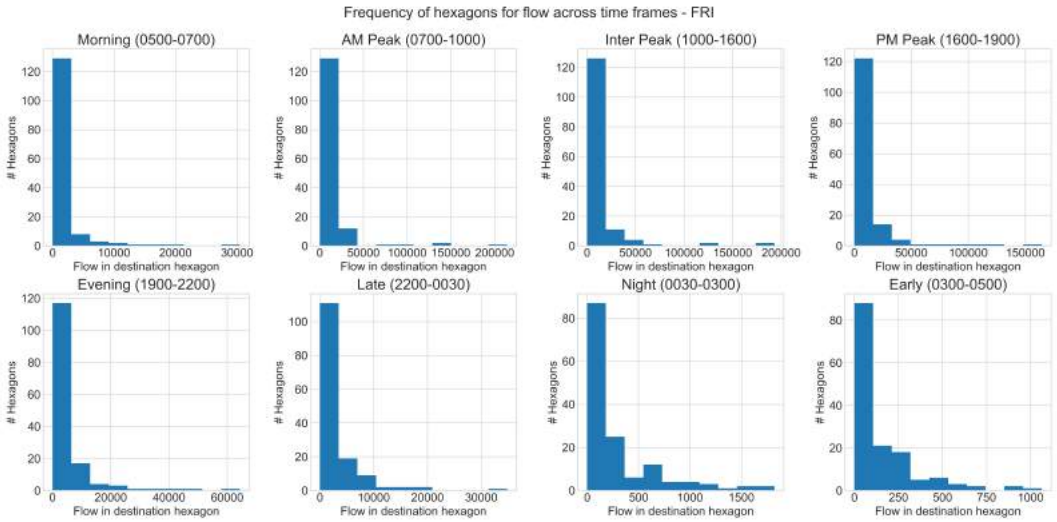

Figure 8.18: Frequency of Hexagons for travel flow across time frames of the day - FRI

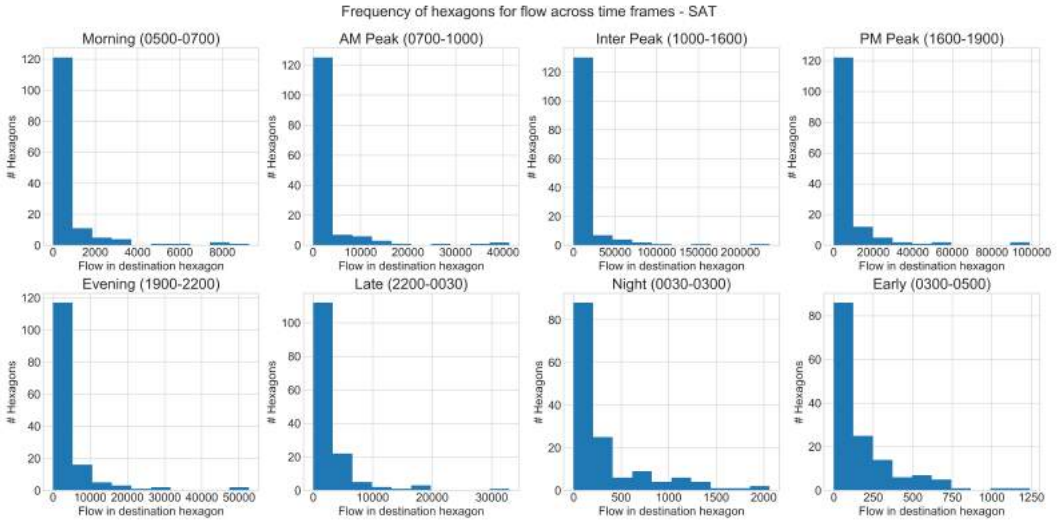

Figure 8.19: Frequency of Hexagons for travel flow across time frames of the day - SAT

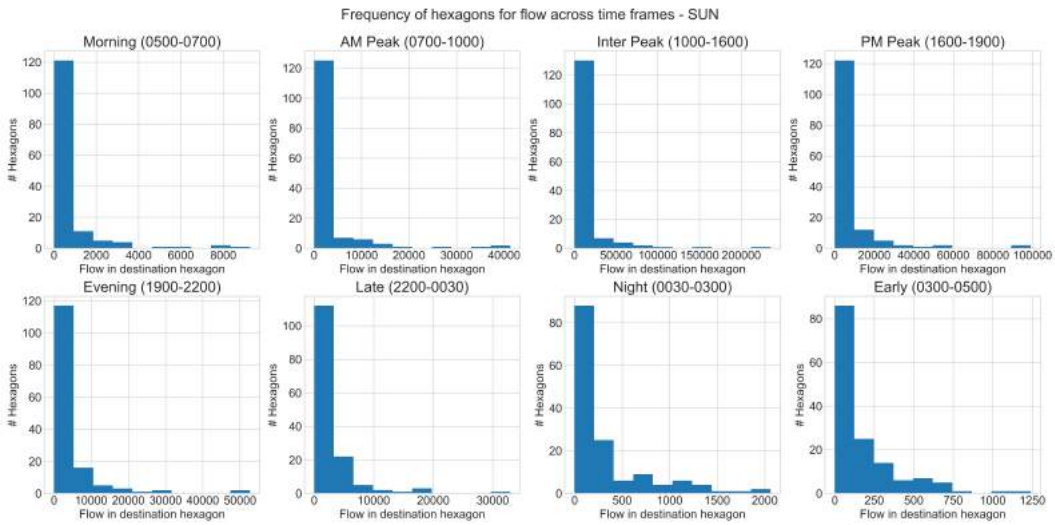

Figure 8.20: Frequency of Hexagons for travel flow across time frames of the day - SUN

Average distance traveled to a destination hexagon

The average distance traveled to a destination hexagon are represented in the figures 8.21, 8.22 and 8.23 for the days FRI, SAT and SUN respectively. They exhibit similar patterns when compared between themselves and similar to the day MTT.

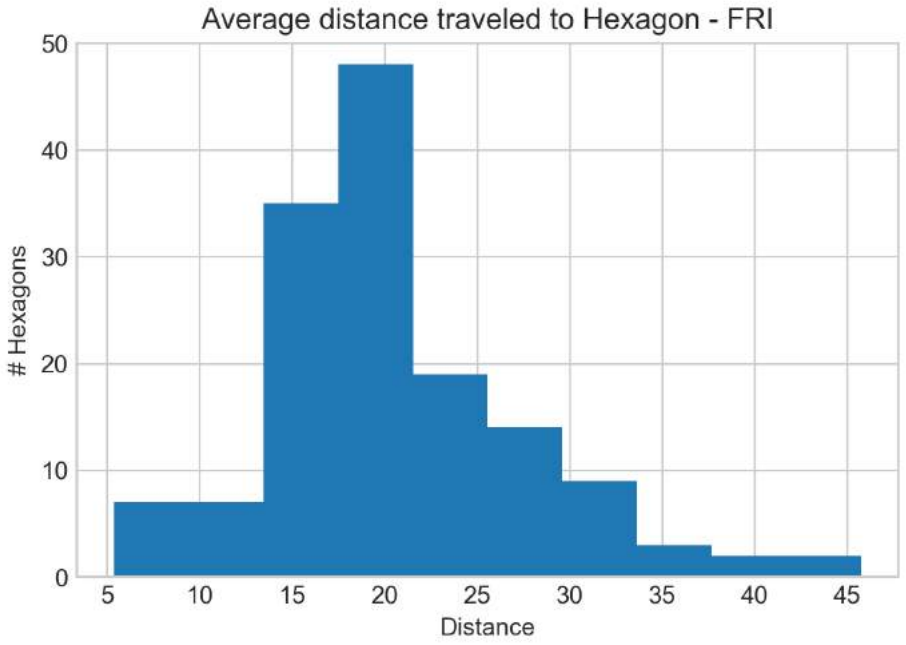

Figure 8.21: Average distance traveled to a destination hexagon - FRI

The plots for Set VI for MTT are presented below.

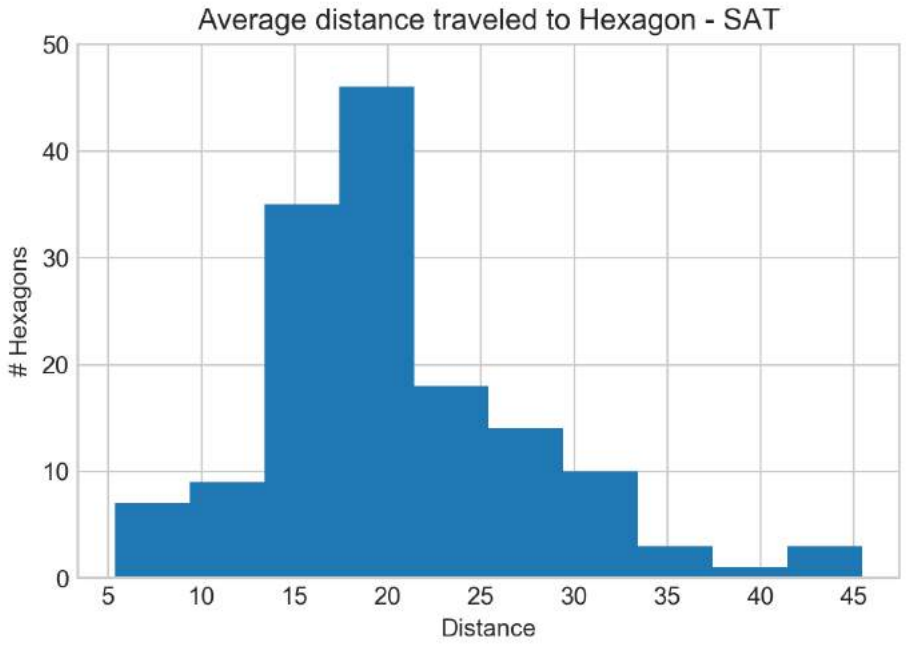

Figure 8.22: Average distance traveled to a destination hexagon - SAT

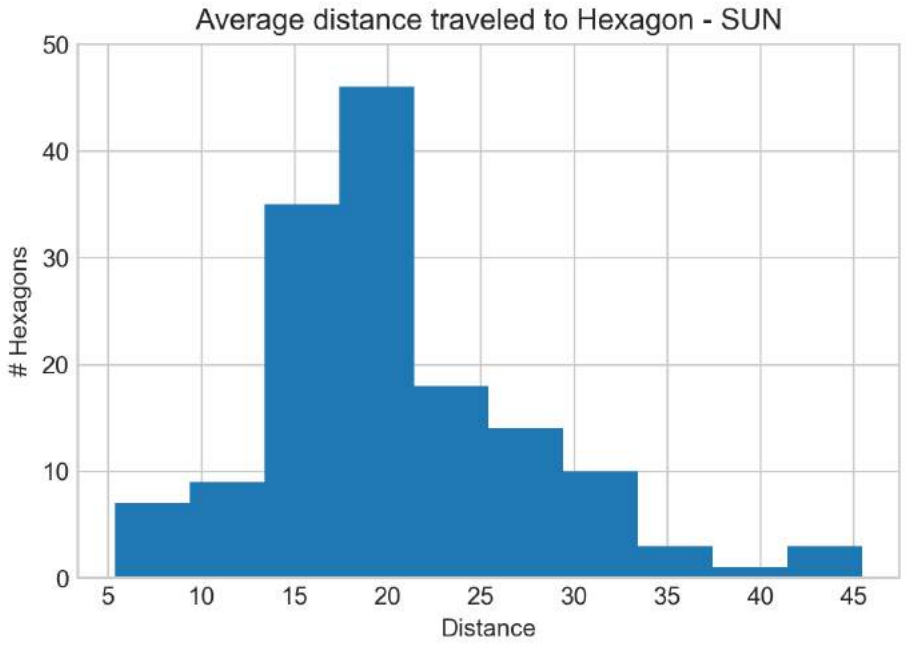

Figure 8.23: Average distance traveled to a destination hexagon - SUN

8.3. Experiment summary

Set III experiments summary - MTT

The experiments summary for the models in Set III (MTT) is presented in the figure 8.23. The R2 values is extremely poor and the results should not be considered further.

	Experiment	RMSE	R2	AIC
0	nbe1	77.50024	0.131635	113173.7
1	nbe2	80.26455	0.118545	113175.0
2	nbe3	78.18864	0.123335	113187.7
3	nbe4	79.75147	0.120674	113188.7
4	nbe5	77.09916	0.131074	113188.6
5	nbe6	74.99839	0.148862	113281.0
6	pe1	74.18736	0.160542	1020522.0
7	pe2	78.82889	0.058416	1020533.0
8	pe_g	78.39461	0.061790	1144567.0
9	pe3	74.54145	0.152808	1029031.0

Figure 8.24: Set III experiments summary - MTT

Set III experiments summary - FRI

The experiments summary for the models in Set III (FRI) is presented in the figure 8.25. The R2 values is extremely poor and the results should not be considered further.

	Experiment	RMSE	R2	AIC
0	nbe1	76.97454	0.131456	112559.1
1	nbe2	80.53796	0.115563	112562.1
2	nbe3	77.79193	0.123078	112576.4
3	nbe4	79.02031	0.120580	112577.8
4	nbe5	76.36056	0.132541	112576.6
5	nbe6	74.33541	0.147876	112686.2
6	pe1	73.47061	0.161386	999488.8
7	pe2	78.16985	0.057162	999511.6
8	pe_g	77.60110	0.063694	1119498.0
9	pe3	73.85517	0.152817	1008953.0

Figure 8.25: Set III experiments summary - FRI

Set IV experiments summary - MTT

The experiments summary for the models in Set IV is presented in the figure 8.26. The R2 values is extremely poor and the results should not be considered further.

	Experiment	RMSE	R2	AIC
0	nbe1	754.4309	0.245637	158297.9
1	nbe2	3216.3970	0.162284	157668.0
2	nb_g	806.8122	0.123077	158760.5
3	nbe3	823.2820	0.194439	157774.1
4	nbe4	1284.6230	0.180859	158097.8
5	nbe5	751.2308	0.237944	157371.2
6	nbe6	742.8290	0.254083	157531.4
7	pe1	707.3618	0.321491	7597886.0
8	pe2	829.6164	0.079520	7597901.0
9	pe_g	803.0375	0.125618	10164620.0
10	pe3	719.8224	0.297669	7855916.0

Figure 8.26: Set IV experiments summary - MTT

Set VI experiments summary - MTT

The experiments summary for the models in Set VI is presented in the figure 8.27. The R2 values is extremely poor and the results should not be considered further.

	Experiment	RMSE	R2	AIC
0	nb1	1734.937	0.267525	150191.5
1	nb2	1764.308	0.265834	150201.6
2	p1	1826.529	0.117636	11089200.0
3	p2	1135.317	0.672404	2759365.0

Figure 8.27: Set VI experiments summary - MTT

8.4. POI Scatter Plot - Linear Regression

Linear regression is fit to the plots to measure the goodness-of-fit metric. For any of the categories, the R^2 value is quite low indicating that all the categories are not following a linear regression and are widely distributed. Thus, linear regression is not suitable for estimating POIs and other methods such as multivariate regression is necessary.

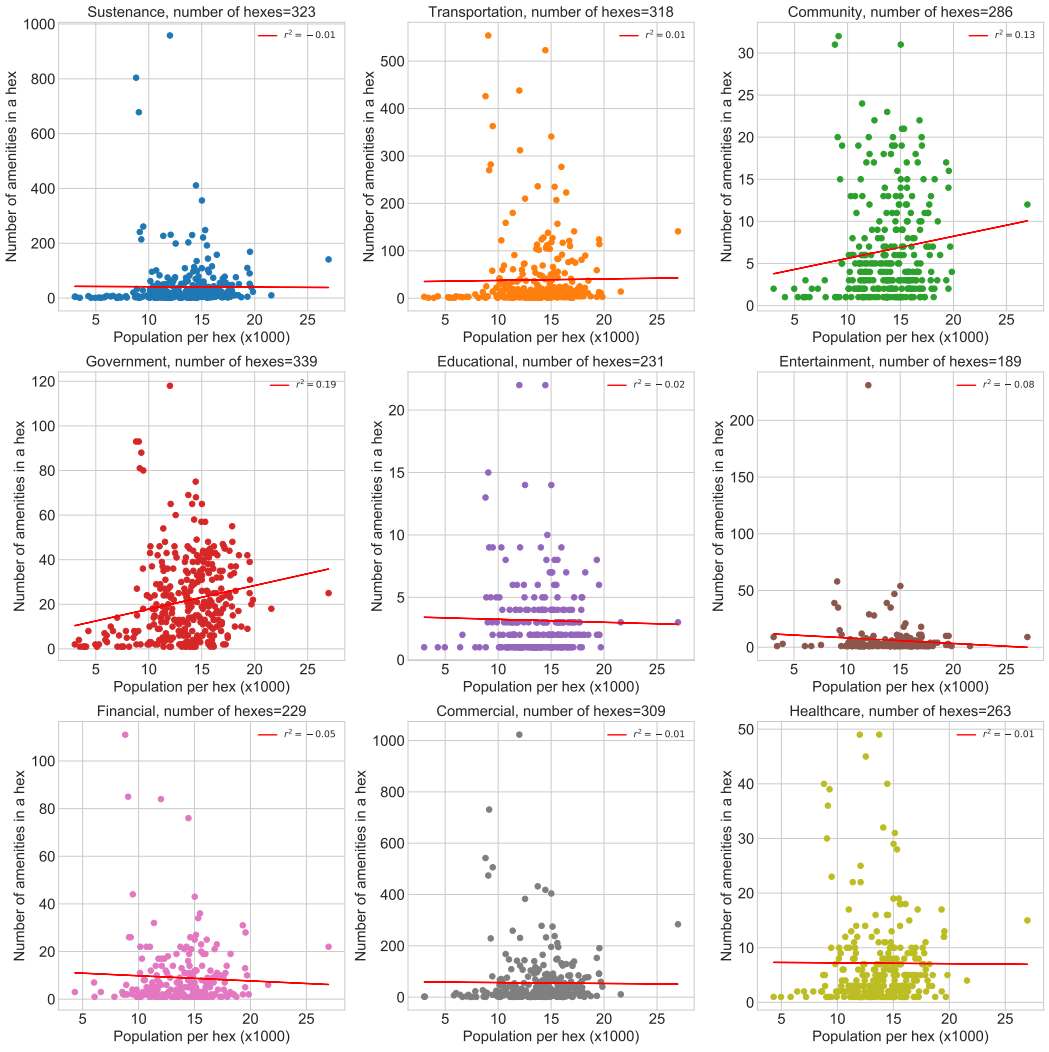

Figure 8.28: Set VI experiments summary - MTT

8.5. Abbreviation table

The table 8.1 below presents the abbreviations used in the report and their corresponding explanations.

Table 8.1: Abbreviation table

Abbreviation	Explanation
MTT	Monday to Thursday
FRI	Friday
SAT	Saturday
SUN	Sunday
POI	Points of Interest
GLM	Generalized Linear models
NB	Negative Binomial
OSM	OpenStreetMap
SSI	Sorensex Similarity Index
AIC	Akaike information criterion
RMSE	Root-mean-square error
R2	Coefficient of determination (Rsquared)
LBSN	Location based social network
OD	Origin- Destination
ML	Machine Learning
GIS	Geographic Information System
TfL	Transport for London
LU	London Underground
LO	London Overground
DLR	Docklands Light Railway
OA	Census Output Areas
MAUP	Modifiable Areal Unit Problem
O	Origin
D	Destination

8.6. Reproducible research

The source code is available at https://github.com/karanizer/Estimating-Mobility-POI. Moreover, almost all of the data sources can be downloaded via notebooks that are provided. Thus, a researcher can operate with raw data, run prepossessing scripts, and verify the results. For the data sources that were retrieved manually, a comprehensive description of the process is provided in the chapter 3.