Optimising Battery Replacement in a Fleet of Scooters

Joshua Price

November 12, 2022

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

CONTENTS

ABSTRACT

The emergence of hireable electric scooters presents a unique mode of travel within cities, and as such have their own unique challenges. Solving these challenges could reduce costs of the service, allowing these scooters to be an effective method to reduce travel time in one's commute. One of these challenges is recharging, and determining which scooters are in most need of battery replacement. Reducing the number of recharges required to satisfy demand will reduce the costs of energy and labour. This research analyses user and trip patterns to develop a metric that measures the effect that replacement of an individual scooter has on overall revenue gained, and develops an algorithm to optimise battery replacement.

In Chapters 2 and 3, data provided from the Beam e-scooter service implemented in Adelaide, South Australia, Australia, is analysed to develop a metric that could identify areas of priority for battery replacements. Out of the classification methods tested, simple binning was chosen to divide the area. It was found that the distribution of revenue obtained from trips starting in each location was quite consistent.

In Chapter 4 a metric is developed that takes estimated demand and revenue distributions in order to estimate the expected increase in revenue when a given number of batteries are replaced in a given location. These estimates are taken from previously observed patterns.

In Chapter 5 this metric is applied to simulated data and simulations are made where batteries are replaced based on the metric. An optimisation heuristic is used to select scooters, and the resulting revenue is compared against that of simulations that approximate how Beam employees select scooters to recharge. Due to the consistency of revenue patterns across the observed area, demand has a much greater impact on whether scooters should be recharged, since this is how Beam employees tend to prioritise known hot spots. However, the metric takes into account when trips end in higher demand locations, and this was observed to result in a small increase in revenue, so a modified approach to how Beam operates may increase profits.

We conclude with a discussion and outlook for future research in Chapter 6.

1. INTRODUCTION

In 2016, Neuron produced the first model of electric scooters and began to deploy them in various countries, to be hired by any of the locals [12]. This research focuses on the behaviour of scooters in South Australia, which has had access to these e-scooters since the beginning of 2019 [13]. The state has provided e-scooters from four different services, however only two operate currently: Beam and Neuron. This research uses data from two services: RIDE, which operated from mid-2019 to mid-2020, and Beam, which began operating in mid 2019, until becoming unavailable due to the Covid-19 pandemic and returning in mid 2020. The RIDE data includes trips and status information in June to August 2019, while the Beam data starts from when Beam services returned in August of 2020.

Once e-scooters become available in an area, it tends to take some time for locals to become accustomed with their existence, and frequent user patterns usually emerge after a couple of months[1]. Initially, e-scooters were only available in the Adelaide Central Business District (CBD), however since this research started both companies currently operating have expanded their area of availability. Due to the time required for patterns to emerge in new areas however, this research is mostly concerned with the central CBD area as it has a sufficient amount of reliable data.

Improving the effectiveness of e-scooters has relevance in society as they provide a potential solution to the 'last-mile' problem. The last-mile problem refers to how the last portion of a trip often takes the longest, usually relating to goods transportation and commuting, as

[1] This information was ascertained from personal discussions with Beam employees in regards to potential considerations for patterns observed in the data. When asked about why very few trips were occurring in newly accessible areas, the response was that consumers take some time to become accustomed to the existence of e-scooters in their neighbourhood, and this pattern is observed even in busy areas like the CBD.

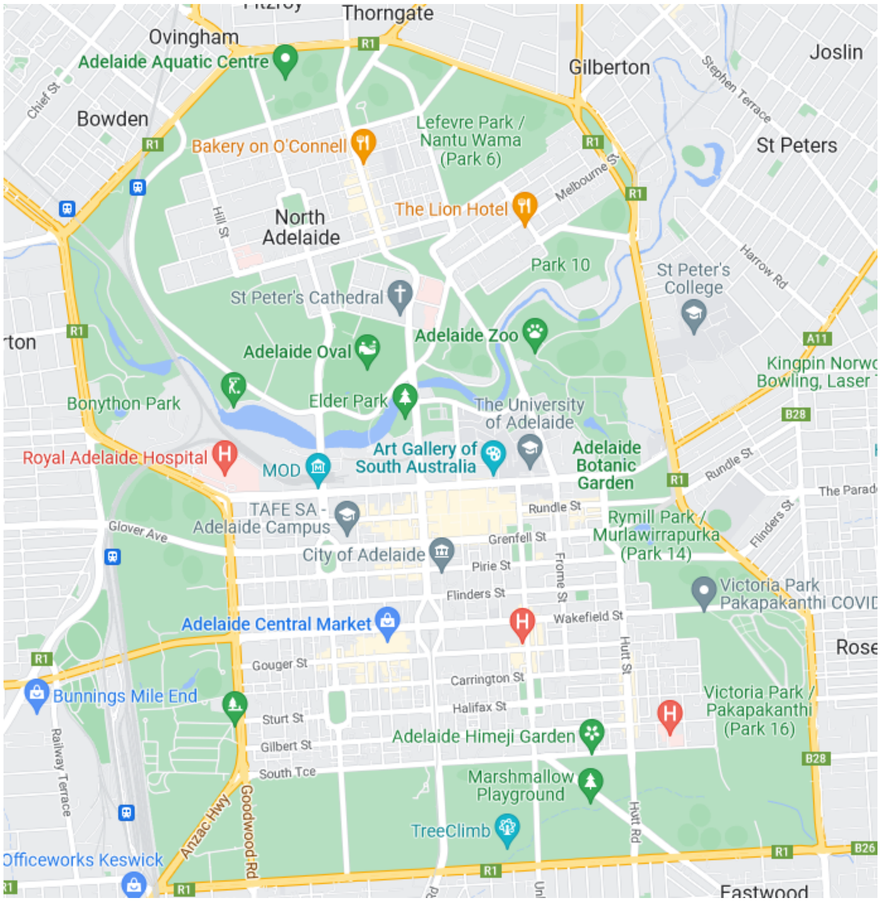


Fig. 1.1: The area of Adelaide observed in this research. The CBD takes up the southern area in this map, while the North Adelaide area across the river is also available for scooter use. Beam increased the area of operations further to the east and south of this map during this research, however that data is not used in this thesis.

these are areas where the destinations are normally focused in the same small geographical regions [19].

For the example of commuting, as of the 2016 census, over 117,000 South Australians were employed in offices located in the CBD [1]. 95.5% of these residents who work in the CBD commute from surrounding suburbs, so at peak hours traffic density will be highest at this location as over 100,000 employees travel there. E-scooters allow commuters to park outside the CBD and ride in, using a small vehicle that takes up less space, reducing congestion.

Providing commuters with an alternate method of travel has a number of benefits apart from avoiding slow traffic. Stop-and-start traffic in areas of high congestion uses more fuel and causes more wear on the vehicles than driving the speed limit. E-scooters could reduce the pollution generated, and the costs and resources used in central areas. In a city like Adelaide, where taller buildings are increasingly constructed to accommodate more workers, congestion will increase as the population of the surrounding area continues to grow. As such, methods of travel like e-scooters may become even more relevant over time.

Another description of the last-mile problem is the use of public transport. Users of public transport are not often dropped off in locations in the near vicinity of their destination. This causes the last portion of the trip to take longer as users typically need to walk the rest of the distance. By providing e-scooters at the locations of public transport stops, users can ride to their destination more quickly.

Commercially available e-scooters come in two varieties: one where the unit is plugged directly into a power socket to be recharged, and another where the battery is removed and placed into a charging dock. The advantage of the latter is that if multiple batteries are owned, the battery can simply be replaced with another, so the scooter can still be used while the battery is charging.

Both of the companies that provided data for this research provide models of e-scooters that contain replaceable batteries as this allows an employee to travel to scooters that are low on charge, and replace the battery with a fully charged one. This means the scooter is recharged immediately, rather than being taken to another location to be recharged, becoming unavailable for some time.

Even though the process of charging is quick, travelling to individual scooters can take time and presents a unique problem. The effectiveness of a replacement could be dependant on where the replacement occurs, not just on the amount of charge left in the battery being replaced. Not only that, but there could be areas where it is more effective to replace a number of batteries at once, which would require less travelling.

1.1 Literature Review

Schellong *et al.* [18] estimate that the global market share for e-scooter rides could potentially reach $40-50 billion by 2025, but they are still yet to be profitable. Maximising revenue gained per replacement could reduce costs, providing a potential solution to this issue.

Noland [15] found that scooters are rarely used for commutes, by observing the times that they are usually hired, and comparing the trip patterns to those of casual users of bikeshare systems. However, trip distances were shorter on weekdays, implying that short commutes could be common. This could imply that public transport users are using e-scooters to solve their definition of the last-mile problem.

Conversely, a survey found that those that did not commute to work were less likely to consider using e-scooters [11]. A possible explanation for this contradiction is that people may expect scooters to be useful for daily travel, but in practice find that they are better suited for less frequent use, due to availability and cost. Mitra [11] cites analysis of real world e-scooter data from McKenzie [10], which supports that users more commonly use scooters to travel to recreational or commercial areas. McKenzie also came to the conclusion that in a city that provided both bike-sharing and scooter-sharing services, commuters primarily used the bikes for travel. However, they make note that at the time of writing the scooter service had only existed for a year, whereas people had been using bikes for over a decade. Bicycles also have their own lanes for use which make them more effective modes of transport, while scooters share either the footpath or the road. In Adelaide, bike-share systems are far less ubiquitous, and there are also very few bike lanes in the central city. It is still possible that scooters could show commuter patterns in the provided data. Capital Bikeshare, which provides the bicycles used in the Washington D.C. data analysed in [10], has two separate pricing plans that were used to classify users. Those that paid a monthly fee for the service showed patterns that would be expected in commuters – high use before 9am and after 5pm – whereas those that paid for each trip showed a more consistent use pattern throughout the day. Beam has recently released a monthly pricing option for users, however the data does not contain pricing information. This monthly option may motivate more users to add e-scooters to their commute.

A network of e-scooters provides unique challenges for any research, let alone the specific battery replacement work presented in this thesis. Despite extensive research into travel and transportation, these scooters have many characteristics that other vehicles do not. Transportation by car for instance, is constantly being observed and simulated. However, e-scooters are not restricted by roads like cars are and so a priority metric can not be made based on those previous simulations. E-scooters can be ridden on side walks and roads, but can also be ridden on paths. The term Free Floating Sharing Service (FFSS) is sometimes used to refer to e-scooters, as the locations they can be found in is much less restrictive than many other vehicles [4].

As vehicles such as cars, trains and bicycles are at least centuries old, critical research done on the patterns of use are old as well. On the contrary, e-scooters are very new and as such there is little research on the areas that make these vehicles unique. Hence, the literature that was reviewed to form the basis of this thesis consists of older work done on other vehicles that may relate to what we want to understand about e-scooters.

Bike sharing systems have been available in many countries for a long time, as modern bike sharing evolved from older rental models dating as far back as 1965 [20], however as many are now electric they may provide similar characteristics as e-scooters. However, when bikes were originally introduced they could only be parked in certain docking stations, which is a station based approach that the free floating one has replaced in Adelaide[4]. Research such as [7], which analyses user patterns to allocate bicycles to meet demand, and Yoon *et al.* [21], which developed an application to predict route information given the origin and destination, forms a theoretical basis for Chapter 4 of this thesis. However, since these bikes are restricted in which locations they are found in, predictions of use are not actually based on geographical location but the station they are parked in. The free floating nature of e-scooters means that not only is location relevant, but coordinates need to be classified before they can be treated in the same way.

One method of classifying locations is a Traffic Analysis Zone (TAZ), which separates a continuous landscape into a finite number of locations, using density of trips. For instance, a dataset used by Gonzales *et al.* [6] contains similar trip information to that provided for this research, in that trips are presented as a pair of pickups and drop-offs. Gonzalaes developed a model that estimates the number of pickups and drop-offs in a given location. While this is not particularly relevant to this research, as knowing the amount of pickups does not provide all of the information needed to estimate battery usage, a TAZ produced with this data is still useful. A TAZ uses a heuristic to develop zone boundaries, which means that it is designed around the information acquired and requirements for classification. For instance, Gonzales *et al.* use census data and major landmarks to inform zoning.

Martinez *et al.*. [9] also divides geographical space using trip data. However, the heuristic used delineates the data in a way that is not appropriate for this research. Their algorithm attempted to reduce trip density in a given zone, which made zones larger if if more trips started within them. The opposite is desired in this work, as demand is to be observed. If a small area had high demand, its zone would be large, implying high demand in the area around it.

There is also a variety of work done in similar areas that attempts to achieve the same results of trip prediction, however they are not useful in terms of battery replacement priority.

For instance, IBM Research Ireland developed a predictive model that estimates trip-time and destination given the location of a taxi [8]. Since trip revenue generated by scooters depends entirely on trip duration, a model such as this would identify likely revenue from individual scooters. However, the dataset used in [8] contained frequent GPS updates, and the trajectory of the taxi was used to determine its likely trip-time. This means that the estimate was available only once a trip had started, whereas battery replacement must occur before a trip occurs.

This research is concerned with two common problems with electric scooters: battery replacement and trip prediction. Both of these problems have been extensively researched since the introduction of e-scooters, however rarely in the same context. Battery replacement is frequently observed in the context of commercially available scooters for personal use, and trip prediction does not usually take into account the battery use of the scooters.

Ciociola *et al.* [2] uses a simulation approach alongside Kernel Density Estimates to create a demand model to determine areas of scooter implementation that could be improved. They found that increasing the number of e-scooters available and battery replacements would satisfy demand and reduce cases where users run out of battery. However, it does not attempt to determine locations where battery replacements occur to maximise their effects, as increasing the number of scooters is sure to increase costs.

Pender *et al.* [16] use a model to optimise the fleet of scooters such that their battery levels meet demand. However similar to [2], Pender *et al.* determines the number of employees required to replace batteries at an efficient level. It follows that hiring more employees and making more replacements would increase costs for the company provisioning e-scooters, and neither of these papers take this into account as they are not concerned with reducing costs. Hence, this thesis, which attempts to reduce costs by optimising the effect of each replacement, compliments these previous works. Pender's work, however, does have some findings that are related to observations made in Chapter 3 of this thesis in terms of battery lost during trips.

Dong [3] presents a method of optimising e-scooter operation using clustering to group scooters. The motivation for clustering here was to simplify required services by grouping scooters in optimal locations decided by the clustering methods. This would mean that e-scooters were required to be placed in these physical locations by both employees and users, which is not possible in our context. Clustering is still used in this thesis to classify locations for analysis in Chapter 2, however binning into a grid of locations was found to be the preferred method out of the clustering algorithms considered. Da Dong considers a travelling salesmen-like approach to optimising replacements, by developing an algorithm that produces a route that passes through each cluster and replaces the batteries of all scooters with battery percentage below a certain threshold. This is what makes Da Dong's work distinct from this thesis, as in Chapter 4 a metric is designed to base the replacement decision not on a threshold but instead from user patterns in that location. The route development does provide potential future work however, as simulations in Chapter 5 do not take into account the distance between replacements. Also, the more detailed clustering methods presented may provide a more effective classification of e-scooter locations.

Another option for geographical zoning is Geohashing, developed by Niemeyer [14]. Geohashing is method to represent the longitude and latitude value of coordinates as a single value that collects locations that are nearby each other. The result is a grid of locations labelled with this value. In this thesis, a similar effect is achieved by rounding coordinate values with a number of significant figures, and then combining those values into one string. Unlike Geohashing, this does not result in values that provide a mathematical representation of the distance between two zones, however this was not required. As such, Geohashing is not used, however is a potential method of interest for future work.

The R programming language [17] is used for all data analysis in this thesis. All code used to create the plots and findings is available in the public Github repository: `https://github.cs.adelaide.edu.au/a1705558/RIDE`.

The remainder of this thesis is structured as follows. In Chapter 2 we clean the data and classify geographical coordinates into discrete locations. In Chapter 3 we develop a predictive model to estimate how much revenue a scooter would generate if it had its battery replaced, using historical pattern from trips that occur in the location obtained from the previous chapter. In Chapter 3 this revenue is combined with demand patterns to develop the priority metric which provides a value for how effective a battery replacement in a given scooter will be. Lastly, in Chapter 5 simulations of scooters and trips are performed to observe if battery replacements using this metric result in more trips being satisfied that return higher profit. Chapter 6 concludes this thesis with an outlook for future research.

2. DATA OUTLINE AND ANALYSIS

A method to determine an optimal way to replace scooter batteries was outlined in Chapter 1. This method will be informed by historical data, to derive a decision making process that is based on patterns that have been observed in the past.

RIDE provided datasets of both observed trips and scooter statuses throughout 2019. This included over 25 thousand trips during a period of time between February and July of that year, and the statuses of 450 scooters at half an hour intervals in during June and July. Beam provided an API, which is a source that provides data when requested and is constantly updating. This allows for data to be obtained from a several distinct data-ssets, with recordings as early as August of 2020 until the present. As it contains the most recent, largest, and most reliable dataset, the Beam API was used for all research once access was granted. Ride data was used up until this point, for initial analysis.

This chapter will describe the main scooter dataset used for research, alongside supplementary datasets used for potential extension. Exploratory data analysis is performed, to outline observed patterns in scooter trips. These patterns can be used to inform the algorithm used to rank scooters based on priority in terms of battery replacement.

2.1 Outline

The Beam API provides a number of distinct datasets, however three are considered for this research:

1. Status: Which contains information regarding events. That is, the starts and ends of either maintenance or user trips, and the scooter's status at those times. This includes

battery percent, location, and the scooter's unique ID.

2. Route: Which contains information for all trips that have occurred. Whereas the Status set contains information at the start and end of a trip (in latitude/longitude), the Route set contains information during the trip, such as the positions of the scooter at certain intervals of time, and values for the overall distance and duration.

3. Free Bike: Which contains the locations of all scooters at the current time.

From these three datasets, the following relevant data is available:

- Trip start and end time (in POSIXct format, which contains data and time as a single value in milliseconds).
- Trip origin and destination (latitude and longitude).
- Trip duration - which can be used to calculate trip revenue (obtained from start and end time).
- Trip distance (in metres).
- Battery charge lost during trips (as a percentage).
- Scooter location outside of trips (latitude and longitude).
- Scooter ID - to keep track of the same scooters throughout trips.
- Demand in individual locations, which is obtained from the frequency of trips starting in those locations.

Since the Status data contains the start and end times of a trip, the duration can actually be ascertained from the Status data alone. The path that the user takes during the trip could be useful for terrain/traffic information in potential future work, but is not observed in this research. As such, the only extra information the Route data provides is the travel distance.

Using the Manhattan distance between the start and end locations, the travel distance could be estimated from the Status data, but it should not be assumed that users will always take the most direct route.

Demand could be naively estimated from the events contained within the Status data, however it is important to gain an understanding of the supply to provide an accurate estimate (ie. how many scooters are available to users in a given location). The free bike API provides this information, however does not provide historical data, only the current locations of scooters at the time it is requested. A script was written that queries the API every half an hour, and the results are saved so a dataset containing information over a long period of time can be observed. The script has been running consistently since late June 2021, and the dataset used here contains location information from then until October 2021.

Once a method for determining an optimal order of battery replacement was developed, the effectiveness was tested on a simulated environment where scooters changed location and their batteries were depleted according to simulated trips. The location data is useful as this allows for a simulation to be designed where the location of scooters is similar to what occurs in reality.

The free bike API does not provide the battery level of scooters at the queried time. This would be useful for the simulation as it provides information on areas that frequently have low battery scooters and how users choose scooters based on the battery percentage of scooters around them. A simulation could then set the battery level of individual scooters to values based on real life patterns, and the decisions users make could influence which trips occur and which scooters are used. However, the battery levels of scooters in the simulation are hence chosen at random as this information is not provided.

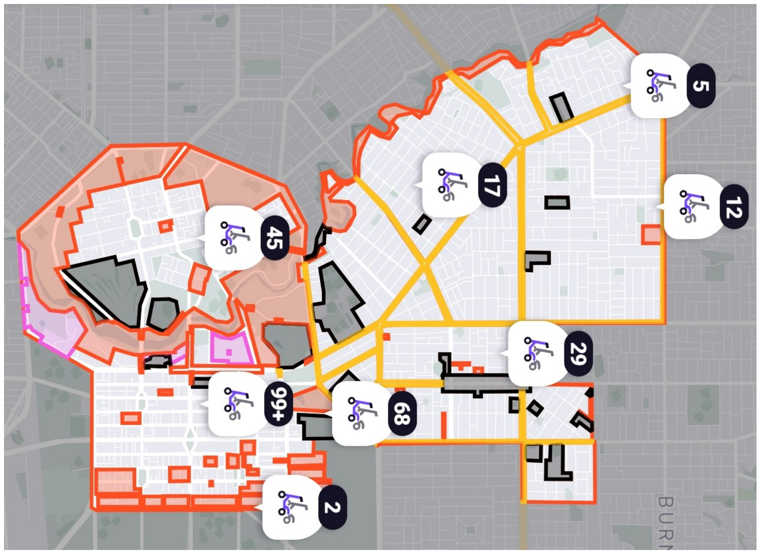


Fig. 2.1: The usable space considered for this research. This is a screenshot of the Beam mobile application, which can only be viewed in landscape.

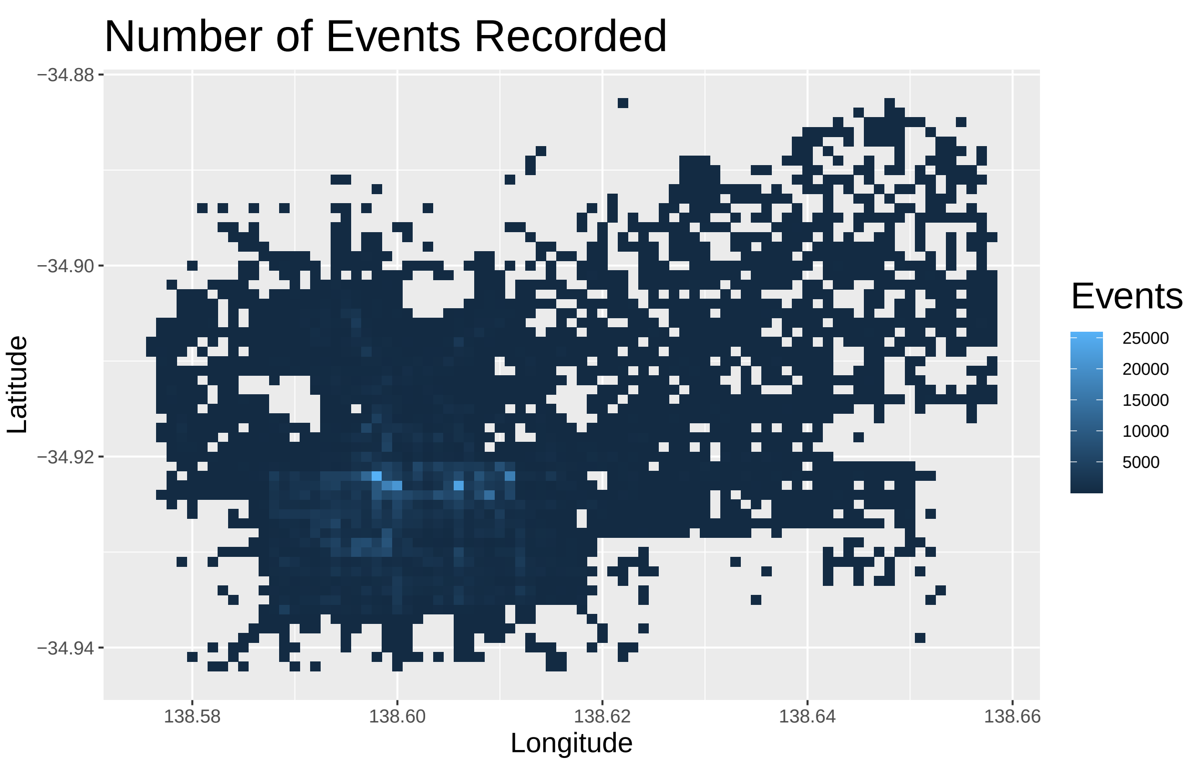


Fig. 2.2: Number of events in each location over a one year period.

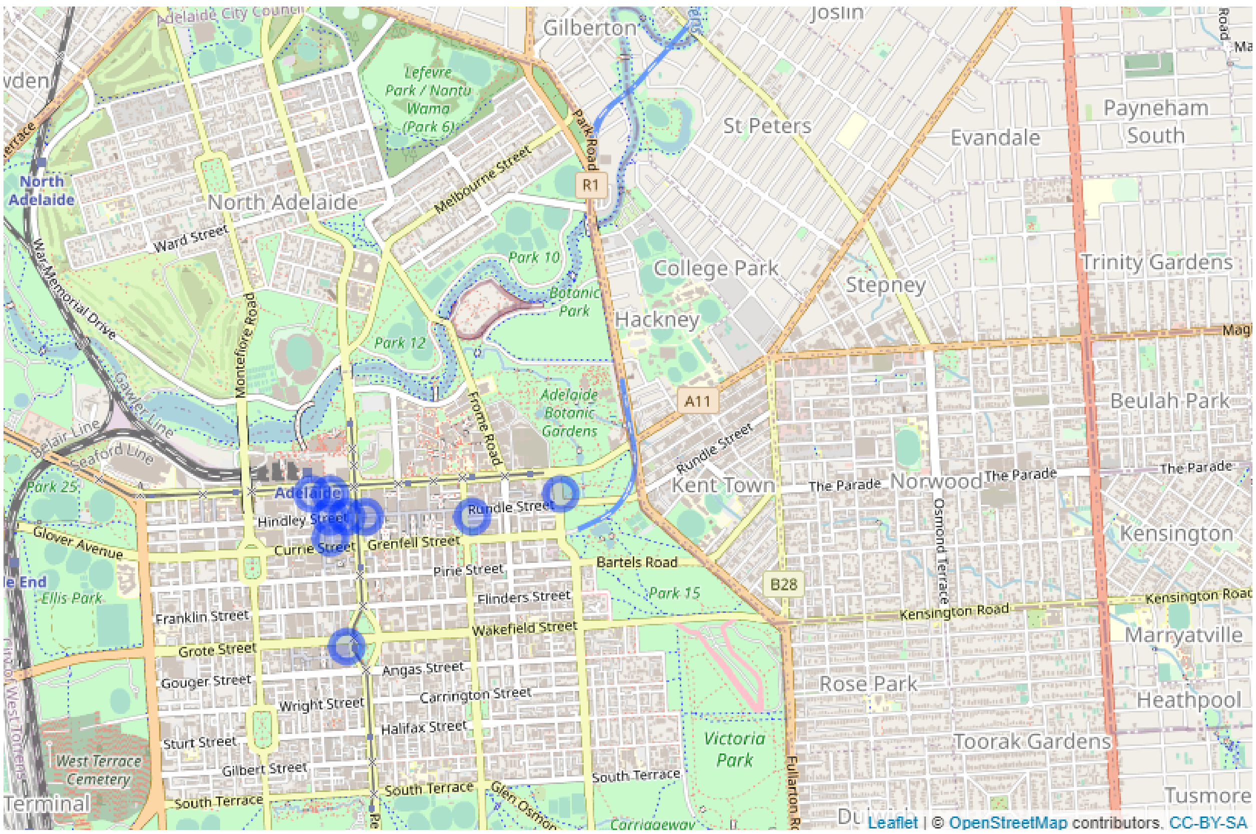


Fig. 2.3: Map of the nine most popular trip starting locations.

2.2 Analysis

2.2.1 Locations

The authorisation provided allows access to data relating to scooters used throughout South Australia. This includes the central area, and also some beaches off the west coast. In these beach areas it is required that scooters park in specific parking locations, whereas throughout the CBD users have much more freedom. As outlined in Chapter 1, this research is concerned with transportation that does not have restrictions such as parking stations. As such, only the central area is considered here.

Figure 2.1 shows this central area. Grey areas are sections where Beam scooters cannot be ridden, whereas they can be ridden in red sections, but must be parked elsewhere. Orange sections represent areas where speed is restricted, usually as they take place on main roads with heavy traffic. The bottom-left section contains the Adelaide CBD, and the area above it contains North Adelaide, a less busy area with apartments, restaurants, and cafes. To the right of both of these, is the Norwood/Payneham area, which is mostly suburbia, but does contain a shopping district near the bottom called The Parade.

Figure 2.2 presents the locations of 680,000 events from a dataset generated from the Status API between October 2020 and August 2021. The distribution of these events reveal the locations of hot-spots within the area defined by Beam, which are all located in the bottom left quadrant of the plot, which contains the Adelaide CBD. The coordinates relate to five different areas in the CBD, the train station, the two ends of Rundle Mall (users cannot ride inside the mall), the Gouger Street market/business area, and the corner of the north and east borders of the CBD.

The rest of the usable area is not as popular as the CBD. This includes the North Adelaide region and the Norwood region, however it is important to note that these regions have been usable for less time than the CBD, with the Norwood region opening in July 2021. These new locations allow users to ride directly from their homes into the city. However, at this

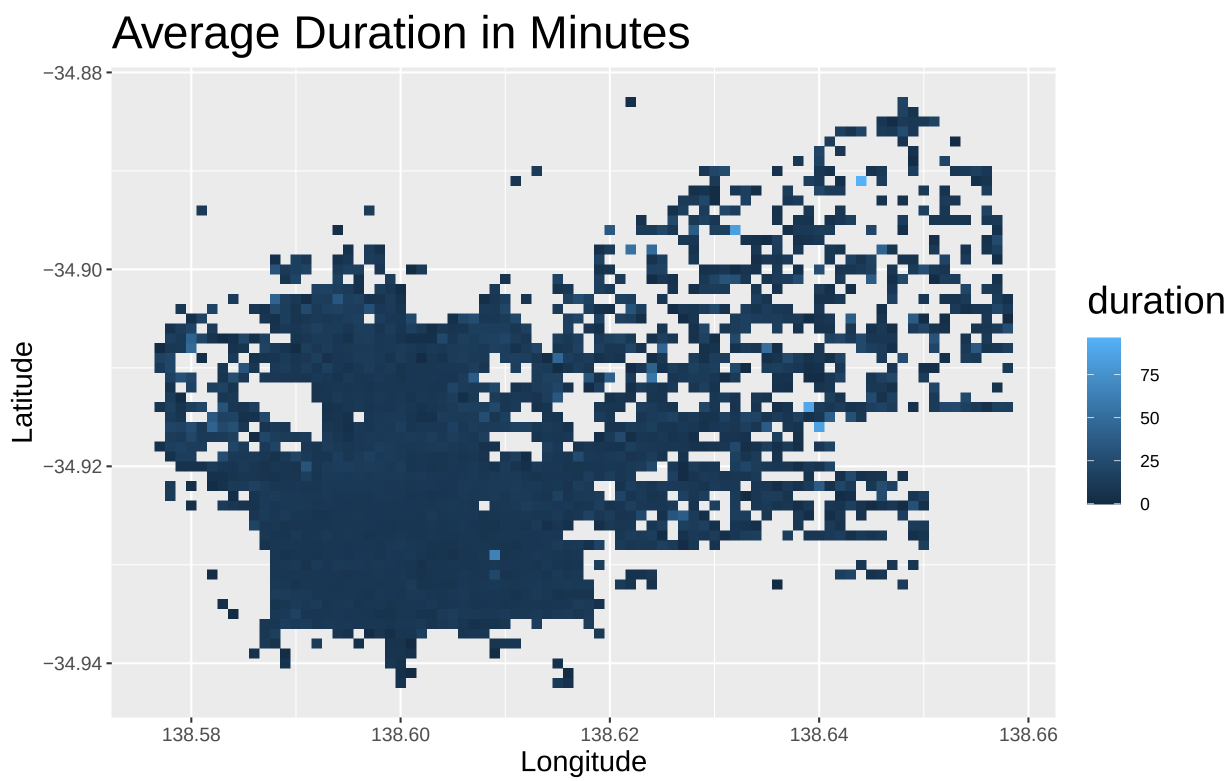


Fig. 2.4: Average Trip Duration in each recorded location.

point in time it seems that, especially in the Norwood region, users are not utilising these new areas. This can be seen in not only the lower demand, but in how the average duration of trips in this region are not higher despite locations in this region being further apart (Figure 2.4).

The Last-Mile problem provides potential implications for research. If users utilise these scooters for their daily commute to work, certain patterns should be present in the data, as trips are repeated. It could be assumed that demand would be higher immediately before and after the common work period (9-5). Also, if a user rides somewhere in the morning, it could be assumed that this was a trip to work, and it would be likely that the reverse of this trip could occur later in the afternoon. However, these types of trips do not occur often enough to produce a pattern.

Figure 2.5 shows that popular destinations for trips almost completely match popular locations for trips to start (Figure 2.3). While one of the most popular locations is the train

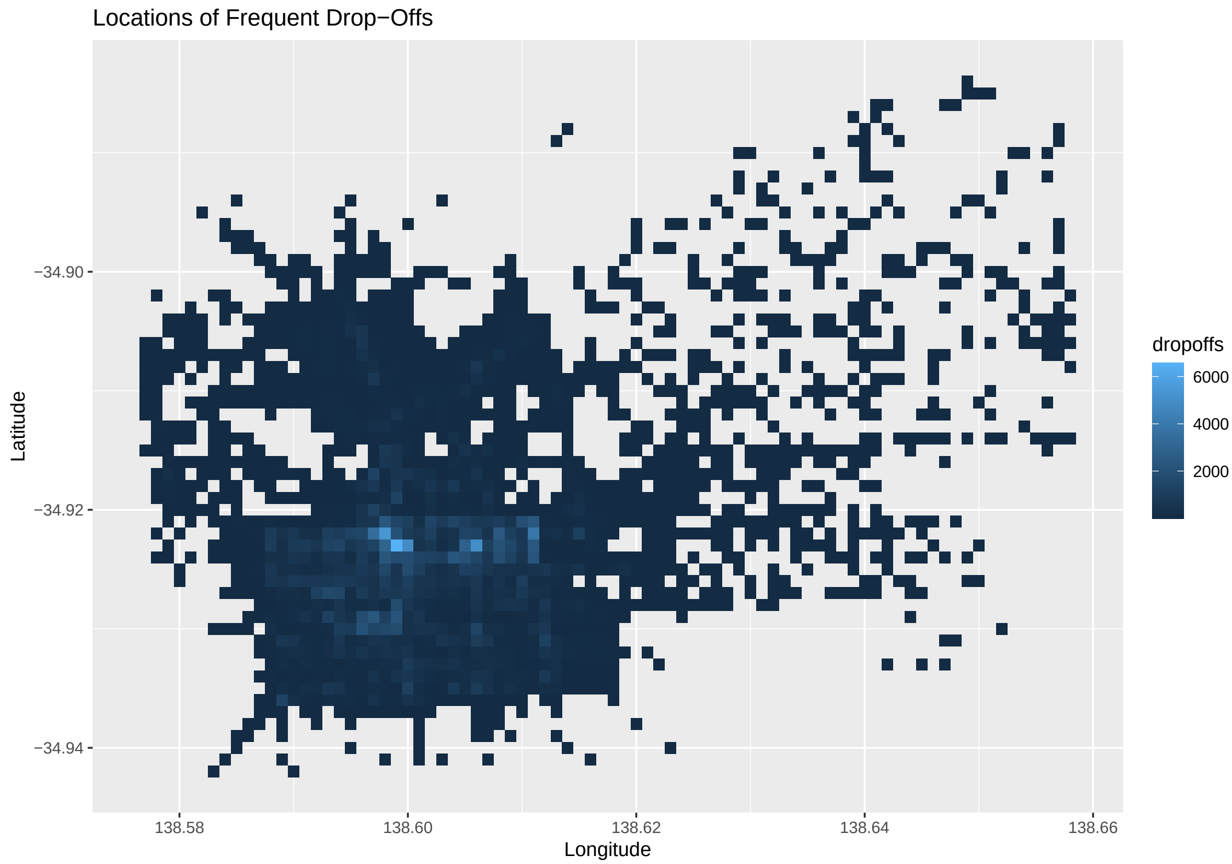


Fig. 2.5: Heat map of dropoffs.

station, which may imply that these scooters solve the public transport definition of the last-mile problem, all other hot-spots are related to commercial or retail areas. It would be expected that if commuting trips were popular, trips would frequently start and end in more work-related areas of the city.

By observing the number of events that have occurred at specific hours during the week in Figure 2.7, information on demand can be ascertained. Trips are much more popular on weekends, which includes Fridays as trips are also much more popular at night (after 5pm). The comparatively few events that occur in the morning on weekdays (before 9am), to the number of events that occur in the middle of the day (9am to 5pm) continue to support the claim that these scooters are used more for leisure than work. While commuting trips may still certainly occur relatively frequently, leisure trips are prime reason scooters are used.

Figure 2.5 shows the patterns of drop-off locations based on where the trips start. In

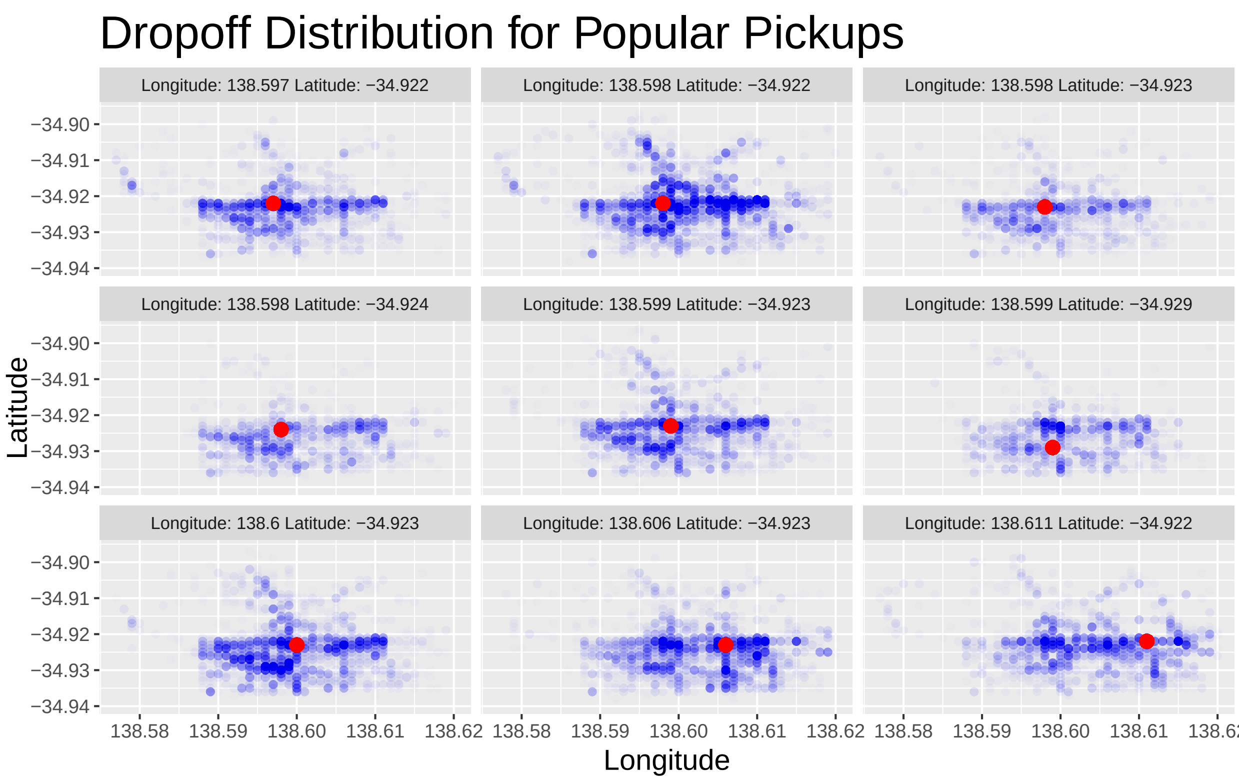


Fig. 2.6: The distribution of trip destinations for given trip origins. Each plot corresponds to one of the top nine most popular places to start trips, identified with a red dot, and blue dots show how commonly trips starting at the red dot have ended there. The darker the blue dot, the more frequently scooters are ridden there. When coordinates are rounded to 0.001 degrees, the top five starting locations are all within a very small area. These are the coordinates of the Adelaide train station. The horizontal line of dark blue dots present in most plots corresponds to North Terrace, the upper border of the Adelaide CBD. It can be seen that scooters are often taken across North Terrace, with some pickup spots at the train station often seeing trips that end in the North Adelaide suburb.

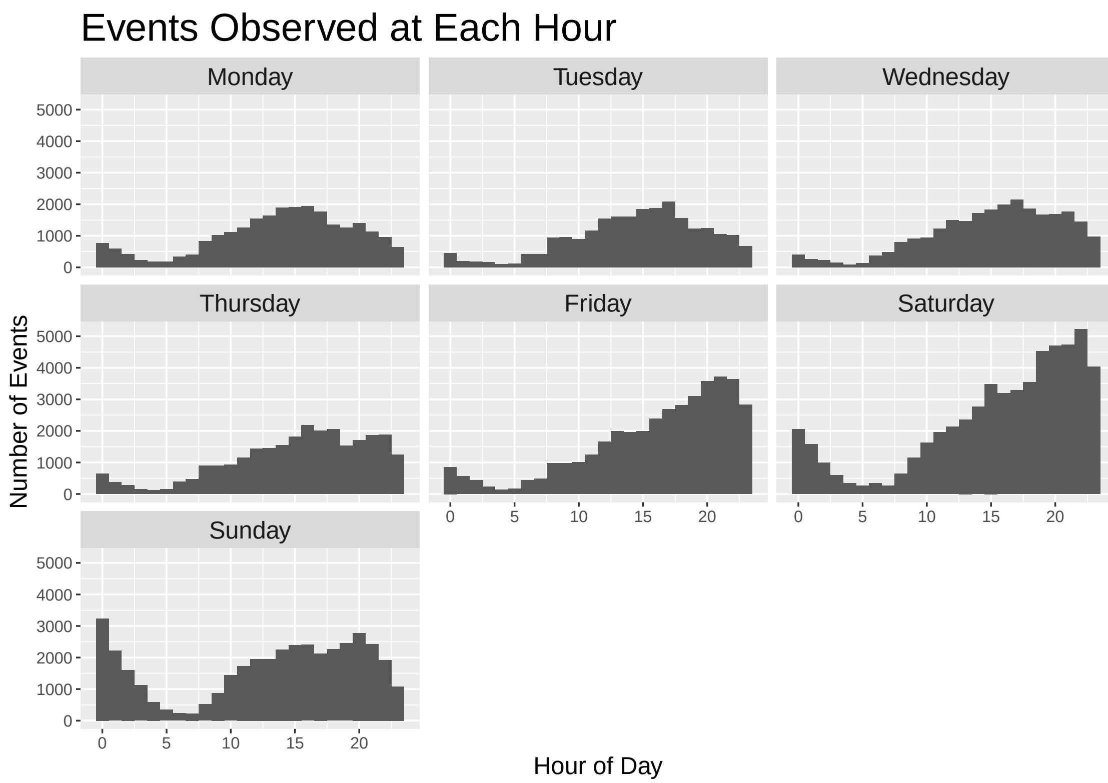


Fig. 2.7: Number of events recorded at each hour of each day over a one year period.

most cases, it appears that trips frequently end a small distance away from where they start, or along North Terrace near the retail areas, represented by the thick horizontal line of blue dots.

However, trips do exist that start at the train station (-34.922,138.598) and end in a variety of locations. Since workplaces are scattered throughout the Adelaide CBD and not concentrated in one specific area, it is possible that patterns such as commutes would be difficult to observe. However, trips most frequently end to the east of this point (the mall), or directly south (the market), further showing that commuting on these scooters is far less common than leisure or non-essential travel.

While this means that assumptions cannot be made based on commute expectations, the patterns observed here and the evidence that scooters are used very frequently for leisure is still useful. It can be expected that trips most commonly start and end in a small number of locations. Demand also appears to heavily depend on time of day and the day of the week, as supported by Figure 2.7. The histograms for Monday, Tuesday, Wednesday and Thursday are all visually similar, but it appears that demand on Fridays is much higher later in the day as people finish work for the week, and Saturdays and Sundays have higher demand during the day.

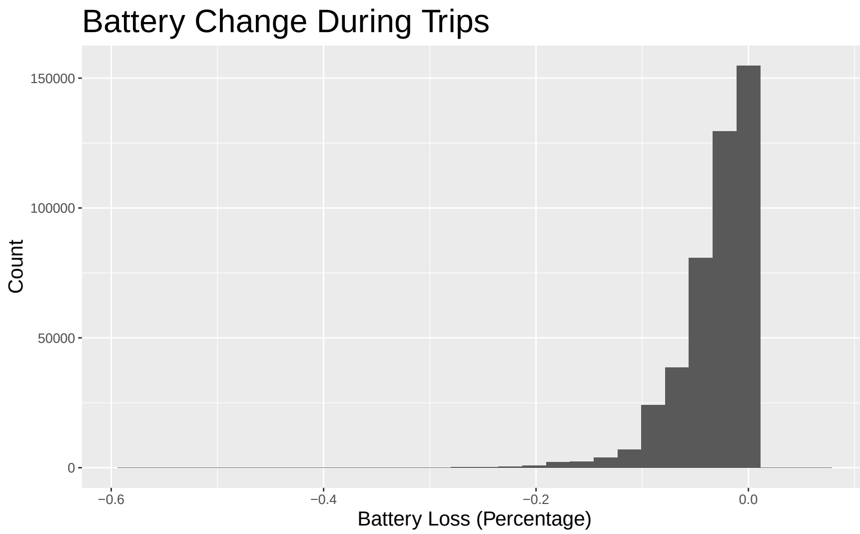
2.2.2 Battery Use

Figure 2.8 shows how batteries have been depleted during trips historically. While there are trips that experience a high battery loss (15-25%), the majority of trips experience less than 5-10% battery loss (Figure 2.8a). It can also be observed that the majority of scooters start trips with more than 40% battery, but more importantly, there are no recordings of trips that occur using a scooter with less than 15% battery (Figure 2.8b). While there are records of trips that end with less than 15%, it is uncommon (Figure 2.8c).

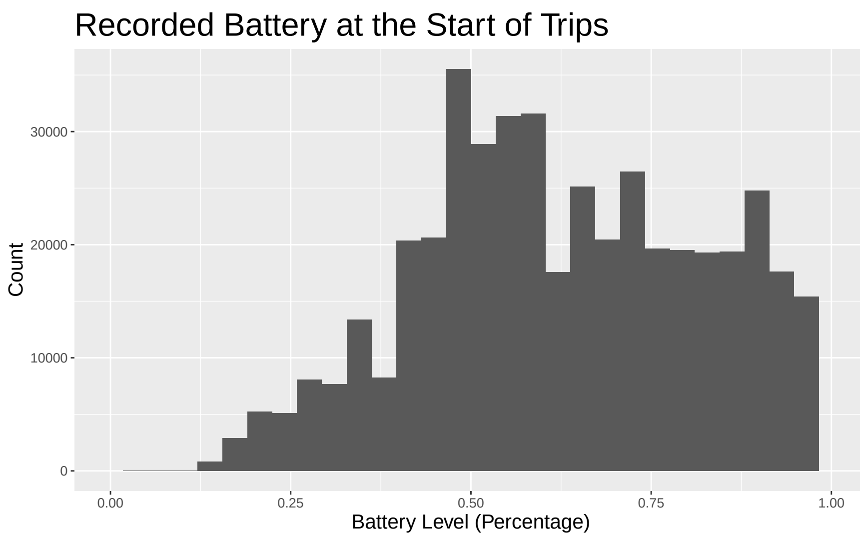
This is important, as the only information provided in the data that is related to battery is given when a trip starts or ends. If no trips start with less than 15% battery, and very few trips end in that state, then there are very few records of scooters with too little battery to perform a trip, as the majority of trips use much less than 15%. It should be expected that scooters exist that are so low on battery that they require charging before a trip is performed, however it is possible that users avoid these. As such, they are effectively non-existent in the dataset as no trips are performed with them, and there is no data on scooters that need their batteries replaced or when they have been replaced.

As can be expected with technology that is used frequently over long periods of time and performs intense actions, there are some issues with the accuracy of the data that is recorded. This is especially true for the batteries, as the recording is a percentage of the battery's maximum potential charge which can decrease as the battery ages, and the observed available energy can depend on the current voltage of the system[1]. The latter means that sometimes a scooter can end a trip with a higher recorded charge than when it started the

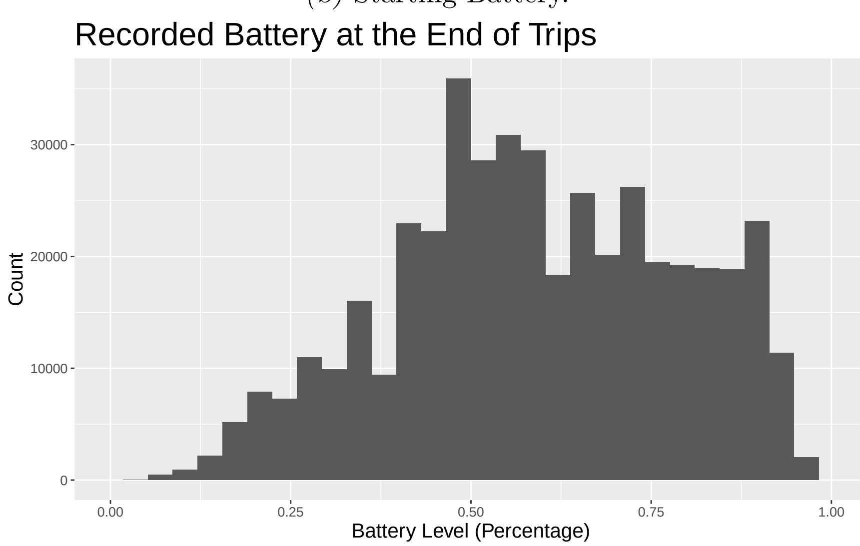
[1] This information, along with other Beam-related information, was ascertained through personal communication with Beam employees, organised throughout 2021 to better understand the data.



(a) Battery Change.

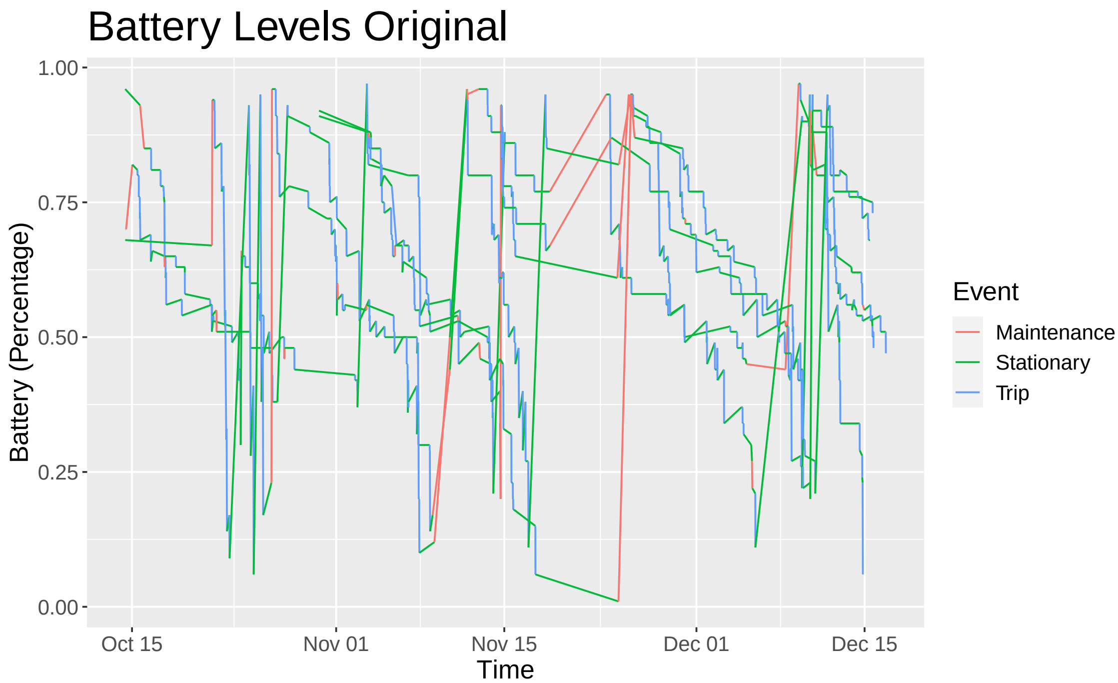


(b) Starting Battery.

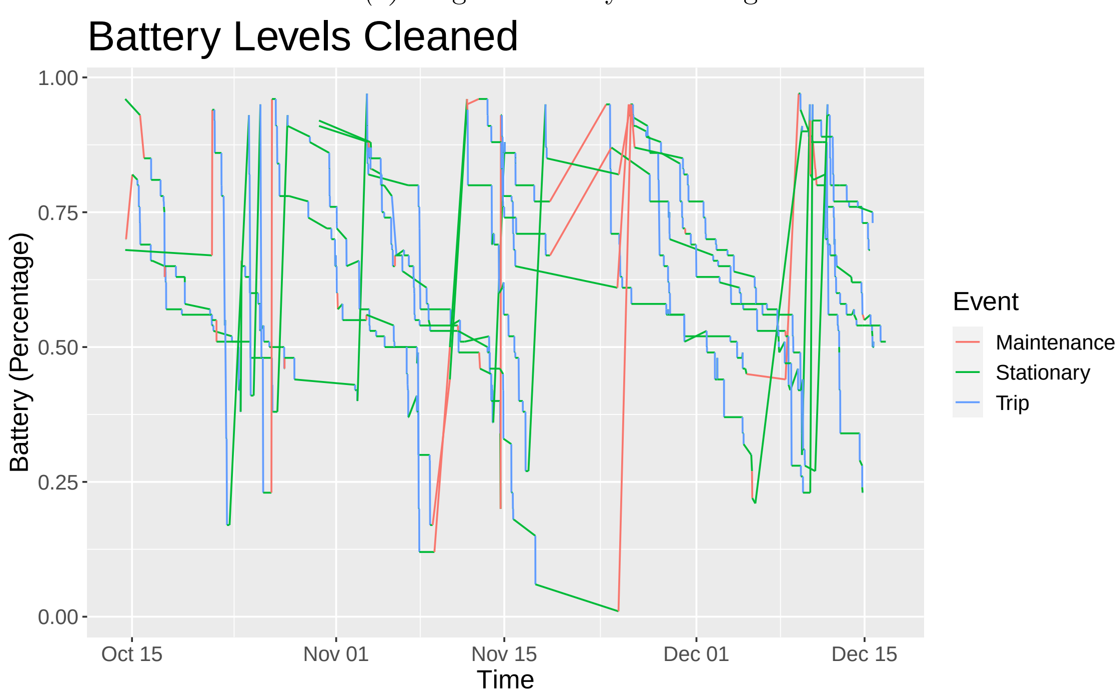


(c) Ending Battery.

Fig. 2.8: Battery Patterns During Trips.



(a) Original Battery Recording.



(b) Modified Battery Recording.

Fig. 2.9: Battery Recordings of five individual scooters over two months.

trip, which is physically impossible, but could also end a trip with a lower recorded charge than it actually has. Despite the true values of battery level being difficult to ascertain, it is possible to clean the data so that the maximum battery change during a trip is 0%.

Figure 2.9 shows a method to clean the dataset that reduces the effects of this behaviour. There are five lines representing individual scooters, and the colour of the line represents what is happening to the scooter at that time. The blue parts represent battery changes during trips, while the green parts describe how the battery behaves while the scooter is stationary. Most important to note is points where the line is purple and drops, but then gradually increases while the line is blue. This is the result of a scooter's high voltage causing the battery percentage recording to be lower at the time of a trip ending than what it actually is. As such, some trips will be recorded as intense on the battery with a shorter duration or length than what would be expected, while other scooters may somehow recharge while stationary.

The data cleaning method that creates the results in Figure 2.9b compares the battery percentage at drop-off to that of the next event. If the battery increases in this time, it is assumed that this is the result of an inaccurate recording, and the drop off percentage is set to the next percentage (when the scooter is used next), so that the change in battery while stationary is 0, instead of positive. In order to take battery replacements into account, the percentage at pick-up is compared to the percentage after the drop-off. If the battery percentage is larger at the start of the next event, it is assumed that the battery was replaced since the drop-off. The current percentage in this case is unchanged. If battery percentage increases during a trip, the drop-off percentage is changed to be equal to the pick-up percentage.

For an event i such that x_i is the battery percentage at a drop-off, we have

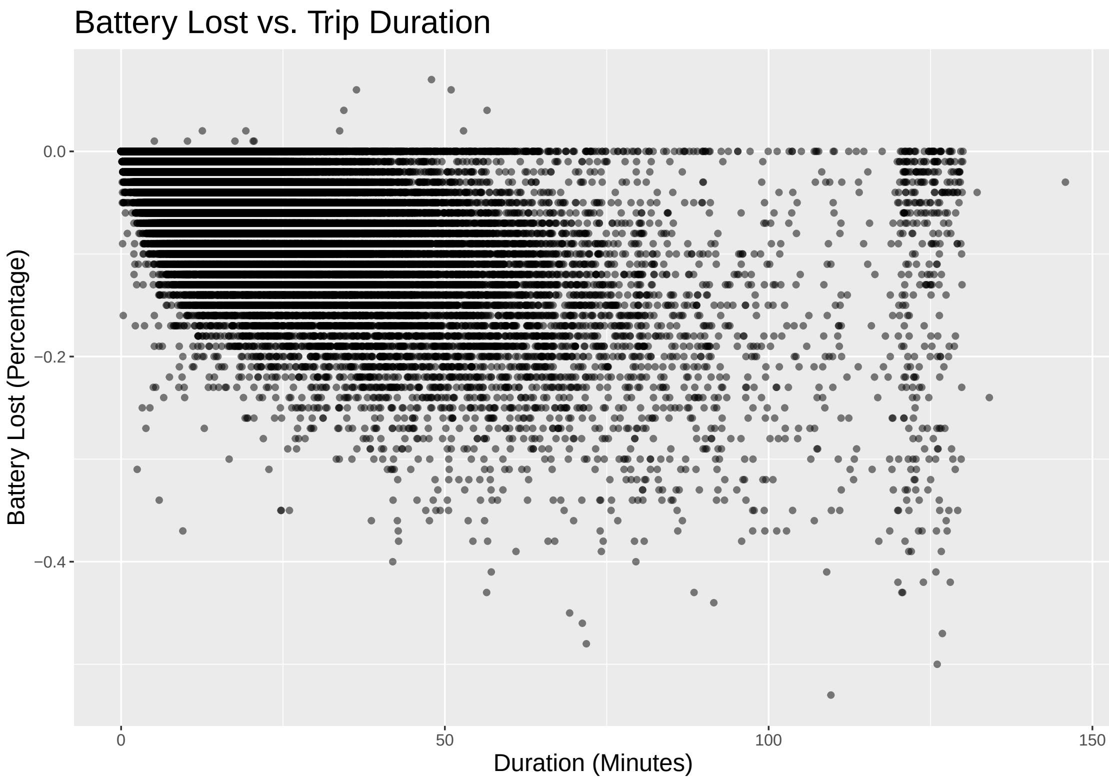


Fig. 2.10: Battery lost during trips.

$$x_i = \begin{cases} x_{i-1} & \text{if } x_i > x_{i-1} \\ x_{i+1} & \text{if } x_i < x_{i+1} < x_{i-1} \\ x_i & \text{otherwise.} \end{cases}$$

This method will not remove every instance of these issues, and this can be seen in Figure 2.9b, as there are still some stationary points where the battery increases slightly. The method is simple, but is able to take into account the fact that battery percentages do not usually increase to 100% when replaced, they just increase by a significant amount. This is also done without having to set a threshold for battery increase before it is considered a replacement. However, inaccurate recordings as a result of complicated scenarios, such as a number of consecutive trips with low battery use, may still be present.

Battery Use In Context

For the purpose of relating battery use to revenue, it is useful to model how much battery is lost over time during a trip. This way, a scooter's current battery level could be used to approximate the "time remaining" for riding. Then, using how much scooters cost to ride per minute, an approximation could be developed of how much revenue a scooter can generate given its current battery percentage. However, Figure 2.10 implies that this would be difficult, as while a higher battery loss implies a higher duration, a higher duration does not seem to imply higher battery loss. This is to be expected, as a scooter could be hired and then remain stationary. While distance most likely has a greater impact on battery loss, the rate of battery used per metre will also vary depending on how fast the scooter is going. As such, a model that takes into account both duration and distance would be more accurate than duration alone.

Other abnormalities in the data include trips experiencing an increase in battery. As this seems to be limited to short trips, it is most likely to be a result of unreliable battery recordings, such as if the battery level is unchanged and the previous recording was lower than it should have been. This should not be possible, so removing this data is valid.

There are also a number of trips with zero battery change, with a variety of different durations. This is shown by the large number of trips in Figure 2.10 that have 0% battery loss but have a duration between 0 and 125 minutes. This could be due to users starting a trip and then forgetting to stop the trip before deciding not to ride, with the trip only ending when they remember to press stop much later. However, it would also be expected that if there were a pattern of users forgetting to end a trip, then there would also be examples of users riding a scooter and then leaving the scooter without ending the trip. Since this extended duration only seems to occur at zero battery change, it is much more likely that it is a result of trips failing to start. Trips fail to start for a number of reasons, if the helmet doesn't unlock, if the scooter isn't responding, or if it is out of battery, but potentially the system records that a trip has begun and the trip is only recorded as finished when an employee is able to fix the problem.

There is also a vertical band in Figure 2.10 of trips with high duration (> 120 minutes) that don't follow the observed pattern of battery loss. Points not in this band have a duration of less than 100 minutes, setting them visually apart, but in terms of battery loss these trips are within the same range. This could be a result of users forgetting to end a trip, and the Beam system may forcefully end trips if the scooter has not moved for a certain duration and the trip time reaches 2 hours. As such, the true duration of the trip is unknown, but could be ascertained from the route data by finding the time that the scooter stops moving.

The trip data-set is cleaned by removing trips with non-negative battery loss, and with a duration higher than 120 minutes. While this will remove short trips that don't have much impact on battery, these trips are not relevant to this research anyway as they don't have an effect on battery replacement. Trips in the vertical band could be kept as mentioned above, however the dataset is still large (245,627 trips) after removal. This clean data (Figure 2.11) more easily demonstrates that trips longer than 50 minutes have a higher average loss in battery, however this pattern becomes less apparent for trips that use higher than 20 percent battery.

2.3 Zoning

This research assumes that a scooter's location will provide information on potential demand and trip duration/intensity, which can then be used in conjunction with a scooter's remaining battery to determine if replacement is a priority. For this to be done, locations need to be divided into a number of "zones" so scooters can be classified by location.

Location data is given as 2-dimensional coordinates, longitude and latitude. While this is highly accurate, using continuous data creates problems as there are potentially infinitely many locations. While there are methods to approximate both duration and demand densities using continuous data, such as parametrisation that uses location as a variable, this adds to the complexity of the approximations. Also, the battery usage estimates developed

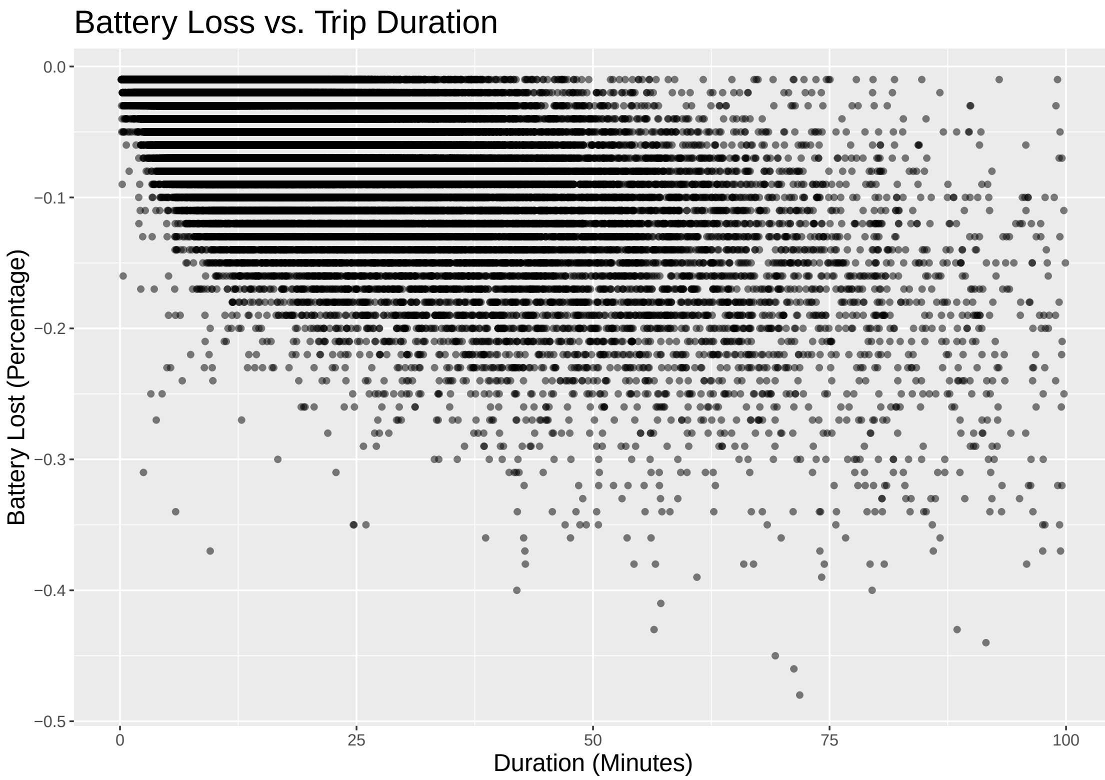


Fig. 2.11: Battery lost during trips (Cleaned).

in Chapter 3, which provide an estimate for the types of trips a scooter can satisfy given current battery level, depend on location as well. In this case, treating location as a continuous variable will imply that slight changes in location would have an effect. Especially in a central city location, it is more likely that distinct 'zones' exist where user patterns are consistent throughout. Hence, the location data is hence classified into discrete regions. Demand is also calculated for each individual zone, however it would still be possible for it to be estimated continuously.

Zones need to be small enough that individual patterns are captured, but not so small that there are an unnecessarily high number of them and the average distance between scooters in neighbouring zones is not too small.

For the purpose of minimising the number of zones, the space considered in this research is halved. This removes the eastern portion of usable area, which contains the Norwood/-Payneham section. This decision comes from both the area's large size and lack of observed popularity with users.

Three potential methods for classifying were considered:

1. Grid the location space by simply rounding the longitude and latitude values to the nearest decimal place;

2. Use a predefined clustering method to define centroids based on areas of high density and classifying points by their closest centroid; and,

3. Define a heuristic that decides zone shapes using boundaries adjusted to observed patterns and required outcome (matching locations with similar patterns, optimising predictability *et cetera*.).

While the heuristic provides potential future extensions to this research, a clustering method was chosen as there are packages that provide effective implementations of them, and its performance was compared to finer and coarser rounding.

2.3.1 Effectiveness Metric

In order to determine how effective a zoning method is, how effectiveness is measured needs to be decided. Effectiveness should be defined in a way that represents how useful it is to the research. In this case, we define the effectiveness of a method by how well it separates patterns such as trip duration and battery use.

An effective zoning method will classify locations into as few zones as possible, while also ensuring that areas with significantly different trip patterns aren't collected into the same classification. This provides zones with as much data as possible, with little loss of information.

Figure 2.12 provides the approximated density functions for the nine most popular locations. These are achieved with a KDE, described in Section 2.3.2. The peaks represent the expected value for revenue from a trip in that location. Trips can commonly be expected

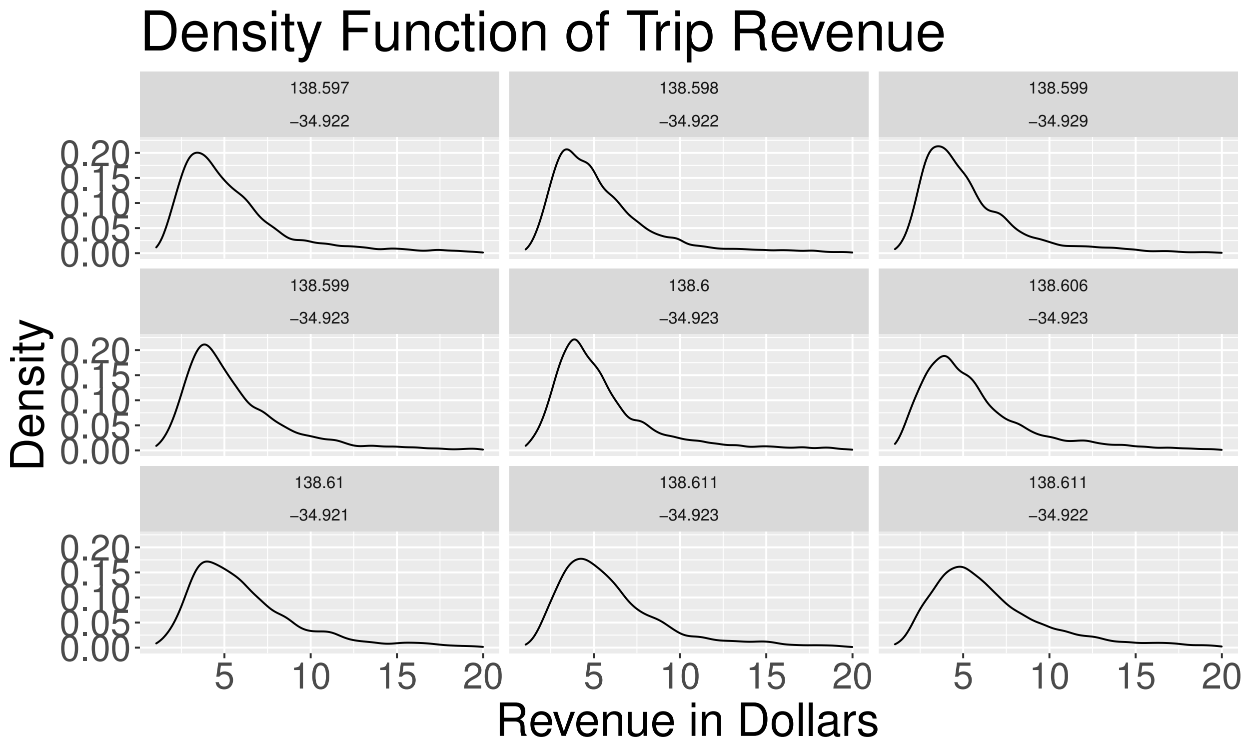


Fig. 2.12: Density function for the revenue of trips in 9 different locations.

to generate around $5 of revenue, however some locations do have a slightly lower expected revenue, suggesting that trips that start in these locations generate less revenue on average.

Some locations have slightly lower peaks than others, but the right tails are slightly wider. This indicates that while the expected value is the same, these locations have a slightly higher chance of generating above average revenue from trips that start in these locations. These difference are small, however, and as such statistical evidence is needed to suggest that these densities are in fact unique. This statistical evidence can also be used to measure the effectiveness of zoning.

2.3.2 KDE Comparison

The revenue densities are calculated using a kernel density estimate (KDE), which is able to approximate an unknown univariate density function using a set of values, $(x_1, x_2, ..., x_n)$, assumed to be independently distributed from the unknown density. This is achieved by selecting a non-negative kernel function, $K(x)$ and a positive smoothing bandwidth parameter, h. We estimate value of the density, f, at point x, with:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

A Gaussian density was chosen as the kernel. Since this density has mean zero, in practice the KDE creates a small Gaussian density centred at each point x_i, with a standard deviation selected automatically by the R stats package [17], and sums the shape of these small densities to form the shape of the overall density $f(x)$.

It is preferred that a classification method is used that divides that scooter space into zones that are distinct in terms of their revenue patterns. Integrated square error (ISE) is used as a metric to quantify the difference in shape between the estimates of two revenue densities.

We take n revenue samples x_i from trips starting in zones j and k:

$$\mathbf{x}_j = (x_{j,1}, x_{j,2}, ..., x_{j,n}) \quad \text{and} \quad \mathbf{x}_k = (x_{k,1}, x_{k,2}, ..., x_{k,n}).$$

We then calculate the densities:

$$f_j(x) \text{ from } \mathbf{x}_j \quad \text{and} \quad f_k(x) \text{ from } \mathbf{x}_k.$$

The ISE of the two densities is then:

$$e_{j,k} = \int_{-\infty}^{\infty} \left(f_j(x) - f_k(x)\right)^2 dx.$$

A difference in density at revenue x results in $f_j(x) - f_k(x)$ to be non-zero. As such, the more dissimilarities that the revenue densities for zones j and k contain, the larger $e_{j,k}$ will be.

We calculate the overall ISE for each zone j by:

$$e_j = \sum_{k=1}^{m} e_{j,k},$$

where m is the number of zones that the classification method generates.

Then, the mean ISE for chosen method α for comparison is:

$$I_\alpha = \frac{1}{m} \sum_{j=1}^{m} e_j.$$

We use a higher I_α as a measure of zone quality.

2.3.3 Methods

Apart from simple gridding, two classification methods were considered:

1. K-means Clustering: and,

2. Mean Shift Clustering.

Both methods are examples of clustering, which are concerned with finding "centroids" which relate to the centres of areas that are all classified under the same label. This assumes that points with the same label will be always close in value, which is useful for location-based classification as it will not result in zones that have more than one location. When applied to trip data, these methods attempt to fit centroids to areas of high density. The centroids converge towards the coordinates of peaks in the demand density, as seen in Figure 2.13.

K-means clustering creates k centroids in random locations and sets the label of each point in the dataset to the label of the closest centroid to that point. Then, the location of the centroid is set as the mean value of all points with the same label. The labels for all points are also reset based on the new locations of the centroids. This is done iteratively until there are no changes in centroid location. The assumption is that there are k zones,

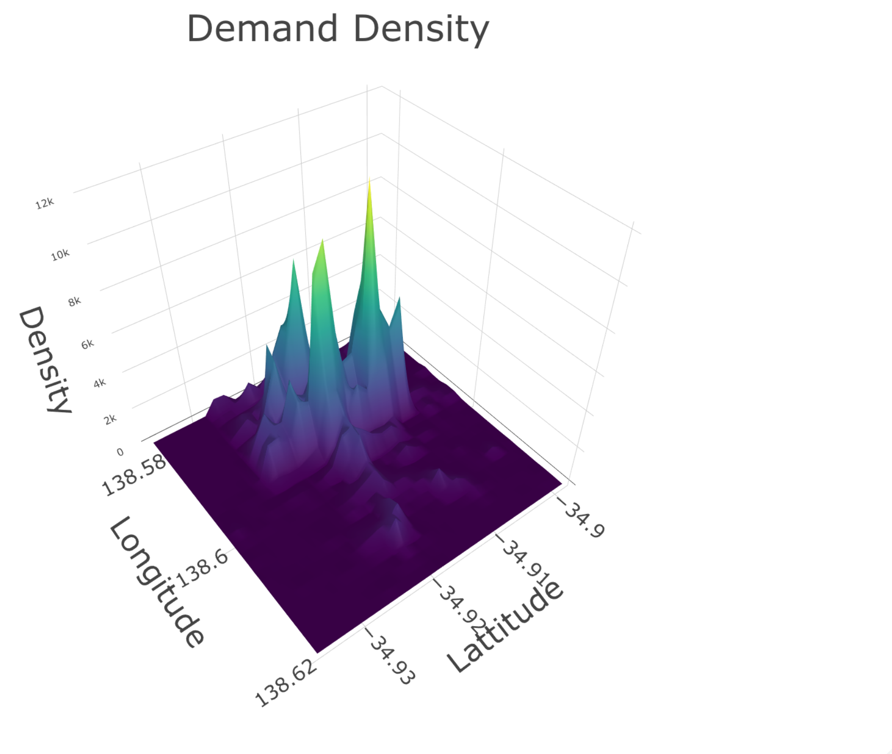


Fig. 2.13: Clustering methods will attempt to fit centroids to these peaks.

and they are all centred around an area of high density. If this is true, the algorithm will converge to the point where the final locations of all centroids reflect this pattern.

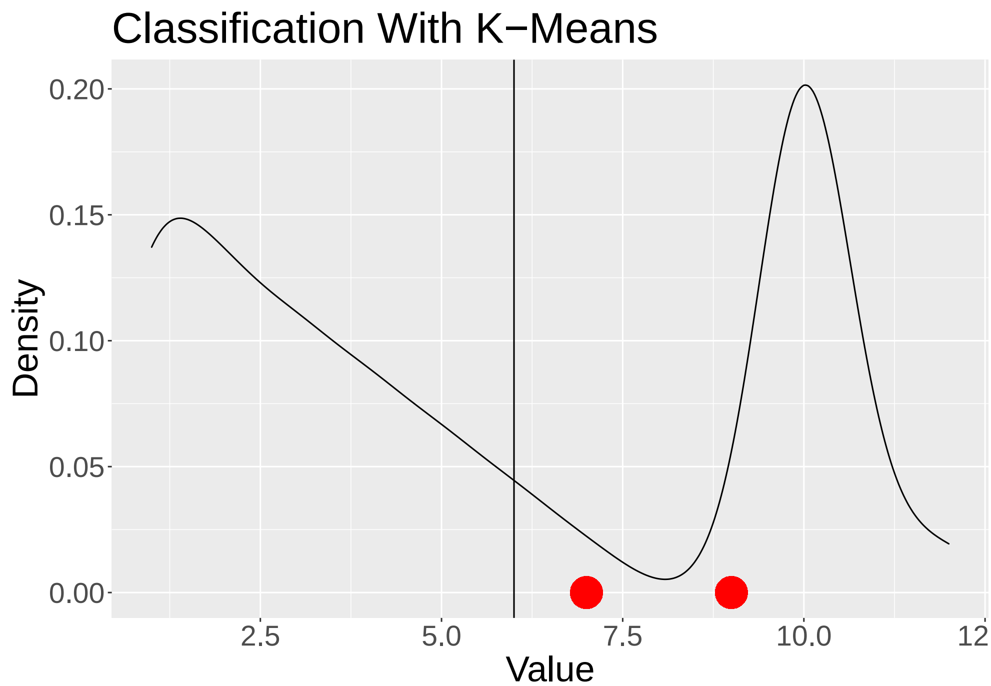
Mean shift clustering first fits a density function to the entire space, and attempts to find the maxima of that density. In this case, a 2-dimensional density is fit to the coordinates of start locations for trips. Mean shift then assigns every point as its own centroid, but shifts these centroids towards the maxima using gradient descent, where the cost function to maximise is the likelihood of a trip starting at those coordinates, to calculate the direction of shift. This is performed iteratively, and at steps where centroids shift to the same value, they are combined into the same label. When the iteration converges, the number of centroids remaining should reflect the number of maxima in the density.

A key difference between K-means and mean shift is that the number of centroids needs to be pre-defined for K-means. A benefit to this is that the ISE can be observed for a different number of zones generated from K-means, but mean shift will always provide what it calculates as the optimal number of zones.

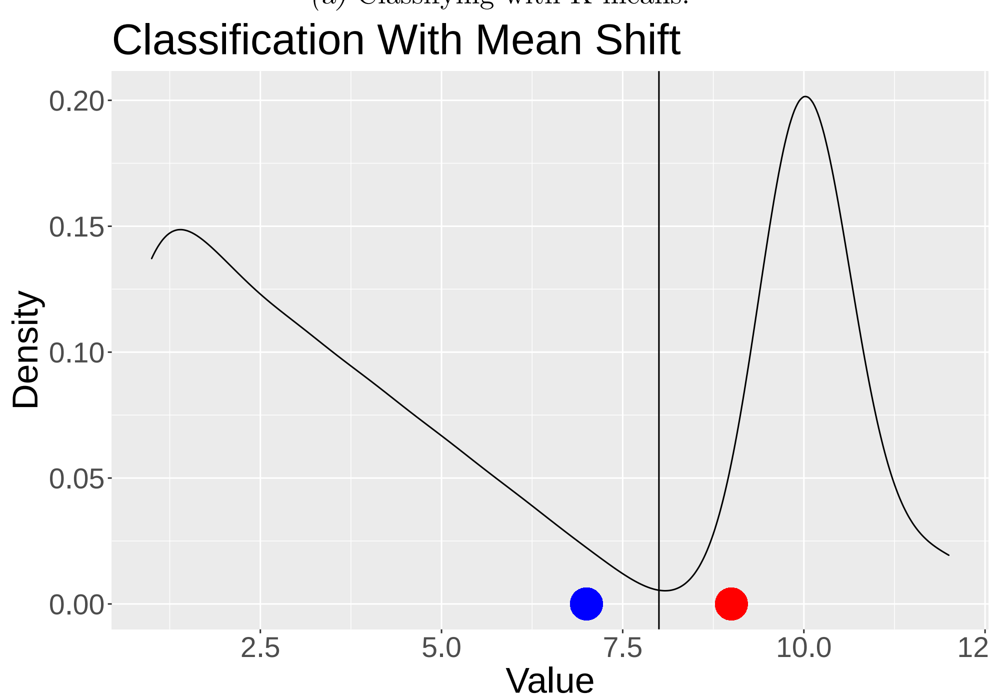
Since real-life data does not usually contain the simple patterns that K-means assume exist, and because the number of centroids is usually unknown, centroids can converge differently based on the value of k but also the starting positions of the k centroids. As such, the results of K-means can change with each execution of the algorithm, unless initial locations are chosen based off of insight. Mean shift on the other hand, will provide the exact same results on each execution on the same dataset.

More differences arise when the context is considered. Popular areas will have a high density as there are many more observed events in those areas and as such the scooter space contains around four very small areas of extremely high density with the surrounding areas having a consistent, low density. These four areas correspond to the four areas previously observed to be the most popular, the train station, the two ends of Rundle Mall, and Victoria Square. Density is an important factor for both methods, as K-means is likely to seek areas of high density as centroids in those locations will have the smallest distance from the largest number of points, and mean shift will seek areas with the most frequent occurrences. Depending on the value of k, K-means may either create too many zones in the vicinity of these areas to account for the higher density, or create centroids that are distant from each other. However, this results in these high density locations having an area much larger than they should, as labels are based on distance and not density. Mean shift may avoid this issue, as the gradient descent will collect locations based on the shape of the density.

This difference is shown in Figure 2.14. Here, a one-dimensional example is presented, where the two dots are being classified. It is expected that both methods will place a centroid on each of the peaks of the density. The shape of the density has one peak on the right that is very thin, whereas the left peak is wider but shorter. There are many reasons for this to occur in terms of demand, but most likely the right peak corresponds to a small popular area, such as a train station, while the left peak represents a large area where key locations are spread out, such as a business district. In this case, the left dot would be in the larger



(a) Classifying with K-means.



(b) Classifying with mean shift.

Fig. 2.14: How the two methods classify points based on centroid location. K-means (top) will classify the two points as red, which corresponds to the right peak, as K-means only takes into account the distance from the peaks. Mean shift (bottom) however, uses the shape of the distribution to classify the left point as blue, corresponding to the left peak. The vertical line, which represents the border between two zones, is located at the midpoint between the two peaks in the K-means scenario, and in the trough between the peaks in the mean shift scenario.

area, whereas the right dot would be in the small area. However, K-means classifies locations based purely on distance from a centroid, so both dots are classified as belonging to the right zone. Mean shift, however, classifies based on gradient descent, and hence the left dot is "shifted" to the left centroid as that is the direction of the density gradient. In this case, mean shift better reflects patterns that occur in the community.

This, however, does not guarantee that zones will have distinct revenue patterns, as classifications are decided based on location density and revenue density. It is likely that the four popular areas have unique patterns to the rest of the city, as the demand patterns are very distinct. However, whether the rest of the zones generated by these methods have unique revenue patterns depends on how those patterns are distributed, and if this distribution happens to be similar to the way the zoning method makes decisions, unless there is a strong relation between demand and revenue patterns.

ISE comparisons are performed with both mean shift and K-means for varying levels of k, alongside different sized grids, to determine the most effective method in relation to revenue densities.

2.3.4 Observation

We calculate the Mean Total ISE (MTISE) for a classification method by finding the mean of the calculated integrated square error for each zone generated by that method. This is done by using histogram function to approximate the revenue density at a given zone, which is normalised to allow for the comparison of shape of the density between zones. The densities are looped through twice, such that every density array can be compared against every other density array. When these arrays are subtracted from each other and the resultant array is summed and squared, it calculates an approximate result for the ISE.

This ISE calculation is used for all classification methods. Mean shift generates 29 zones for the Adelaide CBD based on past trip patterns (Figure 2.15), so K-means is trained for $k = 29$ for similar comparison (Figure 2.16), alongside $k = 70$ (Figure 2.17) and $k = 100$. To

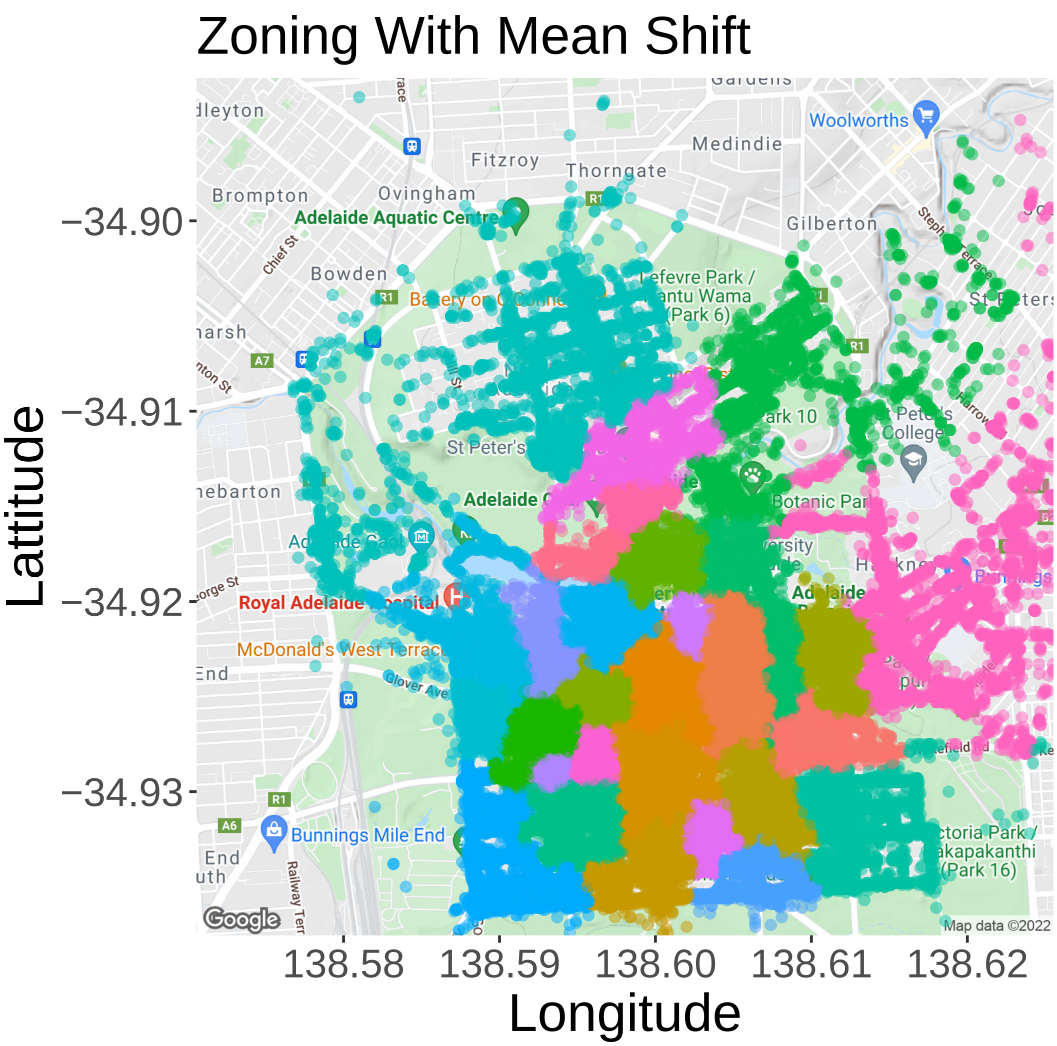


Fig. 2.15: Classification of Points Using Mean Shift Clustering.

create a grid, the 2-dimensional coordinates, provided in degrees of longitude and latitude, are rounded to different decimal places. The chosen rounding values are 0.001, 0.002, 0.004, 0.005, and 0.01 degrees. Due to the nature of longitude/lattitude coordinates, this creates a grid of rectangles instead of squares. However, it is an incredibly simple method of creating a grid.

The mean total ISE of methods is compared against the number of zones that the method creates, as that is the main difference between the methods. However, there are benefits to both having a large number of zones and a small number. Ideally, zones are small enough such that scooters in the same zone can be recharged in one action by an employee, as the number of scooters to recharge in a zone is something that will be observed. If a zone is too large, an employee recharging multiple scooters may be directed to recharge 2 scooters far away from each other, where it may actually be more beneficial and quicker to recharge two scooters that sit on the border of two zones. However if there are too many zones, the number of past trips in the samples of unpopular zones may be small and there might not be enough data to generate an accurate and smooth density estimate. This would result in a higher ISE between zones that doesn't appropriately reflect the true patterns in revenue.

As such, an effective method will have a high mean total ISE, that is not created by a large number of zones, while also having enough zones to effectively capture the region. It is important to note that the decision making process here is hence subjective.

Comparing the mean total ISE for each method as in Figure 2.18 shows the aforementioned patterns. For the grid, using a finer rounding degree causes an initial reduction IMSE as the number of zones increases, however IMSE begins to increase as the number of zones increases past 100. This plot has been developed using a dataset of zones where there is at least 200 samples for each zone, so unpopular zones from certain classification methods are removed assuming that there is not enough data for a reliable result. As such, the effect of many zones is reduced, but does seem to still be present considering how significantly larger the error is on fine grids.

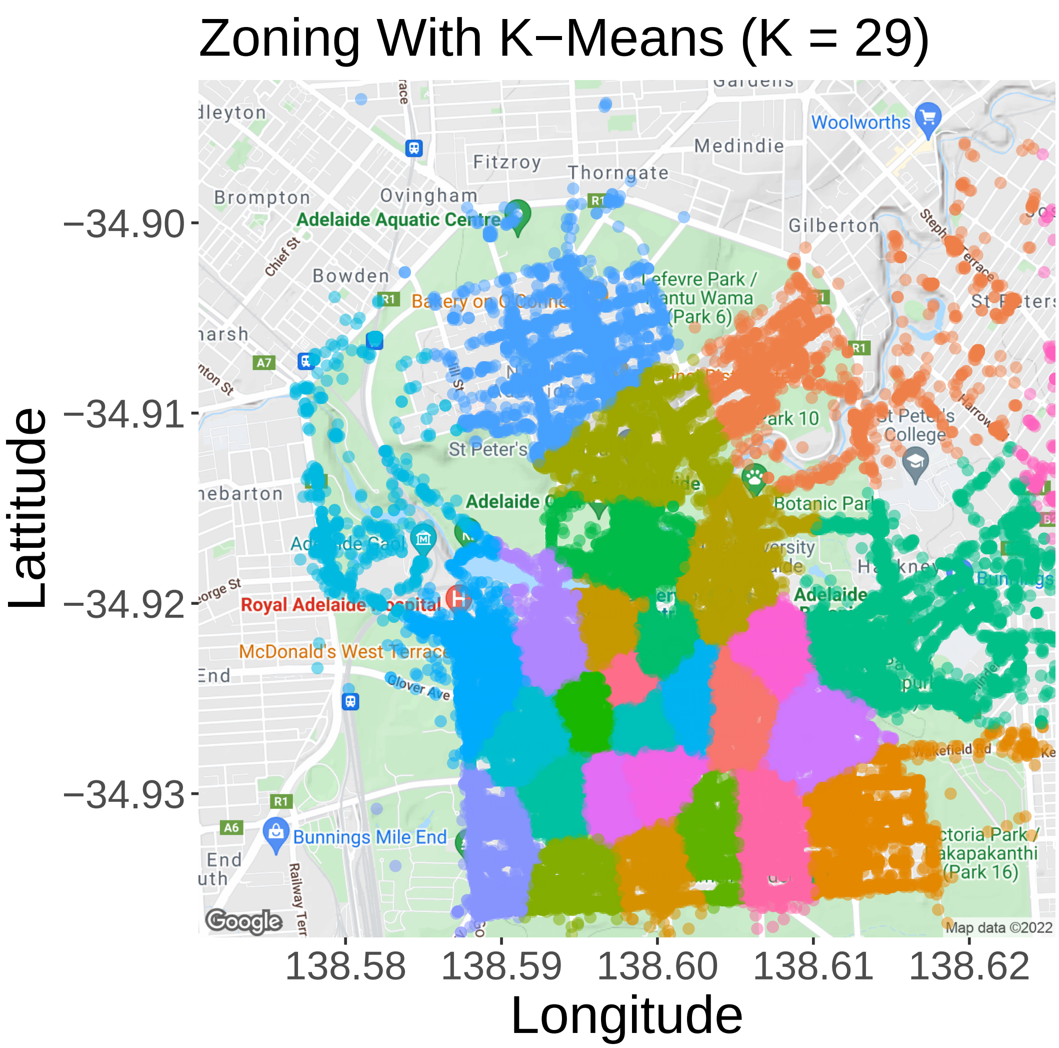


Fig. 2.16: Classification of Points Using K-means with $k = 29$.

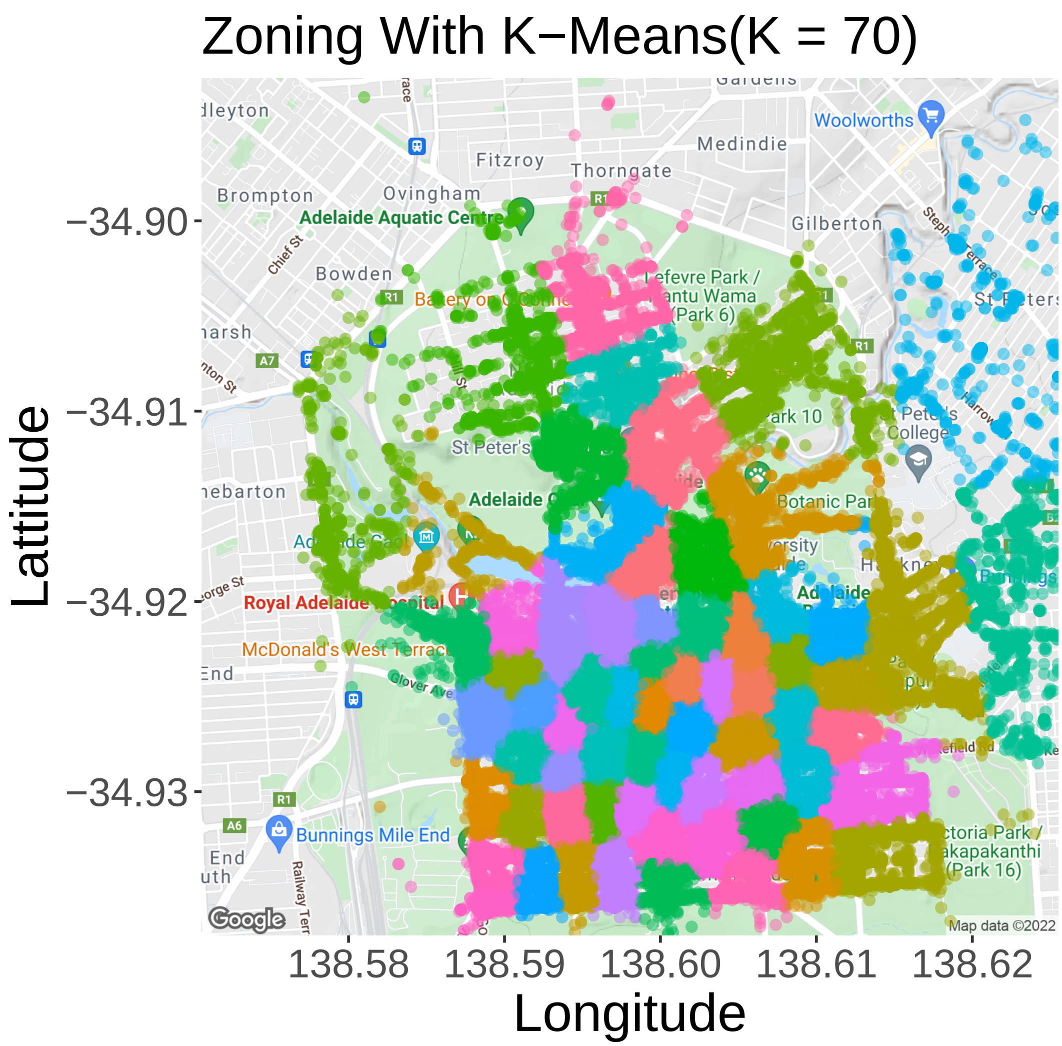


Fig. 2.17: Classification of Points Using K-means with $k = 70$.

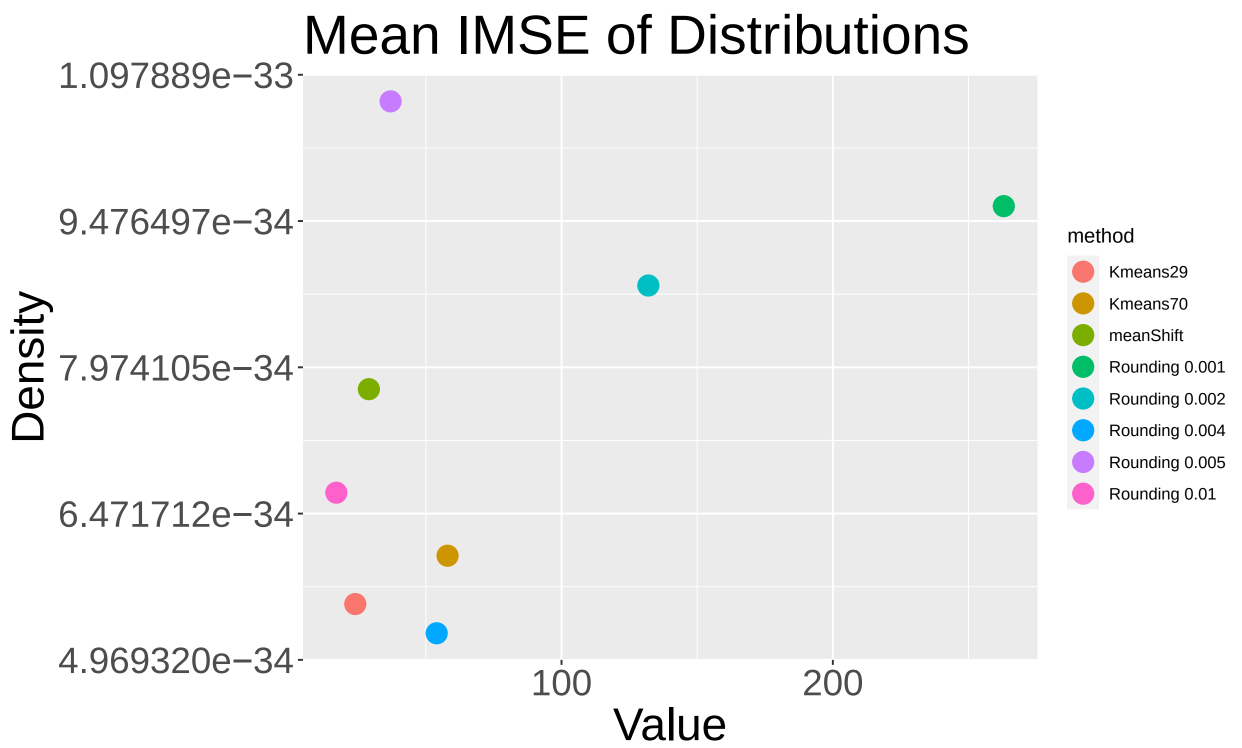


Fig. 2.18: Comparison of mean error between classification methods

With methods that generate less than 100 zones, mean shift appears to consistently out perform K-means. This indicates a potential relationship between demand and revenue patterns, as mean shift is more concerned with separating areas based on likelihood, and hence demand. However, two different datasets were used, one that includes data from 2020, and another from 2021. The 2020 data implies that rounding at 0.005 degrees provides a zoned dataset with the highest error, whereas the 2021 data results in the high error when mean shift is used.

This means that the most preferable classification method depends on the data used, under the assumption that high mean ISE between zones will generate the best results. As such, mean shift and 0.005 degree rounding will both be used, and results gathered in revenue optimisation will be compared against findings here.

2.4 Summary

Enough data has been supplied by Beam in order for battery usage of scooters to be analysed in terms of trip duration and demand. This allows for an expectation of future battery usage to be calculated, provided demand and approximated trip duration in the current area, which is performed in the next chapter. This prediction will be used in the priority metric developed in Chapter 4, which determines individual scooters that are most in need of battery replacement.

This battery usage expectation will be completely informed by the data, and as such the dataset needs to be reliable. In this chapter, the set was cleaned to remove events that were recorded incorrectly, such as trips that experience increases in battery percentage, and events that are not relevant to this research and will only skew approximations, such as trips with high duration and zero battery loss. In terms of duration, two clusters of trips emerged, with the smaller one having a significantly higher average duration. These recordings were removed as they were easy to visually separate from the other cluster, however there is no explanation for why these trips have such a high duration.

A complexity arose in which trips that take the same amount of time may not use the same amount of battery, as users don't always travel at the same speed and batteries of different ages deplete differently. While many of these battery behaviours are difficult to ascertain, it is assumed that some trip duration patterns depend on location as certain types of trips may start in certain areas,

Location data was classified into a finite number of zones to make location-based analysis of trips and demand simpler. As will be seen in Chapter 3, this classification of locations is important for the estimate of battery use per minute, however Chapter 4 shows that methods of demand and battery use are implemented separably, and hence it does not require that demand be estimated in the same way as location. For the sake of simplicity, demand is not defined continuously, however this does present potential future work. Many classification methods were compared, but it was found that simply binning locations into a grid worked best out of the methods considered.

With information on how battery is used each minute in a trip, and locations zoned into finite locations, it is possible to approximate how many 'minutes remaining' of trips a scooter has given its current location.

3. REVENUE INCREASE FROM BATTERY REPLACEMENT

This chapter aims to use the available data to ascertain any connection between predicted revenue and frequency of battery replacement. Battery level provides a restriction on scooter use as a scooter cannot be ridden once this level drops to zero. While there are many factors that influence the rate at which charge is lost, duration is focused on here as it is most related to revenue.

Hence, this chapter develops methods to estimate the duration of trips that can occur when a scooter's battery has a certain amount of charge. Information on a scooter's battery level at the start of a trip could be used to learn what trips are possible given a specific level of battery. However, there is no information on trips that could not occur due to lack of battery level, which would be required for understanding the conditions that result in a trip being 'missed'. Instead, the battery loss of trips is taken, alongside the duration of those trips, and it is assumed that a scooter with a battery percentage lower than that loss cannot satisfy a trip of that duration.

When the distribution for trip duration, and hence revenue, is provided, the expected value of revenue can be calculated. This provides information on the amount of revenue that a scooter's next trip is most likely to generate. Using a replacement threshold based on battery level, the expected revenue before and after replacement can be estimated, and hence shows the expected increase in revenue when replacement is performed.

The expected value of revenue is used to observe the effect of replacement with respect to the likelihood of trips not being feasible given the scooter's current battery percentage. However whether a scooter is actually used depends on demand at its location. Revenue increase is combined with demand in Chapter 4.

As discussed in Chapter 2, approximating battery loss with trip duration is not going to have high reliability. When the distance of a trip is included, a model that approximates battery loss will be more accurate as information such as distance and average speed are more correlated with to battery loss, as a stationary scooter will increase trip duration without providing any extra battery loss.

3.1 Linear Model

A naive method is to train a linear model that predicts trip duration given battery loss. The battery lost during a scooter's next trip is unknown, however by taking a scooter's current battery and applying the estimated linear model, the value obtained represents the "minutes remaining" the scooter has for trips before it runs out of battery. Since most trips use less than 15% battery, this relationship between current battery percentage and remaining minutes is most appropriate for scooters with less than 15% charge. It is assumed that this relationship is linear, and remaining minutes will be estimated using the same model for higher initial battery percentages.

A linear model is estimated as in Figure 3.1. Linear regression minimises the distance between individual data points and a fitted regression line. Due to the sparsity of points in this data the result is a slope that does not seem to follow the general pattern that the data has. By forcing the y-intercept to 0, the resulting gradient is steeper and appears to better reflect how duration may increase with battery use; this is reasonable, as it is impossible for a scooter to execute a trip at 0% battery.

The relationship between trip duration and battery loss is clearly not linear. For a linear model to be accurate, more information would need to be provided. However, since no information presented in the data can be ascertained before a trip occurs (distance, route, user, *et cetera*.), any additional features would provide extra uncertainties for the rest of the research.

The linear model returns a slope coefficient of -257.4722 minutes for each percentage of

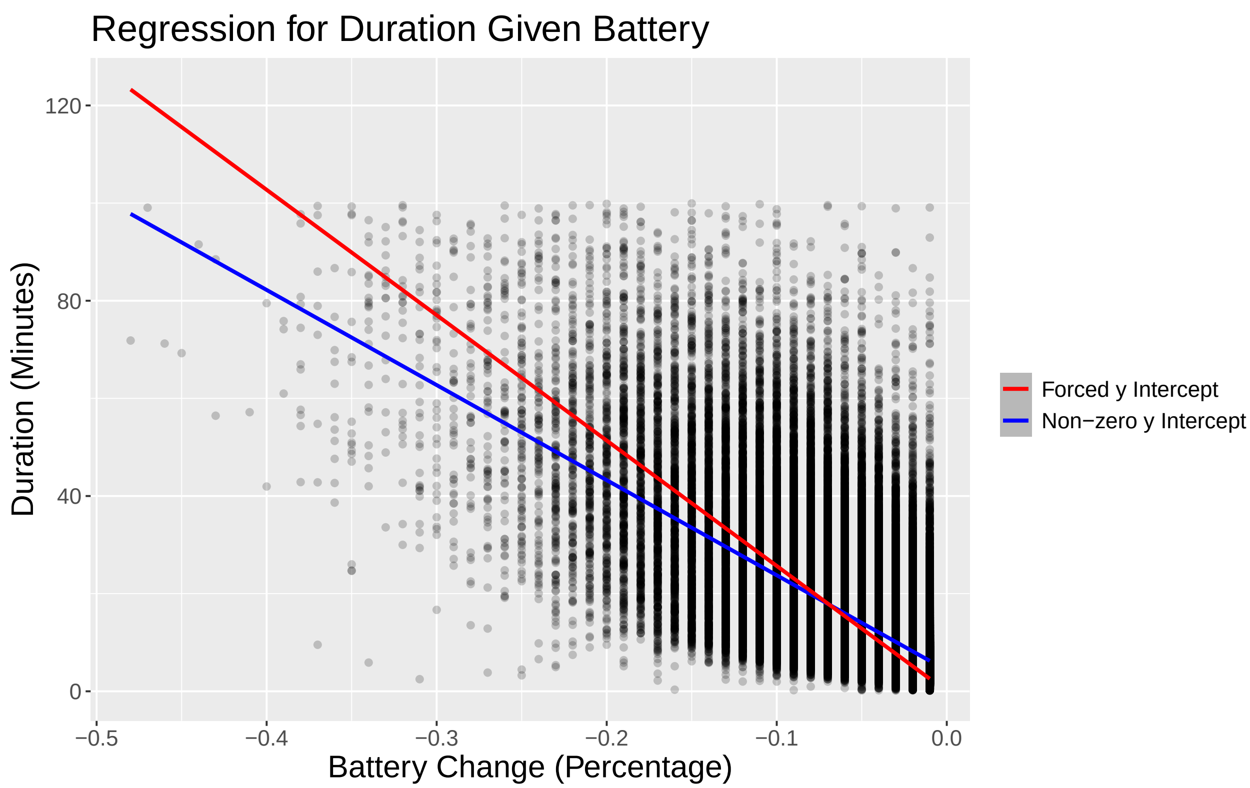


Fig. 3.1: Linear models predicting trip duration from battery lost

battery used. Since battery loss is always negative, but battery level is always positive, the formula for approximating minutes remaining becomes:

$$\text{minutes remaining} = 257.4722 \times \text{battery level}.$$

Scooters generate \$0.38 every minute during a trip, with a starting cost of \$1, so the approximated revenue cap is:

$$\text{rev}_c = 1 + 0.38 \times 257.4722 \times \text{battery level}.$$

The data presents battery percentage as a decimal between 0 and 1, to two decimal places. Hence battery level can be treated as a whole number percentage instead, let batt $\in [0, 100]$:

$$\text{rev}_c = 1 + 0.38 \times 2.57 \times \text{batt} = 1 + 0.98 \times \text{batt}.$$

3.1.1 Method

To calculate the expected increase in revenue of scooter in zone i with battery percentage b, first approximate the time remaining d in minutes as a linear model l_i, trained with the duration of past trips starting in zone i:

$$d = l_i(b).$$

Then the cap on revenue, $cap_{i,b}$ is simply rev_c where the coefficient of batt is estimated with data from trips starting in zone i:

$$cap_{i,b} = 1 + 0.38 \times l_i(b).$$

Then, using revenues generated from trips starting in zone i, approximate the density of trip revenue $f_i(x)$. The revenue of a scooter's next trip is represented as the random variable X_i. The expected revenue from that trip is:

$$E[X_i] = \int_0^{\infty} x f_i(x) dx.$$

The expected revenue bounded by the scooter's battery percentage is:

$$\int_0^{cap_{i,b}} x f_i(x) dx.$$

Hence, the expected revenue increase as a result of battery replacement is:

$$\hat{R}_L = \int_0^{\infty} x f_i(x) dx - \int_0^{cap_{i,b}} x f_i(x) dx = \int_{cap_{i,b}}^{\infty} x f_i(x) dx.$$

3.2 Kernel Density Filtering

Another method that allows for the estimation of effects on potential revenue is to filter the data used for the KDE based on the current battery level of the scooter. In terms of probability, this is similar to calculating the conditional density of revenue, given battery level. Two densities are estimated, one using data from all of the past trips in that location, and another using only data from trips in that location that use less battery than the scooter currently has. The second density is an estimate of the revenue from a trip given the scooter's current battery level, whereas the first is the likelihood once the scooter's battery is replaced. By calculating the expected value for both densities, the difference between the two should represented the expected increase in revenue from battery replacement.

This should result in a very similar value to one that is calculated based on the linear model, but taking into account the full distribution of battery loss. A given trip duration can result in a range of losses, and a given battery level can satisfy trips that are longer than other recorded trips. As such, a rigid limit such as a cap on revenue will not take into account trips with a higher than average duration for that battery use. Common examples of this are trips where a scooter doesn't travel very far in a given time. These trips will generate the same amount of revenue ones where a scooter travels further in the same amount of time, despite using less battery. KDE filtering should capture the likelihood of these trips occurring.

However, it is common for the two densities to be identical. This occurs when the scooter's battery percentage is higher than the battery lost in all trips that have previously started in the scooter's location. Filtering by the scooter's battery percentage returns the same set of data, and the difference between the expected value of both sets is hence zero. A revenue cap may still capture some increase, as it ignores the fact that the scooter has enough charge to satisfy all previously observed trips. The predictive model method **??** might be preferable if there are locations with a small number of samples, as it uses the average use of scooters across all locations, allowing for the possibility that a trip occurs with a higher loss

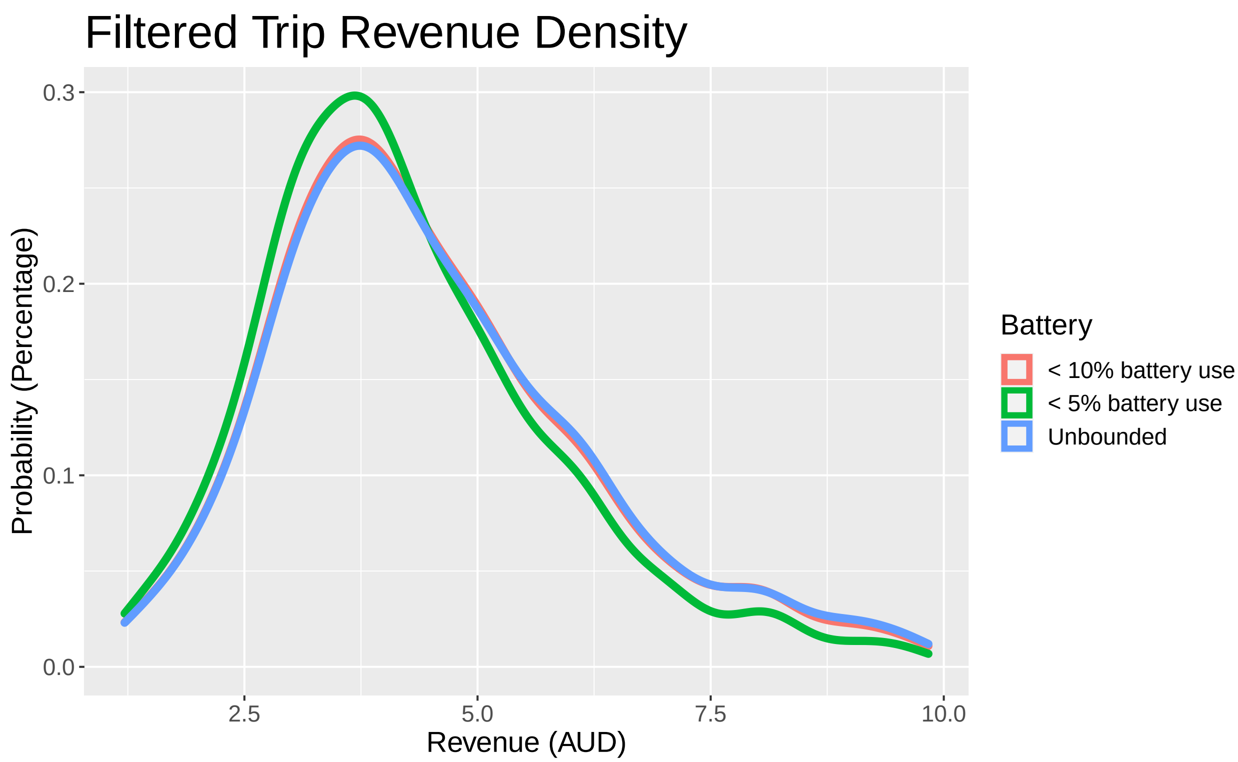


Fig. 3.2: Comparing densities of revenue when they are filtered by battery use

of battery than has been observed up until that time. However, with a large enough number of samples, this problem should not occur often. When sample sizes are smaller, the linear model should result in fewer non-zero increases, whereas the non-zero increases from KDE filtering should be slightly larger.

Figure 3.2 shows how the density of trip revenue changes when only trips that use less than a certain amount of battery percentage are considered. The density is estimated using the revenues of trips that start in a zone containing the Adelaide train station. The trip revenue is ‘bounded‘ by filtering out trips that use more than 10% and 5% battery in the red and green lines respectively.

Bounding causes the peak of the density to be higher, but the right tail is lower. This represents how trips that use less battery percentage will be shorter and generate less revenue, hence high revenue trips are less likely. The peaks all being around $3 implies that the mode, the most likely amount of revenue generated, is the same across all 3 cases. However, the expected value, which takes into account the possibility that trips can generate more revenue than the mode, will be smaller and closer to the mode in the bounded case as the mode is even more likely.

As to be expected, the effect of bounding is more prominent when trips are filtered by a smaller battery percentage (5%). However it is important to note that the effect is almost negligible when filtered by 10%. This is a result of trips rarely using more than 10% battery. This supports the claim made previously that scooters with battery percentage above 10% will have minuscule or even zero predicted revenue increase as the densities are likely to be identical.

3.2.1 Method

For a scooter in zone i with battery percentage b, let X_i be the random variable that represents the revenue in dollars obtained from that scooter's next trip.

From the dataset, let $\mathbf{x}$ be the revenues of all recorded trips, and $\mathbf{y}$ be the corresponding battery losses of those trips. The information for each individual trip is stored as pairs in the form (x_j, y_j). Then, let $(\mathbf{x}, \mathbf{y})_i$ be the subset of those sets that contain information only for trips that start in zone i.

Then, let $(\mathbf{x}, \mathbf{y})_{i,b}$ be a subset of $(\mathbf{x}, \mathbf{y})_i$ for which $y_i \leq b$. Hence $\mathbf{x}_{i,b}$ is a sample of revenues from all trips starting in zone i that use no more than b battery.

Now, let $f_j(x)$ be the function for the density obtained from a KDE using the sample set $\mathbf{x}_i$, and let $f_{i,b}(x)$ be the equivalent but obtained from the sample $\mathbf{x}_{i,b}$. Hence $f_{i,b}(x)$ is the likelihood of the scooter generating x dollars with its current battery percentage, whereas $f_i(x)$ is the likelihood of the scooter generating x dollars when its battery percentage provides no limit on its revenue, or if it was recharged.

If we let $E_\alpha[Z]$ represent the expected value of the random variable z using a density function obtained from subset α, the expected increase of revenue when scooter in zone i

with battery level b is recharged, calculated using KDE filtering, $\hat{R}_K$ is:

$$\hat{R}_K = E_j[X] - E_{i,b}[X].$$

Which can be written as:

$$\hat{R}_K = \int_0^\infty x f_i(x)dx - \int_0^\infty x f_{i,b}(x)dx.$$

Or:

$$\hat{R}_K = \int_0^\infty x\left(f_j(x) - f_{i,b}(x)\right)dx.$$

3.3 Method Comparison

Since both methods approximate the same value, it should follow that:

$$\int_{cap_{i,b}}^\infty x f_i(x)dx \approx \int_0^\infty x\left(f_i(x) - f_{i,b}(x)\right)dx.$$

$f_{i,b}(x)$ is the density function for revenue calculated from trips that use less than b battery. KDE filtering assumes that this density is unbounded, however higher values of x are less likely in this density than in f_i, which is calculating using all trips regardless of battery.

However, when a revenue threshold, the revenue is bounded by $cap_{i,b}$. If this is true, then:

$$\int_0^{cap_{i,b}} x f_i(x)dx = \int_0^\infty x f_{i,b}(x)dx$$

And then the revenue increase estimated using KDE filtering becomes:

$$\hat{R}_K = \int_0^\infty x f_i(x)dx - \int_0^{cap_{i,b}} x f_i(x)dx = \int_{cap_{i,b}}^\infty x f_i(x)dx = \hat{R}_L.$$

Hence, under the assumption made by the linear model method, $\hat{R}_K = \hat{R}_L$.

3.4 Journey Observation

Summarising scooter behaviour into "journeys" is a method of observing revenue changes on a larger scale by collecting trips into groups. By analysing how a scooter is expected to behave in the foreseeable future we can make better decisions on which ones to recharge. Hence, instead of focusing on the next trip a scooter may experience, we focus on the potential for many trips in a short amount of time. Since scooters that are likely to experience multiple trips require more battery than those that might only experience one, while also generating more revenue, the replacement of batteries in these scooters should take priority.

As discussed in Chapter 2, scooters tend to be ridden for very small distances. This means that unless a scooter's battery is extremely low, a scooter will most likely be able to satisfy a users requirements unless that user is planning to ride for a much longer time than the majority of users. Hence, it is expected that the increase in revenue estimated from using either of these methods is very small.

For example, KDE filtering will only return an increase in revenue when the current battery level of a scooter is lower than the maximum loss recorded from trips starting in that location (usually around 15%). Hence there will be no perceived benefit to recharging scooters above that amount.

The average revenue from trips is only $5, and the likelihood of trips generating higher revenue than $10 is incredibly low. $10 is generated from trips that take at least 25 minutes, using the linear model gradient this only requires less than 10% battery on average. $5 is made from 12 minute trips, which use less than 5% battery on average. Hence, the revenue cap will return a small increase in revenue from recharging scooters unless they are very close to running out of battery, and almost no increase if they have more than 10% battery.

However, it is not surprising that revenue increase is small when only the next trip is considered, as the benefits from battery replacement are not immediate. The importance of recharging is not to satisfy the next trip, but to maintain the scooter's availability for the near future, which may involve multiple trips. It is more important that a string of trips is analysed, than a single one.

Analysing this can be complicated, as when a scooter performs a trip it is likely to end in a different location, which is assumed to have its own revenue patterns. Estimating revenue gain means taking account of patterns of both duration and demand to determine trips likely to occur soon after replacement, but if strings of trips are observed, the duration and demand of potential destinations need to be accounted for as well, alongside the likelihood of ending in these destinations. The complexity of these patterns increases exponentially as the number of trips in the string is increased.

One simple way to overcome this, is to take trip strings observed in the data and summarise them individually by summing the important information. Define a "journey" as a scooter's total battery loss and revenue generated from a given time after starting a trip in a given location. Scooters have unique IDs to ensure that each one's behaviour is tracked throughout trips recorded during the time-frame. With enough samples, the distributions of revenue should capture the likelihood that not only the next trip generates high revenue, but also if that trip leads to more high revenue trips. For example, if trips that start in a given location frequently end in an unpopular area, or one where trips are usually short, then the journey revenue density for that area will have a smaller expected value than areas where trips end in popular areas. If locations often result in long chains of trips, where scooters end in popular areas that lead to more popular areas, this is also be reflected in the distributions by having a larger right tail.

Journeys are obtained from the dataset by taking each individual trip, and observing trips that the scooter involved takes for a certain period of time afterwards. Figure 3.3 outlines how journeys are staggered by observing journeys that start in trips that may already be in a previous journey. This allows for a larger sample size. Here, strings of trips are considered in a 24 hour period, and journeys are taken from 4 consecutive trips that used the scooter with ID 4794.

The first journey is taken by considering trips that occur in the 24 hours after the trip

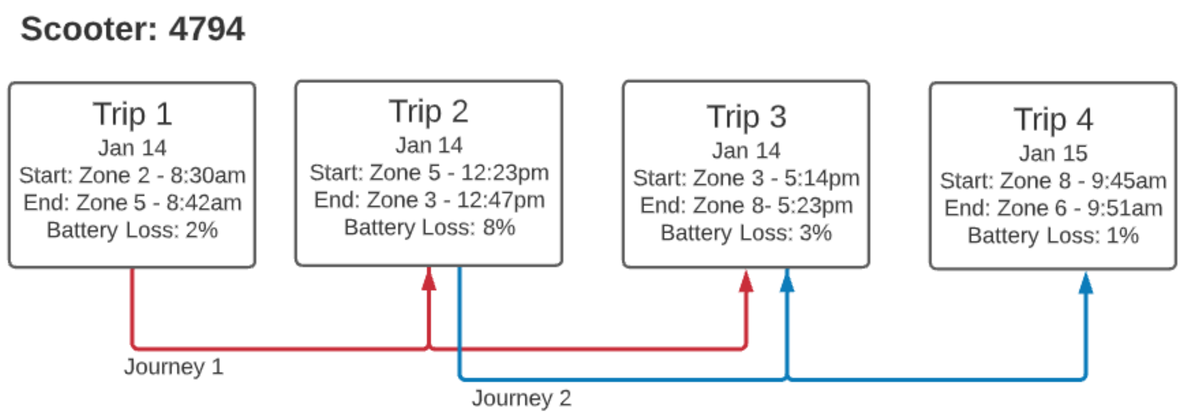


Fig. 3.3: How journeys are gathered from trip data.

currently being considered (Trip 1). Trip 4 occurs more than 24 hours after, and is such not part of Journey 1, but it is part of Journey 2 as it is within 24 hours of Trip 2.

Hence, Journey 1 is a journey that starts in Zone 2 and has a total duration of 45 minutes with a battery loss of 13%, and Journey 2 starts in Zone 5 and has a total duration of 39 minutes with a battery loss of 12%. Journey 1 generated $18.10 in revenue, while Journey 2 generated $15.82.

Once all the journeys are gathered, the scooter IDs are removed and journeys are instead grouped by their starting zone. This results in a dataset of zones which includes their revenue and battery loss patterns over 24 hour periods.

Given a large enough trip dataset, trips can also be grouped by both start zone and time of day to observe expected revenue gain and battery loss for the next period given the current time.

3.5 Discussion

Being able to estimate how much revenue a scooter can be expected to generate with its current battery level is a useful basis for comparing battery replacement priority. With the methods developed in this chapter, the “revenue remaining” for a scooter at its current battery level, and at maximum battery can be approximated. From this, we can estimate the effect replacement has on a scooter's ability to generate profit.

Two methods were developed to achieve this revenue expectation. One, which uses a statistical model to approximate the relation between battery usage and trip duration, the variable that determines trip cost, and another which uses kernel density estimates to compare the likelihood of revenue for trips that use more or less charge than a scooter currently has.

Trip duration directly affects trip revenue, but implies little about the amount of battery used. This is because staying stationary would use very little battery, and hence distance and speed are more useful in predicting usage. However, the model in the first method is trained to predict trip duration from a given battery level, as to provide the required comparison, and including distance in this model would result in a different comparison for each potential distance value for the next trip. Finding the expected revenue would require the integration of a 2-dimensional KDE, making the process significantly more complicated. While this presents possible future work on this problem, it is not included in this thesis. Hence, this method tends to generalise the relationship between battery level and revenue remaining.

This generalisation provides motivation to use the second method, which uses only the distribution of the revenue of previous trips starting in the scooter's current location and takes into account the exact relationship between those revenues and the battery usages. However it was observed that in zones with little data, the results were more unreliable. The statistical model method is able to do more with less data, and hence both methods are considered in the next chapter, where they are used to develop a priority metric.

4. DERIVATION OF THE BATTERY REPLACEMENT LOCATION PRIORITY METRIC

Using the methods from Chapter 3 to predict revenue increase from battery replacement, a metric was developed to compare scooters based on this revenue increase and demand. Scooters were ranked on both revenue generated and the likelihood this revenue is generated in the near future. Higher revenue trips are likely to deplete the battery more, and as such a scooter that is likely to generate high revenue but has a low battery level will be ranked higher than one that is fully charged, as it is in more need of a battery replacement.

The metric requires that the space containing scooters is divided into a finite number of zones as per Chapter 2. By separating data into zones, the metric equation then uses individual revenue and demand patterns to calculate a value representing priority in terms of the impact a replacement in that zone will have on overall revenue. The metric also takes into account the number of scooters in each zone, as available supply in the surrounding area will determine the priority for specific scooters to be recharged.

4.1 Assumptions

It is assumed that when a scooter's battery is replaced, its battery percentage increases to exactly 100%. This is not always the case, as batteries are not always fully charged when they replace another. A battery will replace another as long as it has significantly more charge, however since there is no data on recharges, the distribution of the charge of the replacement battery is unknown. In Chapter 3, the methods outlined assume that when the battery is replaced the charge becomes "unbounded", and is able to satisfy any trip

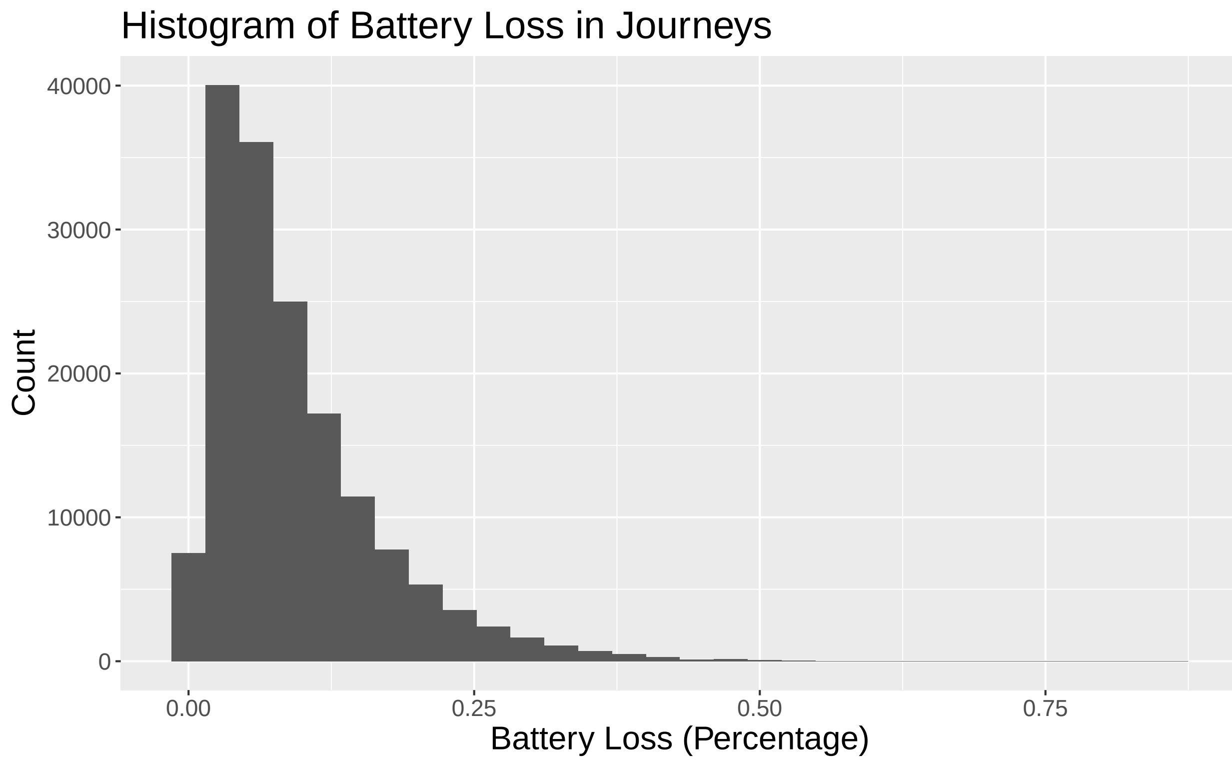


Fig. 4.1: A histogram of battery losses experienced over 16,000 journeys, there are only two recordings that saw more than 80% loss

or journey that may happen next. It is very unlikely for a recharged battery to have less than 80% battery, which will certainly satisfy at least a week's worth of trips. As such, it is reasonable to assume that battery level is at maximum upon a replacement, as long as the period of time observed is not more than a week. For the sake of the simulations performed in Chapter 5, which simulate a weeks worth of trips, the battery level is set to 100% when a replacement occurs.

The assumption that a replaced battery can satisfy any following trip or journey, as all observed replacements result in over 80% battery, the majority of trips use less than 10%, and the majority of journeys use less than 80% (Figure 4.1).

Battery age and health can affect how much charge the battery can store and how quickly it is depleted. Both of these mean that the amount of charge possible is different for each unit. However, with very little information on battery loss, and no information on which battery is in which scooter, it is very difficult to ascertain the quality of a given scooter's battery. Hence it is assumed that the rate of battery loss for each scooter is the same.

It is also assumed that employees replace the battery of the scooter with the lowest battery in the area first. This is a fairly simple assumption, as prioritising scooters with low battery percent is a simple and effective method of performing this task. However, if a zoning method is chosen that creates rather large zones, an employee may prioritise scooters that are closer notwithstanding having slightly higher battery percentage.

It is also initially assumed that users have a tendency to choose the scooter that is most charged when performing a trip. As user behaviour is an added complexity, this assumption allows for a simple derivation that relies only on demand and revenue. However it is unlikely that users choose the most highly-charged scooter as users do not have detailed information on how much charge each scooter has and, since most don't travel that far, they do not usually need a fully charged scooter, so they do not have much reason to walk to another scooter that is more charged. As such, once the initial metric is derived, this assumption is replaced with one where scooter choice is completely random.

4.2 Demand

While the majority of this research has been concerned with linking trip revenue with battery replacement, demand is also important to take into account, as it dictates the likelihood of revenue being generated. Demand is considered due to all of the scooters in a zone being accounted for in the metric. As will be shown in Section 4.3, the number of scooters hired in the next hour dictates how much revenue is generated, and hence revenue is calculated for each possibility of hires in an hour. Then, these revenues are multiplied by the probability of that number of hires, and then the sum of these multiples is calculated.

Demand is represented as $P_{i,j}(n)$, the probability that n scooters are hired at zone i at hour j. As will also be discussed in Section 4.3, demand and revenue patterns are not stratified further by day of week or month of year due to lack of data, but should be done with a sufficiently large dataset.

The maximum value of n is set to the maximum number of trips starting in location i in any given hour. As such, the value varies for each location, and is defined as $n_{\max,i}$.

Demand could be estimated with high accuracy with a detailed algorithm, but three simple estimates are presented here. All three are different methods of calculating the proportion of trips in a given hour.

For calculating any $P_{i,j}$, the dataset of trips is filtered to only contain trips starting in zone i. T_i, a dataset containing the distribution of number of trips performed in an hour is formed by grouping trips by the date and hour they starting in, and counting the number of events. Hence, T_i contains values in $[1, N_i]$ representing the distribution of different values of n at any given hour of the day.

T_i is filtered to only include values of n at hour j, and the demand is approximated for each value of n:

$$P_{i,j}^{(1)}(n) = \frac{\sum I(T_{i,j,k} = n)}{d},$$

where $T_{i,j,k}$ is the kth value of $T_{i,j}$, and I is an indicator variable that is equal to one if $T_{i,j,k} = n$, and zero otherwise. d is the number of days between the earliest observation in T_i and the latest, and hence days with no trips in that hour are also included. Therefore $n \geq 0$. Since more days where trips do not occur will result in $P_{i,j}(n)$ being smaller for all $n > 0$, $P_{i,j}(0)$ is accounted for.

$P^{(1)}$ is the most detailed estimate of demand considered here, as it takes into account both time of day and location. However this will result in a large number of small estimates of revenue increase, especially in areas that are unpopular. As the Adelaide area consists of a few, small, high demand areas surrounded by large areas with significantly less demand, it is expected that that many values of $P_{i,j}^{(1)}$ are equal to zero, especially during low demand areas. For the purpose of battery replacement it is important that batteries are eventually replaced in all scooters in the event that they are required. Using this estimate could result with a metric that has a heavy focus on replacing batteries in those high demand areas only. One method is to automatically rank scooters with 0% battery highly, as to ensure that scooters with absolutely no charge are recharged regardless of demand, which is discussed in Chapter 5. Discussed here are estimates of demand that are less detailed.

For instance, the probability of n trips occurring that start in zone i regardless of the current hour:

$$P_i^{(2)}(n) = \frac{\sum I(t_{i,k} = n)}{d \times 24}.$$

This finds the average hourly demand. Values of $P^{(2)}$ will not get as large as the highest values of $P^{(1)}$, but are less likely to be zero.

For non-zero values of $P_i(n)$, we can remove the cases in which there are no trips:

$$P_i^{(3)}(n) = \frac{\sum I(t_{i,k} = n)}{L},$$

where L is simply the length of T_i.

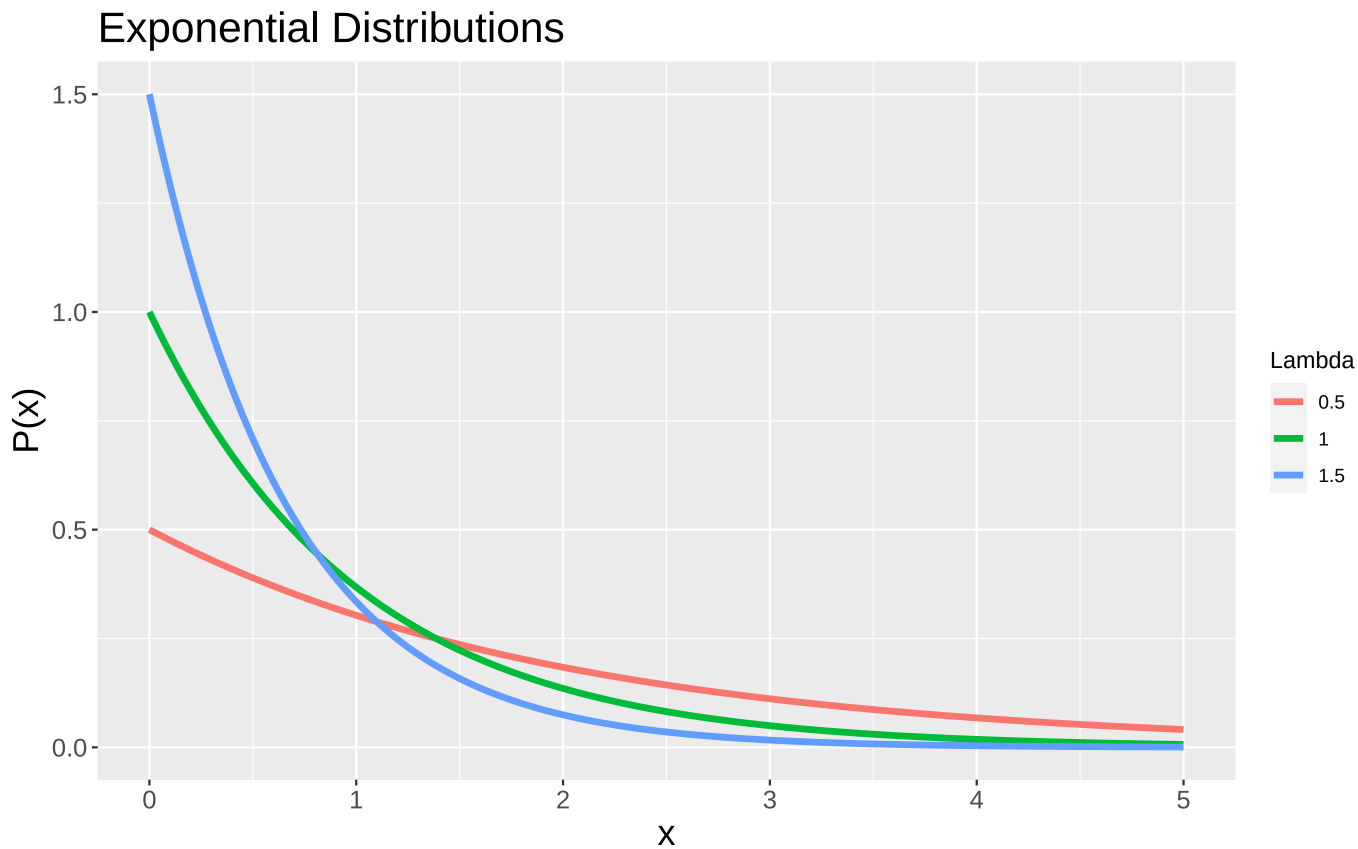
$P_i^3(n)$ is hence only defined for $n > 0$, and is the likelihood of n trips occurring in zone i assuming that a trip occurs in the next hour.

For derivation of the metric in Section 4.3 the superscript is dropped as any of the three approximations can be interchanged into the calculation. In Chapter 5, $P_{i,j}^1$ is used when the metric is implemented.

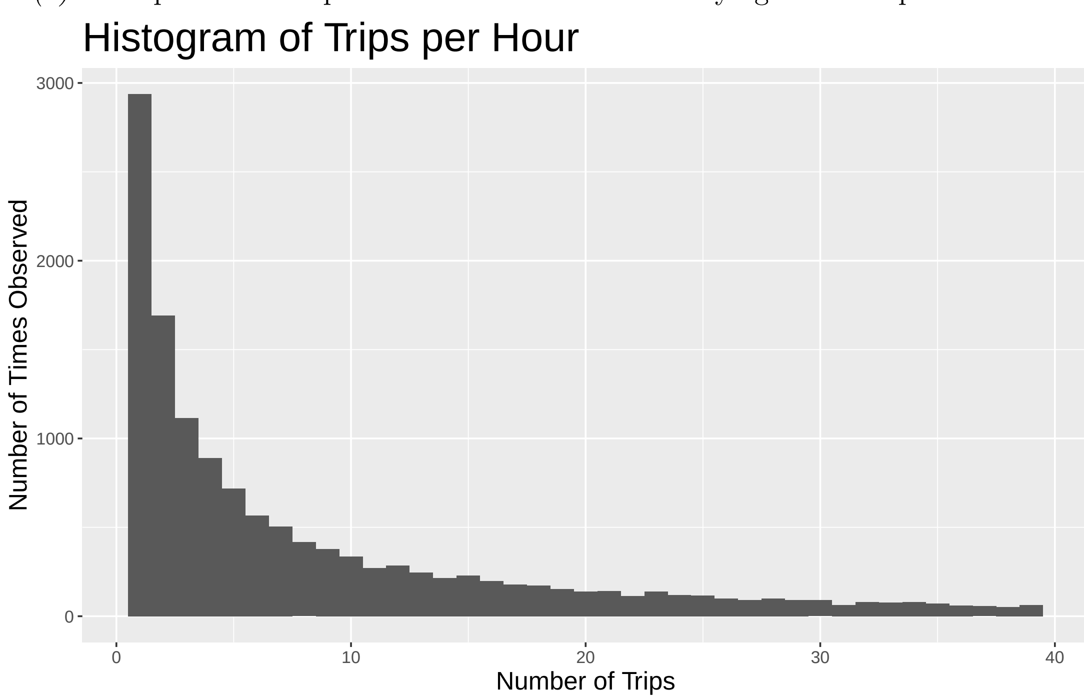
4.2.1 Better approximations of demand

The three approximations shown here are very simple methods to estimate demand. These are easy to calculate, which is useful when a large number of calculations are required, but they do have limitations. For instance, the demand is not defined for values of n larger than what has been seen historically, however it should not be assumed that a random surge in demand is impossible.

As can be seen in Figure 4.2, the likelihood distribution of the number of trips that start at a zone in a given hour follows a similar shape to the exponential distribution. As such, it would be possible to find a parameterisation that allows the distribution to be defined for



(a) Examples of the exponential distribution with varying values of parameter λ.



(b) Histogram of the number of trips recorded each hour from October 2020 to October 2021, for each zone.

Fig. 4.2: Similar to the exponential distribution, the likelihood of a given number of trips occuring in an hour appears to decrease for larger number of trips.

continuous values of n. For zones with higher demand, the value of λ might be larger, as the likelihood of more trips occurring might be greater in relation to that if one trip occurs.

A discrete parameterisation, such as the Poisson distribution, may be better fitting as the data on trips per hour is also discrete. However for all values of the parameter λ, the exponential likelihood is always monotonically decreasing, which is how we expect the likelihood of demand to be. Hence, a discrete approximation of an exponential density should be used.

However, it is assumed that each zone can have even the slightest difference in demand likelihood, and hence this estimate would need to be calculated for each one. A negative binomial distribution, which has a similar shape but two parameters could accommodate this. One parameter would relate to the general pattern of demand throughout the city, whereas the other is calculated for each zone and relates to the specific demand pattern in that area.

Finding a parameterisation for a density, however, is complex and outside the scope of this research. However, since the value of $P_{i,j}$ is unchanged through derivation of the metric, if demand was approximated using a parametrisation, the $P_{i,j}$ would simply be replaced by the estimate.

4.3 Algorithm Formulation

For the sake of exposition, the linear model method will be used to calculate expected revenue increase. Recall from Chapter 3:

$$\hat{R} = \int_{\text{cap}_{i,b}}^{\infty} x f_{i,j}(x) dx \approx E_i[X] - E_{i,b}[X],$$

which relates the revenue increase estimate from the linear model method to the one obtained from KDE filtering. The metric can be changed to calculate revenue increase using KDE filtering by utilising this relationship.

Also recall that $n_{\max,i}$ is the maximum number of trips observed to start in location i in any given hour, and $P_{i,j}(n)$ is not defined for $n > n_{\max,i}$.

Let y_i be the number of scooters available at zone i at the current time. Then, let m_i be the maximum number of scooters that could perform a trip starting in zone i in the next hour.

It follows that:

$$m_i = \min(y_i, n_{\max,i}).$$

The algorithm is formulated by estimating the expected revenue at the current state, $r^c_{i,j}$, and the revenue after h battery replacements are performed, $r^h_{i,j}$. Then, the expected increase in revenue as a result of h replacements is simply:

$$I_{i,j}(h) = r^h_{i,j} - r^c_{i,j}.$$

For $s = 1, 2, ..., y_i$, let $b_{s,i}$ be the battery of scooter s in zone i, in descending order of battery percentage. With the assumption that the most charged scooter is chosen first, the first trip will occur for $s = 1$. The first replacement will always occur for $s = y_i$.

Recall from Chapter 3 that the "revenue cap" is the estimated limit on revenue that a scooter's current battery level provides. Let $\text{cap}_{s,i}$ be the revenue cap for scooter s in zone i, and calculate for $s = 1, 2, ..., m_i$.

From Chapter 3, the equation is

$$\text{cap}_{s,i} = 1 + 0.38 \times l_i(b_{s,i}),$$

where l_i is the linear model of duration given battery loss using data of trips starting in zone i.

4.3.1 Using the User Priority Constraint

To calculate $r^c_{i,j}$, the expected revenue gained in the current hour j at zone i we take each possible value of n and calculate the revenue obtained in each case. Then, we multiply each by the probability of n trips occurring, and sum over all.

$$
\begin{aligned}
r^c_{i,j} = & P_{i,j}(1) \times \int_0^{\text{cap}_{1,i}} x f_{i,j}(x) dx \\
& + P_{i,j}(2) \times \left[\int_0^{\text{cap}_{1,i}} x f_{i,j}(x) dx + \int_0^{\text{cap}_{2,i}} x f_{i,j}(x) dx \right] \\
& + ... + \\
& + P_{i,j}(\geq m_i) \times \left[\int_0^{\text{cap}_{1,i}} x f_{i,j}(x) dx + \int_0^{\text{cap}_{2,i}} x f_{i,j}(x) dx + ... + \int_0^{\text{cap}_{m_i,i}} x f_{i,j}(x) dx \right].
\end{aligned}
$$

This calculates the expected revenue for each value of $n \leq m_i$ by summing over the expected values of revenue for every scooter $s \leq n$, which are each limited by $\text{cap}_{s,i}$. Under the assumption that the most charged scooter is hired first, when $n = 1$ the scooter $s = 1$ is hired, when $n = 2$, the scooters $s = 1$ and $s = 2$ are both hired, and so on.

This revenue is then multiplied by the probability of that many scooters being hired, and since:

$$
\sum_{k=0}^{n_i} P_i^{(a)}(k) = 1 \quad \text{for } a = 1, 2,
$$

the revenue in each case is effectively multiplied by the proportion of instances where k trips have occurred in that zone, scaling the expected revenue by its corresponding likelihood. While the case $n = 0$ reduces to 0 in $r^c_{i,j}$, as the revenue is equal to \$0, low demand still has an effect on the resulting value due to the sum. If $P_{i,j}(0)$ has a high likelihood compared to $P_{i,j}(n)$ for $n > 0$, the values of $P_{i,j}(n)$ will be small and the expected revenue will be small as well.

For the case of $a = 3$:

$$
\sum_{k=1}^{n_i} P_i^3(k) = 1,
$$

as $P^3(n)$ is not defined for $n = 0$, and as such does not account for the likelihood that no revenue is generated. High values of $P_{i,j}(0)$ compared to $P_{i,j}(n)$ for $n > 0$ will have no effect on the expected revenue.

To ensure that this sum holds, the probability $P_i(\geq m_i)$ is used in the last case where $n = m_i$. In the event that the number of scooters available is smaller than the possible demand, $n_i \geq y_i$, the calculation needs to account for the possibility that the demand will be larger than m_i. In these cases, m_i scooters will still be hired, generating m_i trips worth of revenue; and $n - m_i$ trips will be missed, so they generate the same amount of revenue as if demand was equal to m_i. To accommodate this:

$$P_i(\geq m_i) = P_i(m_i) + P_i(m_i + 1) + ... + P_i(n_i).$$

The integral:

$$\int_0^{\text{cap}_{S,i}} x f_{i,j}(x) dx,$$

which represents the expected revenue generated from a trip performed by scooter S at hour j using the revenue patterns observe in zone i, and restricted by the battery of scooter S, is repeated often in the calculation of overall expected revenue.

Hence we define:

$$R_{i,j}(S) = \int_0^{\text{cap}_{S,i}} x f_{i,j}(x) dx.$$

Hence $R_{i,j}$ is the expected revenue from a scooter's next trip if it takes place at hour j in zone i. When a battery is replaced, and $R_{i,j}$ is no longer restricted by battery, we use the substitution:

$$R_{i,j} = \int_0^{\infty} x f_{i,j}(x) dx,$$

to represent how the expected revenue is unbounded.

Replacing One Battery

In the case that one battery is replaced, we let $h = 1$ and the least charged scooter $s = y_i$ has its battery replaced. Hence $b_{y_i,i} = 100$, and the cap on revenue is removed, resulting in the integral of the revenue density for that scooter to be unbounded. We now assume that scooter for which $s = y_i$ is hired first, $s = 1$ is hired second and so on.

The new revenue is:

$$\begin{aligned} r^1_{i,j} &= P_{i,j}(1) \times R_{i,j} \\ &+ P_{i,j}(2) \times [R_{i,j} + R_{i,j}(1)] \\ &+ ...+ \\ &+ P_{i,j}(\geq m_i) \times [R_{i,j} + R_{i,j}(1) + ... + R_{i,j}(m_i)] \,. \end{aligned}$$

Hence the expected increase in revenue is:

$$\begin{aligned} I_{i,j}(1) = r^1_{i,j} - r^c_{i,j} &= P_{i,j}(1) \times [R_{i,j} - R_{i,j}(1)] \\ &+ P_{i,j}(2) \times [R_{i,j} + R_{i,j}(1) - (R_{i,j}(1) + R_{i,j}(2))] \\ &+ ...+ \\ &+ P_{i,j}(\geq m_i) \times [R_{i,j} + R_{i,j}(1) + ... + R_{i,j}(m_i) \\ &\quad - (R_{i,j}(1) + R_{i,j}(2) + ... + R_{i,j}(m_i))] \,. \end{aligned}$$

Using the following:

$$R_{i,j} - R_{i,j}(l) = \int_0^{\infty} x f_{i,j}(x)dx - \int_0^{\text{cap}_{l,i}} x f_{i,j}(x)dx = \int_{\text{cap}_{1,i}}^{\infty} x f_{i,j}(x)dx,$$

$I_{i,j}(1)$ simplifies to:

$$
\begin{aligned}
I_{i,j}(1) = P_{i,j}(1) &\int_{\text{cap}_{1,i}}^{\infty} x f_{i,j}(x) dx \\
&+ P_{i,j}(2) \int_{\text{cap}_{2,i}}^{\infty} x f_{i,j}(x) dx \\
&+ ...+ \\
&+ P_{i,j}(\geq m_i) \int_{\text{cap}_{m,i}}^{\infty} x f_{i,j}(x) dx.
\end{aligned}
$$

Despite the scooter of lowest battery, $s = y_i$, recharged, the integrals are concerned with the scooters of highest battery. For instance, if $y_i > m_i$, the scooter being replaced will not be relevant in calculation of the increase in revenue from replacing that scooters battery.

This should be expected, as while recharging the lowest charged scooter results in the largest battery change, under the assumption that users select the most charged scooter first the actual impact comes from the fact that scooter $s = 1$ is not being chosen. In the $n = 1$ case, it does not matter which scooter is recharged, there is only a greater effect of recharging a lower charged scooter if more scooters are hired at once. Hence, $I_{i,j}(1)$, the impact on revenue, will increase as the likelihood of higher demand increases.

In the $n = 2$ case, the first scooter is hired whether or not a scooter is recharged, and hence that scooter's revenue does not need to be considered when observing the increase. Here, the recharged scooter is chosen over the second scooter, and the increase in revenue comes from the difference between that scooter's expected revenue and the second scooter's.

For each value of n, only 1 integral needs to be calculated.

Replacing Two Batteries

When two scooters are recharged, the revenue distributions for both scooters are the same, so it does not matter which one is chosen in the case that $n = 1$.

The revenue expectation after the recharges is:

$$
\begin{aligned}
r_{i,j}^2 = & P_{i,j}(1) \times R_{i,j} \\
& + P_{i,j}(2) \times 2R_{i,j} \\
& + P_{i,j}(3) \times [2R_{i,j} + R_{i,j}(1)] \\
& + ... + \\
& + P_{i,j}(\geq m_i) \times [2R_{i,j} + R_{i,j}(1) + ... + R_{i,j}(m_i)] \, .
\end{aligned}
$$

Hence the expected increase in revenue for recharging two scooters is:

$$
\begin{aligned}
I_{i,j}(2) = r_{i,j}^2 - r_{i,j}^c = & P_{i,j}(1) \times [R_{i,j} - R_{i,j}(1)] \\
& + P_{i,j}(2) \times [2R_{i,j} - (R_{i,j}(1) + R_{i,j}(2))] \\
& + P_{i,j}(3) \times [2R_{i,j} + R_{i,j}(1) \\
& \quad - (R_{i,j}(1) + R_{i,j}(2) + R_{i,j}(3))] \\
& + ... + \\
& + P_{i,j}(\geq m_i) \times [2R_{i,j} + R_{i,j}(1) + ... + R_{i,j}(m_i) \\
& \quad - (R_{i,j}(1) + R_{i,j}(2) + ... + R_{i,j}(m_i))] \, .
\end{aligned}
$$

This simplifies to:

$$\begin{aligned}
I_{i,j}(2) = P_{i,j}(1) &\int_{\text{cap}_{1,i}}^{\infty} x f_{i,j}(x)dx \\
&+ P_{i,j}(2)\left[\int_{\text{cap}_{1,i}}^{\infty} x f_{i,j}(x)dx + \int_{\text{cap}_{2,i}}^{\infty} x f_{i,j}(x)dx\right] \\
&+ P_{i,j}(3)\left[\int_{\text{cap}_{2,i}}^{\infty} x f_{i,j}(x)dx + \int_{\text{cap}_{3,i}}^{\infty} x f_{i,j}(x)dx\right] \\
&+ ...+ \\
&+ P_{i,j}(\geq m_i)\left[\int_{\text{cap}_{m-1,i}}^{\infty} x f_{i,j}(x)dx + \int_{\text{cap}_{m,i}}^{\infty} x f_{i,j}(x)dx\right].
\end{aligned}$$

It follows that this is the increase in revenue when the recharged scooters are chosen after the last two scooters that would be originally chosen. The $n = 1$ case is unchanged, and the $n = 2$ case is simply the integral of the revenue density to the right of the cap for the two chosen scooters. Then, for higher demand, the two scooters considered are those with lower battery.

From these derivations, the function $I_{i,j}(a)$, which calculates the expected increase in revenue when a scooters are recharged in zone i, can be defined as:

$$I_{i,j}(a) = \sum_{z=1}^{m_i} P_{i,j}(z) \sum_{k=z-a+1,k>0}^{j} \int_{\text{cap}_{k,i}}^{\infty} x f_{i,j}(x)dx. \tag{4.1}$$

For approximations of demand that do not take into account the time of day, $I_i(a)$ is:

$$I_i(a) = \sum_{z=1}^{m} P_i(z) \sum_{k=z-a+1,k>0}^{j} \int_{\text{cap}_{k,i}}^{\infty} x f_i(x)dx,$$

where $f_i(x)$ is the density estimate using trip data starting in zone i regardless of time.

A simple replacement is used to calculate $I_{i,j}(a)$ using KDE filtering.

KDE Filtering

$I_{i,j}(a)$ is in the form:

$$I_{i,j}(b) = \sum_{z=1}^{m} P_{i,j}(z)\hat{R}_{i,j}(b),$$

here $P_{i,j}(z)$ is demand, and $R_{i,j}(b)$ is the increase in revenue. Demand does not depend on the method used to calculate expected revenue increase, so a metric that uses KDE filtering can be derived by using the equivalent estimate for $\hat{R}_{i,j}(b)$.

Note that the integral:

$$\int_{\text{cap}_{k,i}}^{\infty} x f_{i,j}(x)dx$$

in equation 4.1 is the same as the estimate of revenue increase using a linear model calculated in Section 4.1.1.

Recall from Section 4.3 that the equivalent estimate for revenue increase using KDE filtering is:

$$\hat{R}_K = \int_{0}^{\infty} x f_{i,j}(x)dx - \int_{0}^{\infty} x f_{i,j,k}(x)dx,$$

where $f_{i,j}$ is the density for revenue estimated using a KDE generated using all trips starting in zone i at hour j, and $f_{i,j,k}$ is generated using all trips starting in zone i that use less battery than scooter $s = k$ currently has.

Hence the metric for calculating the expected increase in revenue with KDE filtering when replacing a batteries in zone i is:

$$I_{i,j}(a) = \sum_{z=1}^{m_i} P_{i,j}(z) \sum_{k=z-a+1,k>0}^{z} \left(\int_{0}^{\infty} x f_{i,j}(x)dx - \int_{0}^{\infty} x f_{i,j,k}(x)dx \right).$$

For approximations of demand that do not take into account time of day, $I_i(a)$ calculated using KDE filtering is:

$$I_{i}(a) = \sum_{z=1}^{m} P_{i}(z) \sum_{k=z-a+1,k>0}^{z} \left(\int_{0}^{\infty} x f_{i,j}(x)dx - \int_{0}^{\infty} x f_{i,j,k}(x)dx \right).$$

4.3.2 Relaxing the User Priority Constraint

Now that a basic form for the priority metric has been defined, we can begin to relax the assumption that users choose the most charged scooter first. There are two problems with the assumption of scooter selection that provide incentive for its removal.

First, this is not likely common practice. We presume that people are more likely to choose a scooter that is closest to them, especially considering that the battery level of a scooter is unknown until they open the app to hire a scooter. Since scooters can satisfy trips with low battery percentage, it also not required for users to take into account battery level when selecting a scooter.

Second, as seen in Section 4.3.1, when this assumption is made, it does not matter which scooter is recharged when calculating revenue increase. The increase in revenue comes from satisfying trips that could not be satisfied by the most charged scooters in that zone, which is likely to be small.

Relaxing the user priority constraint will not only represent user patterns better, but will also allow for larger zone sizes, as assuming that users prioritise one scooter over the rest also assumes that users are willing to travel large distances, which is unlikely.

It is now assumed that the chance that any scooter s is chosen in a region i, is purely random. Recall that the number of scooters available in zone i at the current time is represented as y_i. Hence the chance that any scooter is chosen is $\frac{1}{y_i}$. For multiple scooters, since the same scooter can't be hired twice, the probability of a specific combination of n scooters is $\frac{1}{C_n^{y_i}}$. C refers to the binomial coefficient, for which the formula is:

$$C_k^n = \begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}.$$

The formula calculates the number of possible ways to choose k objects from a sample of n objects, without replacement.

Recall that the number of scooters to be considered in zone i is denoted by m_i, in the event that the supply of scooters in that zone, y_i, is larger than the maximum numbers of scooters hired in i, $n_{\max,i}$.

The current expected revenue for zone i from Section 4.3.1 is now modified to include the probability of individual scooters being selected by users. For each case in which a scooters are hired, the expected revenue is calculated for each combination of a scooters in zone i, and is divided by $C_a^{y_i}$, the number of combinations, as the probability of each combination occurring is uniform.

Hence, $r_{i,j}^c$ becomes:

$$\begin{aligned} r_{i,j}^c = & P_{i,j}(1) \times \frac{1}{y_i} \sum_{a=1}^{y_i} R_{i,j}(1) \\ & + P_{i,j}(2) \times \frac{1}{C_2^{y_i}} \sum_{a=2}^{y_i} \sum_{k=1}^{a-1} \left[R_{i,j}(a) + \int_0^{\text{cap}_{k,i}} x f_{i,j}(x) dx \right] \qquad (4.2) \\ & + ... + \\ & + P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}} \sum_{a=m_i}^{y_i} \sum_{k=m_i-1}^{a-1} ... \sum_{l=1} \left[R_{i,j}(a) + R_{i,j}(k) + ... + \int_0^{\text{cap}_{l,i}} x f_{i,j}(x) dx \right]. \qquad (4.3) \end{aligned}$$

Now, the expected increase in revenue is calculated by subtracting this from the expected revenue when a battery is replaced.

Replacing One Battery

When one battery is replaced, for $n = 1$ the last scooter is removed from the sum, and instead the unbounded revenue expectation is added to it. When $n > 1$, the other scooters hired are also random. This presents two cases, one where one of the scooters has had its battery replaced, and one where none of them have. This results in:

$$
\begin{aligned}
r^1_{i,j} &= P_{i,j}(1) \times \frac{1}{y_i}\left[R_{i,j} + \sum_{a=1}^{y_i-1} R_{i,j}(a)\right] \\
&+ P_{i,j}(2) \times \frac{1}{C_2^{y_i}}\left[\sum_{a=1}^{y_i-1}(R_{i,j} + R_{i,j}(a))\right. && (4.4)\\
&\quad \left. + \sum_{a=2}^{y_i-1}\sum_{k=1}^{a-1}(R_{i,j}(a) + R_{i,j}(k))\right] && (4.5)\\
&+ \dots + \\
&+ P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}} \\
&\quad \times \left[\sum_{m_i-1}^{y_i-1}\sum_{m_i-2}^{a-1}\dots\sum_{l=1}(R_{i,j} + R_{i,j}(a) + R_{i,j}(k) + \dots + R_{i,j}(k))\right. && (4.6)\\
&\quad\quad \left. + \sum_{a=m_i}^{y_i-1}\sum_{k=m_i-1}^{a-1}\dots\sum_{l=1}(R_{i,j}(a) + R_{i,j}(k) + \dots + R_{i,j}(k))\right].
\end{aligned}
$$

First note that (see 4.4):

$$\sum_{a=1}^{y_i-1} R_{i,j} = (y_i - 1)R_{i,j},$$

and (see 4.6):

$$\sum_{m_i-1}^{y_i-1}\sum_{m_i-2}^{a-1}\dots\sum_{l=1} R_{i,j} = C_{m_i-1}^{y_i-1}R_{i,j}.$$

Also, in the subtraction $r^1_{i,j} - r^c_{i,j}$, parts of the sums cancel. For instance, in the $n = 2$ case, $r^c_{i,j}$ and $r^1_{i,j}$ contain:

$$\sum_{a=2}^{y_i}\sum_{k=1}^{a-1}(R_{i,j}(a) + R_{i,j}(k)) \quad \text{and} \quad \sum_{a=2}^{y_i-1}\sum_{k=1}^{a-1}(R_{i,j}(a) + R_{i,j}(k))$$

respectively.

The difference between these two expressions is that $r^c_{i,j}$ has one more value for j than $r^1_{i,j}$ does. This means that when $r^c_{i,j}$ is subtracted from $r^1_{i,j}$, all of the terms cancel except for the $j = y_i$ term. Hence the result is:

$$(4.5) - (4.2) = -P_{i,j}(2) \times \frac{1}{C_2^{y_i}} \sum_{k=1}^{y_i-1} \left(R_{i,j}(y_i) + R_{i,j}(k)\right).$$

This simplifies to:

$$(4.5) - (4.2) = -P_{i,j}(2) \times \frac{1}{C_2^{y_i}} \left((y_i - 1)R_{i,j}(y_i) + \sum_{k=1}^{y_i-1} R_{i,j}(k).\right)$$

Using these findings, the expected revenue increase is:

$$\begin{aligned}
I_1 = & P_{i,j}(1) \times \frac{1}{y_i} \left[R_{i,j} - R_{i,j}(y_i)\right] \\
& + P_{i,j}(2) \times \frac{1}{C_2^{y_i}} \left[(y_i - 1)R_{i,j} - (y_i - 1)R_{i,j}(y_i) + \sum_{a=1}^{y_i-1} R_{i,j}(a) - \sum_{k=1}^{y_i-1} R_{i,j}(k)\right] \\
& + \ldots + \\
& + P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}} \Bigg[C_{m_i-1}^{y_i-1} R_{i,j} - C_{m_i-1}^{y_i-1} R_{i,j}(y_i) \\
& \quad + \sum_{m_i-1}^{y_i-1} \sum_{m_i-2}^{a-1} \ldots \sum_{l=1} \left(R_{i,j} + R_{i,j}(a) + R_{i,j}(k) + \ldots + R_{i,j}(k)\right) \\
& \quad - \sum_{k=m_i-1}^{y_i-1} \ldots \sum_{l=1} \left(R_{i,j}(k) + \ldots + R_{i,j}(k)\right)\Bigg].
\end{aligned}$$

This further simplifies to:

$$\begin{aligned}
I_1 = & \frac{P_{i,j}(1)}{y_i} \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx \\
& + \frac{P_{i,j}(2)}{C_2^{y_i}} (y_i - 1) \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx \\
& + \ldots + \\
& + \frac{P_{i,j}(\geq m_i)}{C_{m_i}^{y_i}} C_{m_i-1}^{y_i-1} \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx.
\end{aligned}$$

The constant at the front of each integral can also be simplified as:

$$\frac{1}{C^{y_i}_{m_i}} C^{y_i - 1}_{m_i - 1} = \frac{(y_i - m_i)!m_i!}{y_i!} \frac{(y_i - 1)!}{(y_i - 1 + m_i + 1)!(m_i - 1)!} = \frac{m_i}{y_i}.$$

Hence:

$$I_1 = \frac{P_{i,j}(1)}{y_i} \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx + \frac{2P_{i,j}(2)}{y_i} \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx + ... + \frac{m_i P_{i,j}(\geq m_i)}{y_i} \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx$$
$$= \int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx \sum_{a=1}^{m_i} \frac{j P_{i,j}(j)}{y_i}.$$

The sum is the probability that j scooters are hired multiplied by the probability that the recharged scooter is used when j scooters are hired. The result is simply the revenue obtained from replacement multiplied by the probability that replacement has an effect on revenue in the next hour.

Replacing Two Batteries

When two batteries are replaced, the integrals for which $n > 1$ now have three cases. One where both recharged scooters are hired, another where only one is hired, and one where none of the scooters are hired. There are double the number of combinations for the case in which one recharged scooter is hired, as it can be either one of the two scooters.

The expected revenue when two batteries are replaced is:

$$
\begin{aligned}
r_{i,j}^2 &= P_{i,j}(1) \times \frac{1}{y_i}\left[2R_{i,j} + \sum_{a=1}^{y_i-2} R_{i,j}(a)\right] \\
&+ P_{i,j}(2) \times \frac{1}{C_2^{y_i}}\left[2R_{i,j} + 2\sum_{a=1}^{y_i-2}(R_{i,j} + R_{i,j}(a))\right. \\
&\left. + \sum_{a=2}^{y_i-2}\sum_{k=1}^{a-1}(R_{i,j}(a) + R_{i,j}(k))\right] \qquad (4.7) \\
&+ ... + \\
&+ P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}}\left[\sum_{a=m_i-2}^{y_i-2}\sum_{k=m_i-3}^{a-1} ... \sum_{l=1}(2R_{i,j} + R_{i,j}(a)\right. \\
&\quad + R_{i,j}(k) + \ ... + R_{i,j}(l)) \\
&+ 2\sum_{a=m_i-1}^{y_i-2}\sum_{k=m_i-2}^{a-1} ... \sum_{l=1}(R_{i,j} + R_{i,j}(a) + R_{i,j}(l) \qquad (4.8) \\
&\quad + ... + R_{i,j}(k)) \qquad (4.9) \\
&\left. + \sum_{a=m_i}^{y_i-2}\sum_{k=m_i-1}^{a-1} ... \sum_{l=1}(R_{i,j}(a) + R_{i,j}(k) + ... + R_{i,j}(l))\right]. \qquad (4.10)
\end{aligned}
$$

However, this assumes that it is possible for more than m_i scooters to be hired. If $m_i = y_i$, lines (4.8) and (4.9) are omitted as they assume that only one of the recharged scooters is hired, but $m_i = y_i$ means all available scooters are hired. Note that m_i cannot be larger than y_i.

Noting the following:

$$\sum_{a=1}^{y_i-2} R_{i,j} = (y_i - 2)R_{i,j}, \qquad (4.11)$$

$$nonumber \sum_{a=m_i-2}^{y_i-2}\sum_{k=m_i-3}^{a-1} ... \sum_{l=1} R_{i,j} = C_{m_i-2}^{y_i-2}R_{i,j}, \text{ and} \qquad (4.12)$$

$$\sum_{a=m_i-1}^{y_i-2}\sum_{k=m_i-2}^{a-1} ... \sum_{l=1} R_{i,j} = C_{m_i-1}^{y_i-2}R_{i,j}. \qquad (4.13)$$

and subtracting individual segments in $r^2_{i,j}$ and $r^c_{i,j}$, from $n = 2$ (4.7-4.2) we have:

$$\begin{aligned}
&\sum_{a=2}^{y_i-2}\sum_{k=1}^{a-1}(R_{i,j}(a)+R_{i,j}(k))\\
&-\sum_{a=2}^{y_i}\sum_{k=1}^{a-1}(R_{i,j}(a)+R_{i,j}(k))\\
&=-\sum_{a=y_i-1}^{y_i}\sum_{k=1}^{a-1}(R_{i,j}(a)+R_{i,j}(k))\\
&=-(y_i-1)\sum_{a=y_i-1}^{y_i}R_{i,j}(a)-2\sum_{a=1}^{y_i-2}R_{i,j}(a),
\end{aligned}$$

as the combinations in which j is not y_i or $y_i - 1$ are cancelled out. There are $(y_i - 1)$ instances for each value of a as when scooter $s = y_i$ or $s = y_i - 1$ are hired, there are $(y_i - 1)$ other scooters that can be hired alongside it.

For $n = m_i$, the increase in revenues for combinations where neither recharged scooters are hired (4.10) - (4.3):

$$\begin{aligned}
&\sum_{a=m_i}^{y_i-2}\sum_{k=m_i-1}^{a-1}...\sum_{l=1}\left[R_{i,j}(a)+R_{i,j}(k)+...+R_{i,j}(l)\right]\\
&-\sum_{a=m_i}^{y_i}\sum_{k=m_i-1}^{a-1}...\sum_{l=1}\left[R_{i,j}(a)+R_{i,j}(k)+...+R_{i,j}(l)\right]\\
&=-\sum_{u=y_i-1}^{y_i}\sum_{a=m_i-1}^{y_i-2}\sum_{k=m_i-2}^{y_i-2}....\sum_{l=1}\left[R_{i,j}(u)+R_{i,j}(a)+R_{i,j}(k)\right.\\
&\quad\left.+...+R_{i,j}(l)\right]\\
&\quad-\sum_{a=m_i-2}^{y_i-2}\sum_{k=m_i-3}^{y_i-2}....\sum_{l=1}\left[R_{i,j}(y_i)+R_{i,j}(y_i-1)\right.\\
&\quad\left.+R_{i,j}(a)+R_{i,j}(k)+...+R_{i,j}(l)\right]\\
&=-C_{m_i-1}^{y_i-2}\sum_{a=y_i-1}^{y_i}R_{i,j}(a)-C_{m_i-2}^{y_i-2}\left[R_{i,j}(y_i)+R_{i,j}(y_i-1)\right]\\
&\quad-2\sum_{a=m_i-1}^{y_i-2}\sum_{k=m_i-2}^{y_i-2}....\sum_{l=1}\left[R_{i,j}(a)+R_{i,j}(k)+...+R_{i,j}(l)\right] \qquad (4.14)\\
&\quad-\sum_{a=m_i-2}^{y_i-2}\sum_{k=m_i-3}^{y_i-2}....\sum_{l=1}\left[R_{i,j}(a)+R_{i,j}(k)+...+R_{i,j}(l)\right]. \qquad (4.15)
\end{aligned}$$

Note that line (4.14) cancels with line (4.9) and line (4.15) cancels with line (4.8) when the unbounded integrals are removed as per (4.12) and (4.13).

It is important to note again that this is only valid for $m_i < y_i$, as in line (4.14), there is an assumption that when m_i scooters are hired, one of the replaced scooters may not be one of those. For this to occur, y_i must be greater than m_i. When $m_i = y_i$, the increase in revenue equals:

$$\begin{aligned}
&-C_{m_i-2}^{y_i-2}\left(R_{i,j}(y_i)+\int_0^{\text{cap}_{y_i-1,i}}xf_{i,j}(x)dx\right)\\
&-\sum_{a=m_i-2}^{y_i-2}\sum_{k=m_i-3}^{y_i-2}....\sum_{l=1}\left(R_{i,j}(a)+R_{i,j}(k)+...+\int_0^{\text{cap}_{l,i}}xf_{i,j}(x)dx\right)
\end{aligned}$$

as the case where only one recharged scooter is hired is removed.

Hence the value of $I_{i,j}(2)$ when $m_i < y_i$:

$$\begin{aligned} I_{i,j}(2) = & P_{i,j}(1) \times \frac{1}{y_i}\left[2R_{i,j} - R_{i,j}(y_i) - R_{i,j}(y_i - 1)\right] \\ & + P_{i,j}(2) \times \frac{1}{C_2^{y_i}}\Bigg[2R_{i,j} + 2(y_i - 2)R_{i,j} + 2\sum_{a=1}^{y_i-2} R_{i,j}(a) \\ & -(y_i - 1)\sum_{a=y_i-1}^{y_i} R_{i,j}(a) - 2\sum_{a=1}^{y_i-2} R_{i,j}(a)\Bigg] \\ & + P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}}\Bigg[2C_{m_i-2}^{y_i-2}R_{i,j} + C_{m_i-1}^{y_i-2}R_{i,j} \\ & -C_{m_i-1}^{y_i-2}\sum_{a=y_i-1}^{y_i} R_{i,j}(a) - C_{m_i-2}^{y_i-2}\left(R_{i,j}(y_i) + R_{i,j}(y_i - 1)\right)\Bigg]. \end{aligned} \tag{4.16}$$

Which simplifies to:

$$\begin{aligned} I_{i,j}(2) = & P_{i,j}(1) \times \frac{1}{y_i}\left[\int_{\text{cap}_{y_i,i}}^{\infty} xf_{i,j}(x)dx + \int_{\text{cap}_{y_i-1,i}}^{\infty} xf_{i,j}(x)dx\right] \\ & + P_{i,j}(2) \times \frac{y_i - 1}{C_2^{y_i}}\left[\int_{\text{cap}_{y_i,i}}^{\infty} xf_{i,j}(x)dx + \int_{\text{cap}_{y_i-1,i}}^{\infty} xf_{i,j}(x)dx\right] \\ & + \ ... + \\ & + P_{i,j}(\geq m_i)\frac{C_{m_i-1}^{y_i-2} + C_{m_i-2}^{y_i-2}}{C_{m_i}^{y_i}}\left[\int_{\text{cap}_{y_i,i}}^{\infty} xf_{i,j}(x)dx + \int_{\text{cap}_{y_i-1,i}}^{\infty} xf_{i,j}(x)dx\right]. \end{aligned} \tag{4.17}$$

Using the fact that:

$$C_0^{y_i-1} + C_1^{y_i-1} = 1 + y_i - 2 = y_i - 1$$

$I_{i,j}(2)$ can be simplified further:

$$I_{i,j}(2) = \left(\int_{\text{cap}_{y_i,i}}^{\infty} xf_{i,j}(x)dx + \int_{\text{cap}_{y_i-1,i}}^{\infty} xf_{i,j}(x)dx\right)\sum_{a=1}^{m_i} P_{i,j}(j)\frac{C_{j-2}^{y_i-2} + C_{j-1}^{y_i-2}}{C_2^{y_i}} \quad m_i < y_i \tag{4.18}$$

When $m_i = y_i$, line (4.16) becomes:

$$P_{i,j}(\geq m_i) \times \frac{1}{C_{m_i}^{y_i}} \left[2C_{m_i-2}^{y_i-2} R_{i,j} - C_{m_i-2}^{y_i-2} \left(R_{i,j}(y_i) + \int_0^{\text{cap}_{y_i-1,i}} x f_{i,j}(x) dx \right) \right] \tag{4.19}$$

$$= P_{i,j}(\geq m_i) \frac{C_{m_i-2}^{y_i-2}}{C_{m_i}^{y_i}} \left[\int_{\text{cap}_{y_i,i}}^{\infty} x f_{i,j}(x) dx + \int_{\text{cap}_{y_i-1,i}}^{\infty} x f_{i,j}(x) dx \right]. \tag{4.20}$$

When $y_i = m_i$, $m_i - 1 > y_i - 2$. This means, as long as $C_k^n = 0$ for $n < k$, line (4.20) is equivalent to line (4.17) for $y_i = m_i$, and equation (4.18) holds for all $m_i \leq y_i$.

From $I_{i,j}(1)$ and $I_{i,j}(2)$, it follows that:

$$I_{i,j}^L(b) = \sum_{z=1}^{m_i} P_{i,j}(z) \frac{\sum_{k=0}^{b-1} C_{z-b+k}^{y_i-b}}{C_z^{y_i}} \left[\sum_{l=0}^{b-1} \int_{\text{cap}_{(y_i-l),i}}^{\infty} x f_{i,j}(x) dx \right]. \tag{4.21}$$

For KDE filtering:

$$I_{i,j}^K(b) = \sum_{z=1}^{m_i} P_{i,j}(z) \frac{\sum_{k=0}^{b-1} C_{z-b+k}^{y_i-b}}{C_z^{y_i}} \sum_{l=0}^{b-1} \left[\int_0^{\infty} x f_{i,j}(x) dx - \int_0^{\infty} x f_{i,j,y_i-l}(x) dx \right]. \tag{4.22}$$

Where $f_{i,j,l}(x)$ is the revenue density estimate from trips starting in zone i at hour j that use less battery than the current battery percentage of scooter $s = y_i - l$.

4.4 Histogram Normalisation

Generating a KDE and a function that calculates the integral of another function are both computationally intensive tasks. In the next chapter, the effectiveness of the derived priority metrics is tested by running simulations of the scooter environment, that make decisions on which scooters to charge at each hour. In these simulations $I_i(b)$ is calculated multiple times, and hence the running time for a simulation can be incredibly long. $I_i(b)$ is calculated for each zone, at every hour during a simulation. If multiple simulations are to be compared, it is important to find a more efficient method of generating results. Using fundamental properties of KDEs and integrals, the complexity of the algorithm can be reduced.

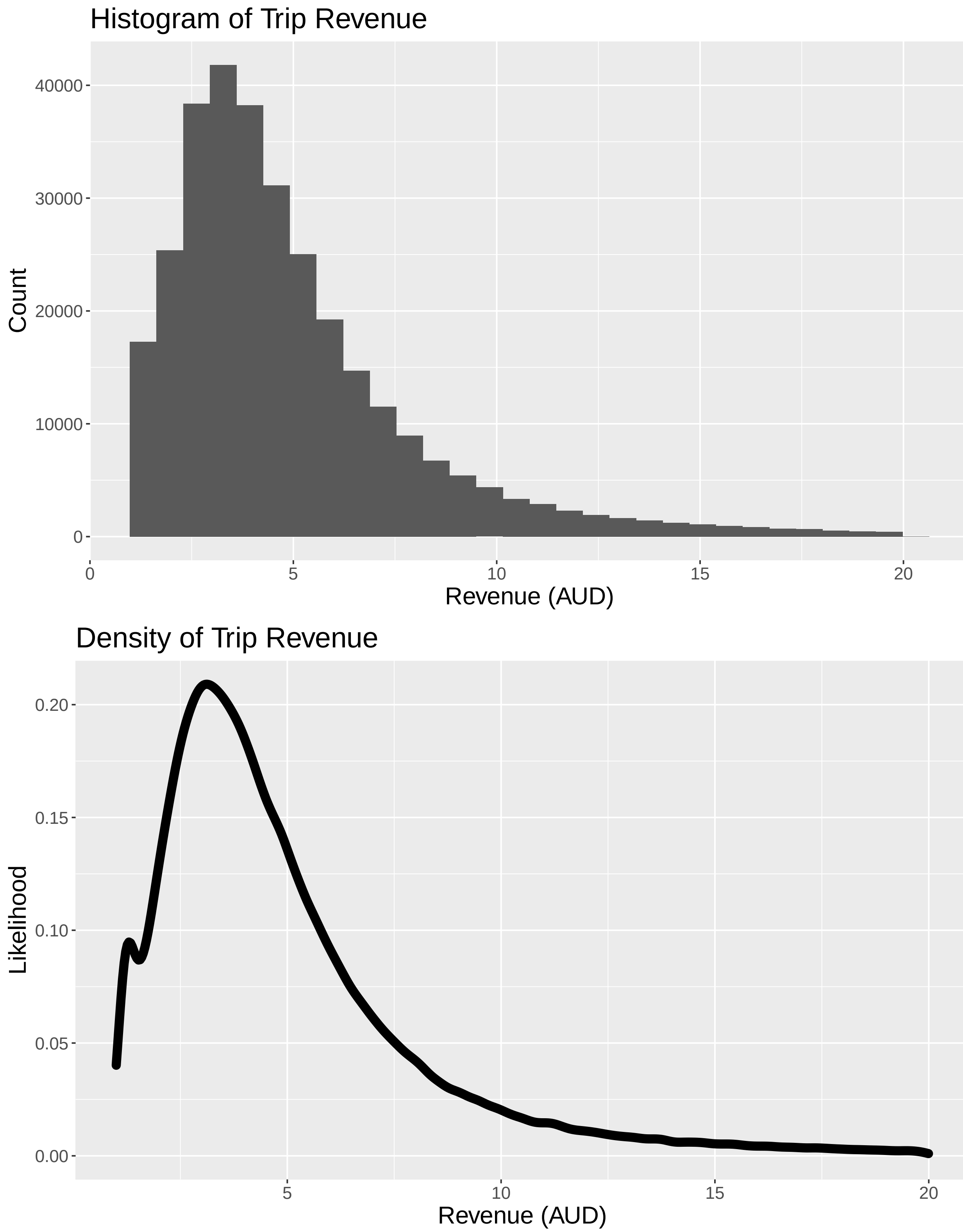


Fig. 4.3: Histogram of revenue observed over all trips (top) vs the estimated density function of trip revenue using the same data and a kernel density package (bottom).

As outlined in Chapter 2, a KDE approximates the shape of the density for given data by summing the shapes of small densities created at each point in the data. As seen in Figure 4.3, in practice a KDE attempts to smooth the shape of the histogram of the data.

Integrals can be approximated using Riemann sums, representing the shape of the curve with rectangles of appropriately small width. This is computationally less intense than calculating the integral of the density function obtained from a kernel estimate.

However, since a KDE smooths binned data, and the data needs to be converted back into a binned format for the Riemann sums, the KDE can be ignored altogether and instead the histogram can be used to create these sums.

However the probability is constrained such that the sum of all probabilities is equal to one. By normalising the histogram bins so that the values in all bins sum to one, it approximates the Riemann sums of the revenue density.

Due to the smoothing nature of a KDE, the probabilities of revenues just outside the boundaries of previously observed values may be non-zero. For example, in Figure 4.3, the maximum observed trip revenue is around \$20, however if a value slightly higher than \$20 was input into the function returned by the KDE, it may be above zero, although minuscule. If the probability of this boundary was higher, values just outside the boundary may be slightly higher as well, as a smoothing estimate struggles to capture sharp changes. An estimate using histogram normalisation will not have this aspect, as each probability is calculated from histogram bins. Since a trip revenue higher than \$20 has never been observed, the probability will be exactly zero.

A KDE approximation will allow for the likelihood of trips generating revenue higher than what has been observed previously, while histogram normalisation will not. Since it is not impossible for trips to generate above \$20, this makes a KDE approximation more appropriate at estimating revenue patterns. However, as can be seen in 4.3 the probability of trips generating above $\sim$ \$18 is infinitesimally small, and as such it is highly unlikely that this will affect results. Histogram normalisation reduces computation time, which allows simulations to be run multiple times easily for a more substantial amount of results, which is much more useful in practice.

4.5 Discussion

In this chapter, an equation to apply a value to individual scooters based on battery level and location was developed, which will be used to rank scooters in order of battery replacement priority. It was further simplified into a sums of the products of three separable parts. These parts were the demand, likelihood of the given scooter being chosen, and revenue increase expectation.

As they are separable, future work that improves the estimations of those three parts could replace the original estimates for a more reliable priority measure. For instance, the likelihood of a scooter being chosen is assumed to be random, however information such as placement, visibility, and user patterns could be included for an informed estimate. The demand approximated in this chapter is also very simplistic, and in the next chapter it is found that demand is the most influential variable on priority. A parameterised approximation of the demand density would allow for demands higher than what has been seen before to be considered. More future work involving demand exists such as demands at different days of the week or times of the year.

With the priority metric developed, the effectiveness needs to be observed to determine if using this metric results in more revenue as expected. This is done in the next chapter, where simulations are run using different methods of replacement priority.

5. SIMULATION-BASED TESTING

Chapter 4 derived multiple algorithms that could be used to provide a priority metric for each scooter at a given time given their current battery level and replacement. Ideally, the KDE filtering method of expected revenue increase, which removes any assumption about user behaviour and uses kernel density filtering to account for how revenue patterns change with battery percentage, should best reflect the expected behaviour most accurately. However, it is not proven that there is no pattern for user behaviour in terms of scooter selection, and as such there is the potential for the Revenue Cap method to provide more effective values for deciding which scooters to recharge first.

Not only do individual methods need to be compared against each other, it also needs to be shown whether any of these methods are better than the methods currently being used by companies such as Beam. As such, many comparisons need to be made to decide the most effective decision-making process.

These comparisons are achieved by simulation. Data on the historical availability of scooters can be used to create a simulated supply with scooter locations and number that reflects what is commonly provided. Then, data on past trips can be used to simulate trips with the relevant information: origin, destination, battery loss and revenue. If a trip were to occur in a location with no scooters available (due to lack of supply or charge), this is recorded and at the end of the simulation the number of trips missed and the total revenue lost over that time can be estimated.

The aim of these simulations is to determine which method results in the least loss of revenue.

Throughout this chapter, “simulate” always refers to the same process of random selection described here. For all relevant information, the data provided gives a historical report of how that information is distributed at different times and locations. For instance, we have information on all trips, and know the number of trips at each time on each day and where those trips occur. If a single trip is simulated, a trip is randomly selected from the list of trips that have previously occurred at that hour of the day, and that chosen trip is assumed to occur, or at least be in demand, at the location that trip started in. This is a simplistic method of simulation, however the motive behind this method is that events that occur more frequently will take up more of the dataset, and hence be selected more often during the random process.

However, before trips can be simulated, the environment of scooters and employee behaviour needs to be simulated.

5.1 Environment Simulation

In this chapter, the "environment" is defined as any behaviour of scooters that takes place without any user influence. Simulating trips allows for modelling of revenue gain with demand patterns considered, however the environment needs to be simulated to account for availability in terms of both scooter supply and battery percentage. When demand is simulated against availability, the effect of battery replacement can be explored.

Since users only act on scooters during trips, the environment therefore refers to four things:

1. Supply of scooters;

2. Replacement of batteries;

3. Maintenance when scooters need repair; and,

4. Redistribution to match explored demand.

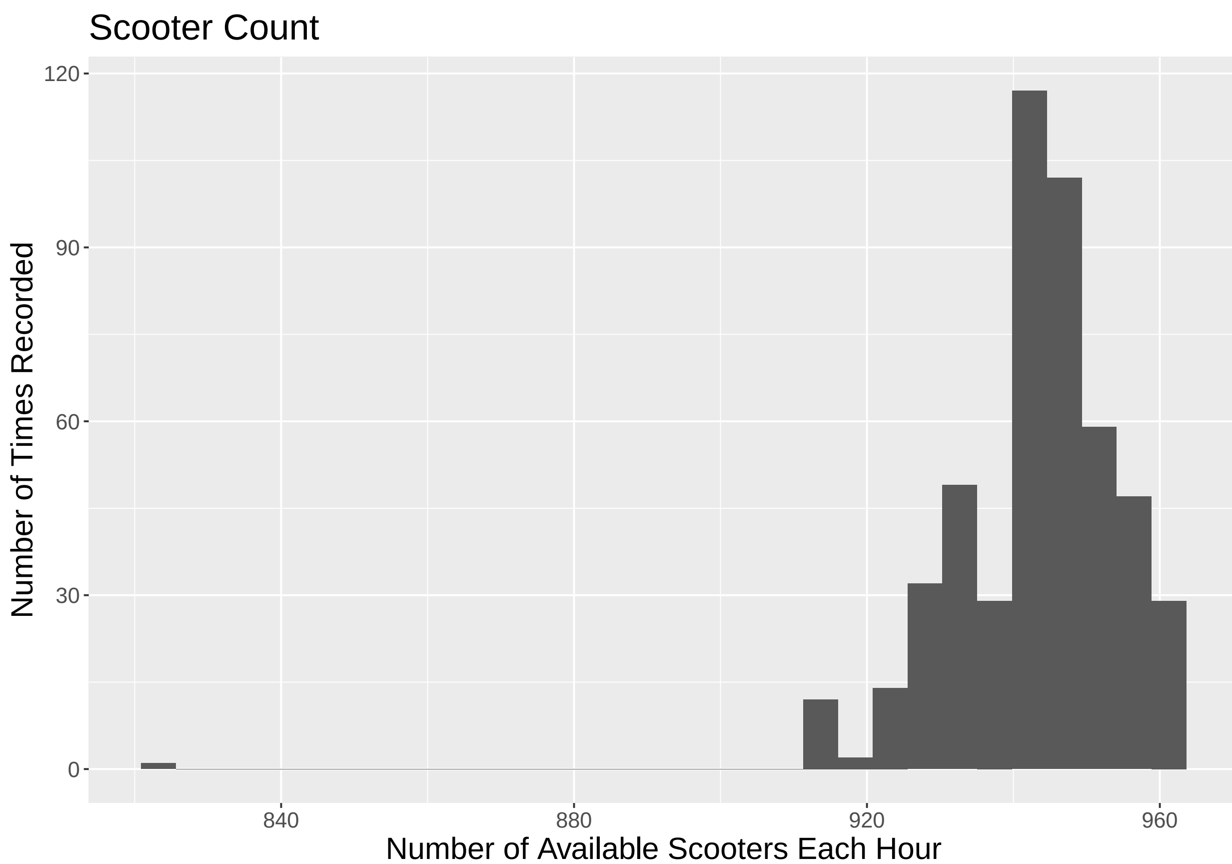


Fig. 5.1: Number of available scooters at each hour throughout April to July 2021.

As mentioned in Chapter 2, supply can be ascertained from the “free bike” dataset, which provides the locations of all available scooters at the current time. By frequently requesting this dataset and storing the results, historical data on scooter distribution at a variety of times can be generated. At the start of the simulation, the initial supply of scooters in each location is simulated from this dataset by choosing a number of scooters to be present, and randomly choosing locations for that many scooters from the data. From the data gathered, the number of total scooters was consistently 2148, however the number of scooters that were not recorded as “disabled” was on average 940, as in Figure 5.1. Since this number fluctuates, it could also be simulated. However, since it does not change once the simulation has started, it will always be set to 940 so that simulations can be effectively compared.

Replacement refers to the process of replacing batteries. At each hour of the simulation, scooters have their batteries replaced with the number of scooters and their locations decided based on previous observations of replacement.

Battery replacement is difficult to exactly simulate, as replacement events are not recorded in the provided data. As outlined in Chapter 2, information on battery percentage is only available at the start and end of trips. Using this, a replacement is assumed to occur when a trip occurs involving a scooter that has a much higher battery level than at the end of its previous trip. This means that the exact time and increase in battery percentage is unknown, as battery level will decrease gradually when a scooter is stationary. However, the number of recharges that occur each day is known, and in cases where the overall time between trips is low the number of recharges at each hour is also known. This is useful for simulation, as the number of recharges per hour is needed for utilising the priority algorithms because the number of scooters to select needs to be specified.

The location of battery replacement is known, as it usually does not change between trips unless redistribution occurs. This is useful to understand locations that Beam usually prioritise themselves, which allows for a simulation of Beam's current method that can be compared against the algorithms developed in this research.

Redistribution refers to the process of employees changing the location of scooters that may be in low demand areas, in an attempt to satisfy potential demand in areas with low supply. There is no data on redistribution, however this can sometimes be observed in trip data when scooters start trips in locations that they did not end their previous trips in. However, this would not capture all redistributions as a scooter could be moved and then still not be used for a trip.

For the sake of observing revenue as a result of battery replacement, however, redistribution is not relevant. While it could reduce the amount of trips lost due to lack of supply, it is difficult to simulate without sufficient data and attempting to redistribute effectively without this data is a problem in itself. Hence, the simulations in this research attempt to minimise revenue loss in a redistribution-less system. As such, redistributions are not included in the simulations. This will most likely result in a high revenue loss, as scooters will enter low demand areas and rarely leave them, but for the sake of comparison this is not of great concern.

Maintenance events are recorded in the same way as trips however they are not simulated as the effect that maintenance has on the total revenue during the simulation is small as it very rarely results in a change in battery percentage. As the scooter is moved to the warehouse during maintenance, it does result in a change in location when it is dropped off, however this is a similar affect to redistribution.

5.2 Trip Simulation

To estimate the revenue of a system that uses a specific decision-making process in terms of battery replacement, trips are simulated at each hour. The number of trips that occur each hour, the distribution of these trips throughout the Adelaide area, and the destinations of each of these trips has been shown in Chapter 2 to vary greatly and are difficult to predict. As such, the simulation code does not attempt select specific trips that are likely to occur, but instead randomly chooses trips from the dataset and assumes that they occur. This choice is done with replacement, as a trip occurring does not prevent a similar trip from occurring in the same location.

Random selection should still reflect demand patterns. Trips that occur often, appear often in the dataset and have a higher chance of being randomly chosen. With enough accurate and clean data, this should sufficiently represent the user patterns in the observed area.

The two-step process to simulating trips is as follows, at each hour:

1. Simulate the number of trips that occur in the hour; and,
2. Select that many trips from the dataset.

With enough data, it is possible to estimate patterns that occur throughout the day. At each hour of the simulation, by filtering the dataset to only include trips that have started at that hour of the day, any possible changes in demand in terms of starting location will be reflected in the results. Whether or not this is done depends on which demand equation is chosen for the priority metric from Chapter 4. If the metric takes into account the current time, the simulation needs to reflect this.

5.3 Simulation Process

The supply is simulated once at the start of the process, while battery replacements and trips are simulated at each step, which represents one hour. All of the trip information mentioned in this research is relevant for the simulation. This is because the status of each scooter is changed by the results of the trip.

Since the randomly selected trip comes from the dataset, all of the relevant information is known; where the trip starts and ends, how much battery level is lost, and how much revenue was obtained from the trip.

Upon selection of a trip, the location of the start of the trip is observed, and if the simulation contains supply at that location we can observe if any of the scooters have enough battery percentage to satisfy the trip. We take the battery lost when that trip originally occurred, and if none of the scooters have a battery percentage higher than that loss, it does not occur and the trip is recorded as missed. If any of them do, a scooter from that group is randomly chosen and the trip is satisfied, generating the same amount of revenue as it did originally.

The selected scooter is then moved to the destination of that trip, and its battery percentage is decreased by the amount of battery lost in that trip.

It was shown in Chapter 3 that battery loss is not consistent or strictly proportional to trip duration or distance. Trips that may start and end in the same location and take the same amount of time may experience a significantly different loss of battery level. Randomly selecting trips should take account of this variation by not attempting to calculate a single value of loss given other trip information.

Since scooters change location during trips, the environment of scooters is affected by trips. Replacements are randomly chosen just like trips, but unlike trips they cannot be "missed". As such the decision to replace batteries is affected by the current supply. In the case where battery replacement is randomly chosen to reflect the current decision-making process employed by Beam, that random selection only occurs on a dataset of replacements in locations where scooters are currently available. The priority metric takes into account the current battery level of available scooters, and as such will always be based on the current supply.

The simulation is run for a given number of steps, and the resulting revenue gained and lost can be summed and observed.

A potential consequence of this, is that the number of trips simulated to occur per hour will be slightly less than what is actually observed. It is expected that a large number of trips throughout the simulation will not be able to occur due to lack of availability, as such the number of trips to occur each hour is less than the number selected, as not all of them are satisfied. However, this number is based off of the number of satisfied trips, and if missed trips were to be accounted for in reality, the number of trips demanded each hour will be higher than what is calculated here. Since missed trips are not visible in the data, it is impossible at this time to actually estimate this number, and as such the demand value estimated for this research is the closest approximation with the data provided.

5.3.1 Data Sampling

Five datasets are obtained to sample from during simulation:

1. Trips - A dataset of trips including start location, end location, start time (in hours), battery lost and revenue;

2. Trip Rates - A dataset with rows for each hour of every day, and the number of trips observed in that hour - to sample demand;

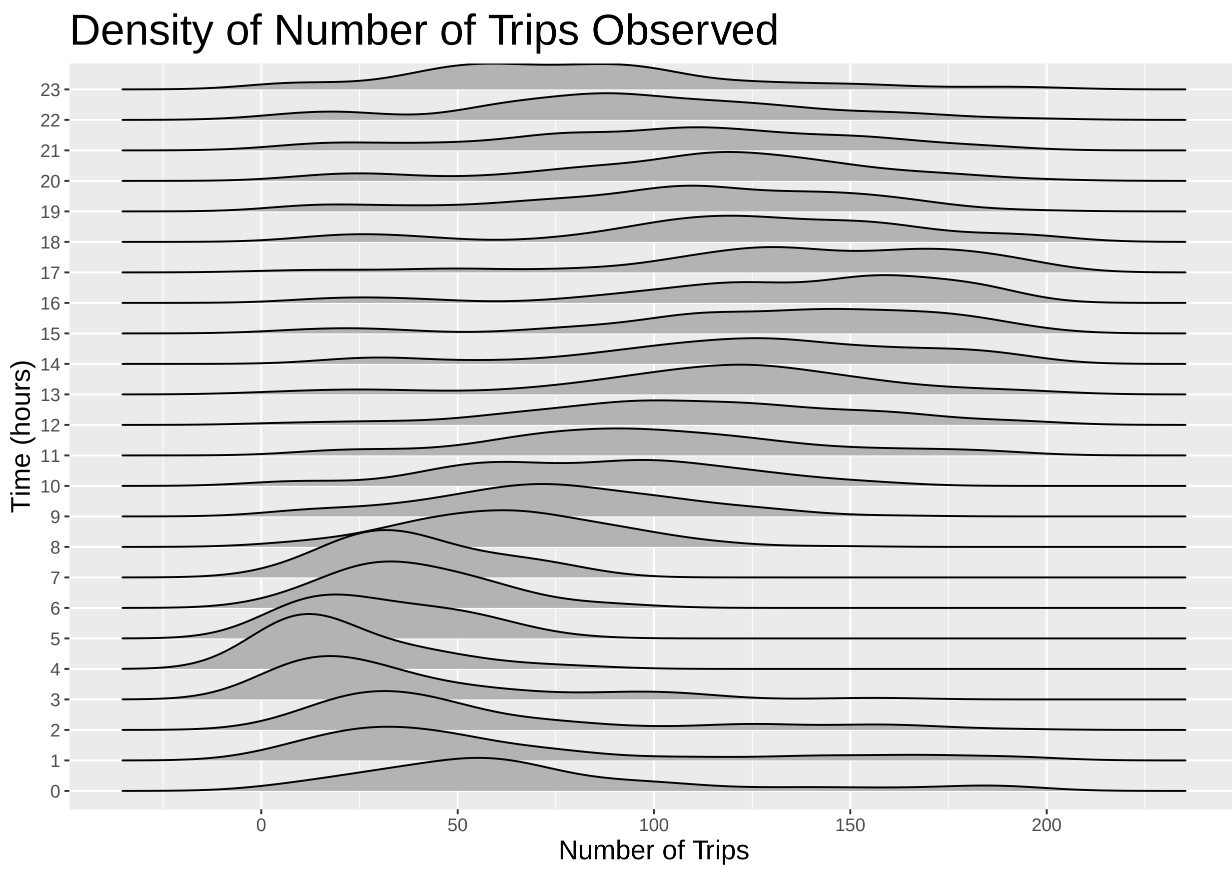


Fig. 5.2: Densities of observed number of trips for each hour of the day.

3. Loss Rates - A dataset that includes every observed period where a scooter is stationary, the time between the end of a trip and the start of that scooter's next trip, and the amount of battery lost in that timeframe;

4. Replacements - A dataset of observed battery replacements including the location and the time in hours that they occur; and,

5. Locations - A dataset containing the locations of scooters at every hour.

At each hour, the Trip Rates dataset is sampled from to estimate the demand. A single row that has the same hour at the current time is randomly chosen, and the corresponding number of trips is presumed to occur.

That number of trips is then sampled from the Trips dataset, and scooters are chosen and moved around as explained previously.

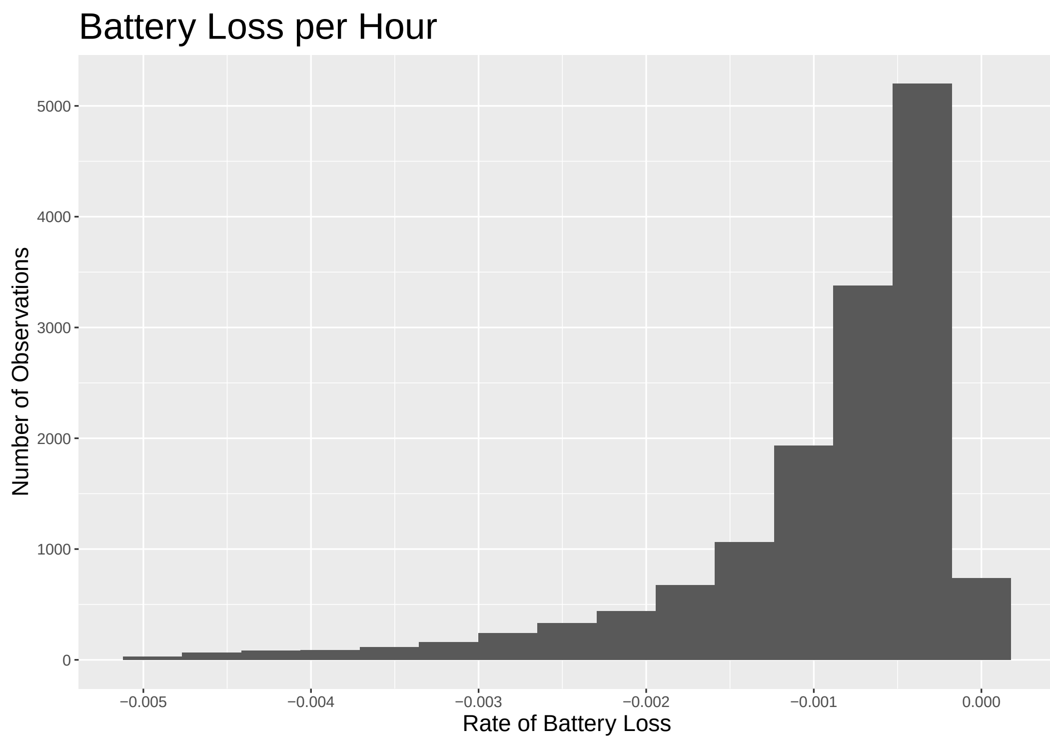


Fig. 5.3: A histogram of the battery percentage loss per hour. The mode for this rate is as small as possible without being zero.

The Loss Rates dataset provides observed rates of battery loss while stationary. Dividing the battery lost by the number of hours stationary results in battery lost per hour. As seen in Figure 5.3, this rate is very close to 0%/hr and is negligible. However for long running simulations this may still be useful. The variance of this loss is taken into account, and is hence also sampled. At every hour, for every scooter, this dataset is sampled and the scooter's battery percentage decreases by whatever rate is sampled.

The Replacements dataset is used similarly to the first two datasets. To estimate the number of battery replacements each hour, this dataset is summarised to return the number of replacements observed at each hour of every day. Then, a row is sampled with the current hour, and the corresponding number of replacements occur in the simulation. This is done for all simulations, however this dataset is then used again when simulating the process Beam uses to replace batteries. Once the number of replacements is obtained, the individual replacements are sampled to estimate the decisions Beam employees make.

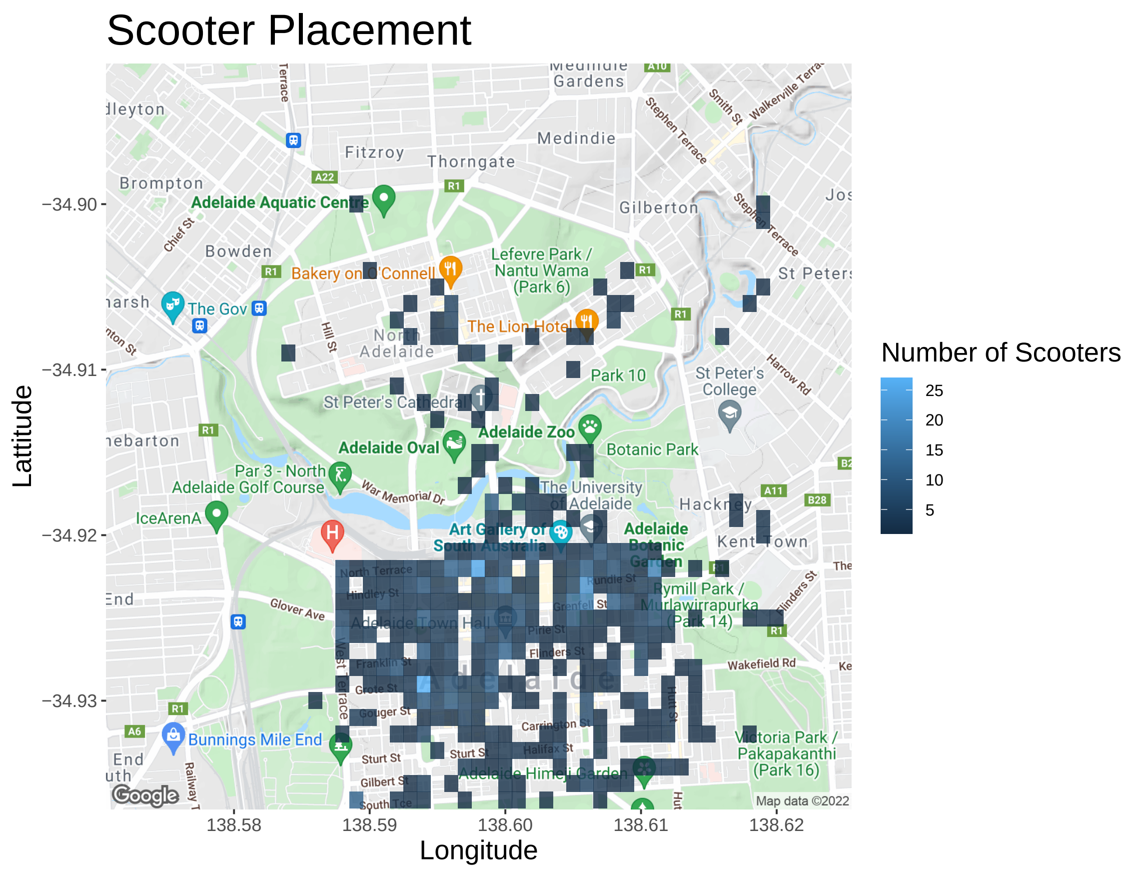


Fig. 5.4: Example of the initial distribution of scooters.

Sampling the previous battery replacements will not entirely reflect the decision-making process employed by Beam, as they will take into account scooters' current battery level before replacements are chosen. However, since there is no data available for the battery level of all scooters at a given time, it is impossible to ascertain why specific scooters were chosen for battery replacement. Sampling replacements should at least highlight specific areas that employees tend to favour when selecting scooters.

At the start of the simulation, the Locations dataset is sampled from to obtain the initial locations of scooters.

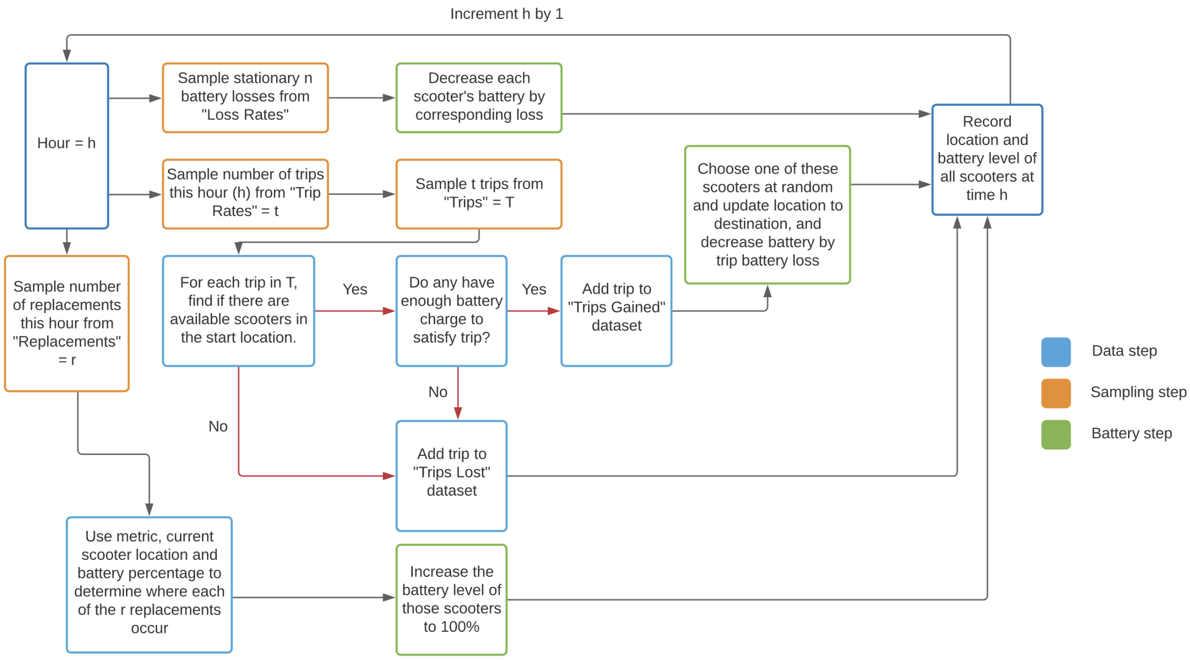


Fig. 5.5: The basic simulation procedure.

5.4 *Simulations*

The simulations are run for each potential method of battery replacement. This includes the priority metrics developed in Chapter 4 and method assumed to be currently employed by Beam. The simulations are run multiple times to obtain the variance of revenue generated, however for the comparisons to be meaningful, they need to be made on the same run. What this means is that when the results of two methods are compared, the results need to be from simulations that include the same initial scooter replacement, the same trips, and the same number of replacements at each hour. This way, it can be observed which method was able to better satisfy the demand of that specific simulation. If methods were compared using the results of simulations on different samples, it would be unclear if the results were different due to the method used or if the sampled trips did not require as many battery replacements.

The overall procedure for simulations can be seen in Figure 5.5.

Storing the trips lost and gained in their own datasets allows a variety of observations to

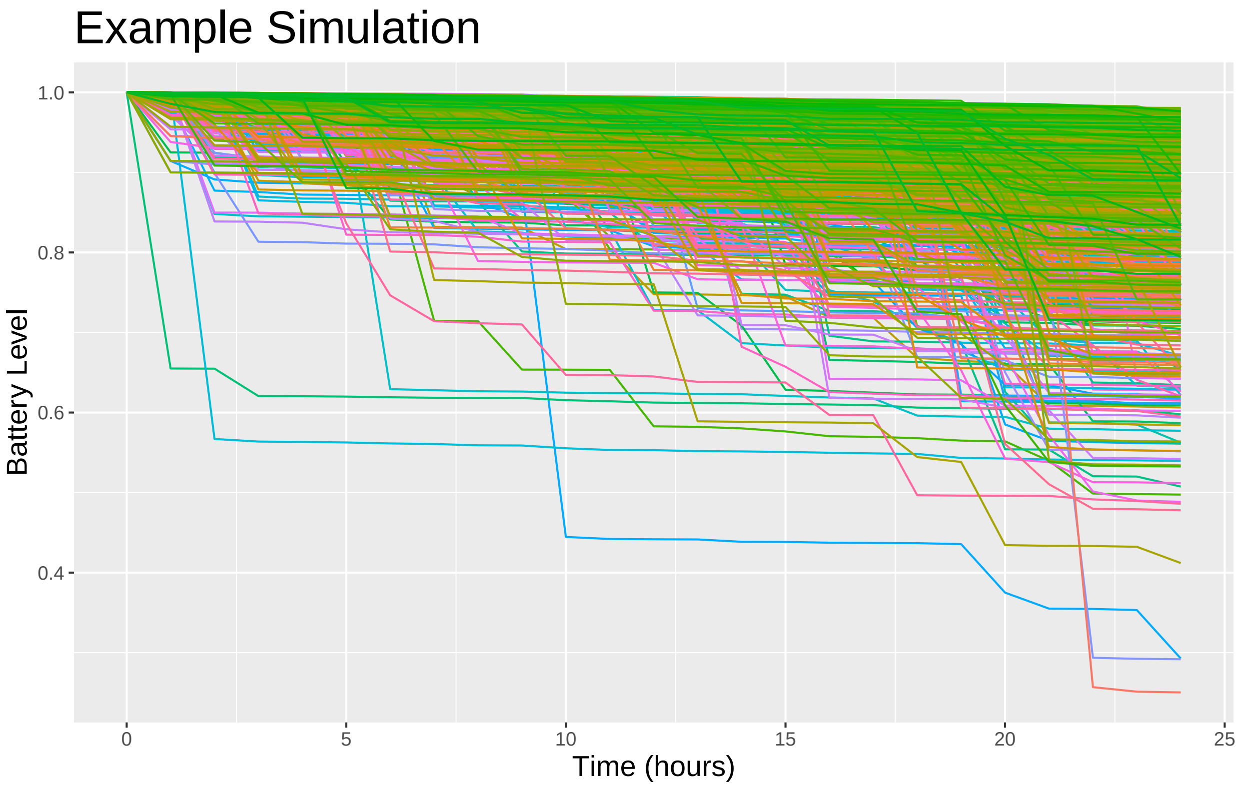


Fig. 5.6: Battery levels of 1200 simulated scooters over 24 hours.

be made. Since the trips contain the revenue generated, the overall revenue lost and gained can be calculated. Locations that frequently result in missed trips can be observed as well.

Replacements are made with the current scooter locations in mind. This is because replacements cannot be "missed" like trips can. For instance, the priority metric is calculated for all locations where scooters are and multiple replacements are considered for locations where there are multiple scooters. When replacements are sampled from the replacement dataset, they are sampled only from locations where scooters currently take place. However, like trips, replacements are stored in a dataset as well to observe where these more frequently occur.

An example of the simulation results is presented in Figure 5.6. Each colour represents a scooter, and its battery level throughout the day can be tracked. As there are hundreds of scooters, it is difficult to observe individual scooters, but the overall pattern is clear. All of these scooters begin at full charge (100% battery), and only 6 of them drop below 50%

within one day. A considerable number of scooters do not experience any trips and stay above 90% battery. For the rest of the simulations, a simulation like this is run initially to provide a spread of battery percentages in the simulation environment.

The few scooters that result with less than 50% battery provide incentive for the analysis of 'journeys' as discussed in Chapter 4. These scooters have experienced an above average number of trips in one single day, and if these scooters tend to start the day in similar locations, then those are the ones that should be prioritised in terms of battery replacement.

5.4.1 Priority Metric Simulation

At every hour of the simulation, the priority metric from Chapter 4 is applied to every location in the simulation environment in which scooters are located.

Recall that the priority metric is defined as:

$$I_{i,j}(b) = \sum_{z=1}^{m} P_{i,j}(z) Q_i(z,b) R_{i,j}(b),$$

where $P_{i,j}$ is the chosen demand estimate, in zone i and hour j, $R_{i,j}(b)$ is the chosen revenue increase estimate when b batteries are replaced, and y_i is the number of scooters in zone i. $Q_i(z,b)$ is the probability that a recharged scooter is chosen if z scooters are hired and b scooters are replaced. Q depends on the assumptions made, if the most-charged scooter is chosen first or if the scooter is chosen at random.

At hour j simulation iterates through each i, and for every possible number of replacements $b = (1, ..., y_i)$, $I_{i,j}(b)$ is calculated.

At the start of each hour the number of replacements that will occur is sampled. Let this number be c, and the c replacements need to be decided from the list of values $I_{i,j}(b)$ for all possible i and b. This creates an optimisation problem, as replacing the batteries of many scooters in one location may be profitable, but potentially not as profitable as replacing a few batteries in a variety of locations. This is similar to the knapsack problem, in which there is limited space (c), and each item has a different value (revenue increase), alongside with different sizes (number of replacements b).

The solution chosen here is a simple 'greedy' solution, in which the revenue increase is divided by the number of replacements b, to provide a 'rate' of revenue increase by replacement. Since there are multiple values of I for each zone, as there are multiple cases of battery replacement, for each i, the value of b that has the highest rate for that zone is chosen. Then, each zone is ordered in descending order of rate, and then the top replacements of which each the sum of each b is equal to c are the ones performed that hour.

5.4.2 Linear Model Method of Priority

Priority is calculated with the following equation from Chapter 4:

$$I^L_{i,j}(b) = \sum_{z=1}^{m_i} P_{i,j}(z) \frac{\sum_{k=0}^{b-1} C^{y_i-b}_{z-b+k}}{C^{y_i}_z} \left[\sum_{l=0}^{b-1} \int_{\text{cap}_{(y_i-l),i}}^{\infty} x f_{i,j}(x) dx \right], \tag{5.1}$$

where $\text{cap}_{(y_i-l),i}$ is the limit on revenue provided by the battery of scooter $y_i - l$ in Zone i.

As mentioned in the beginning of this section, these simulations are identical except for the method used to decide which scooter batteries are replaced at each hour. At each hour, the trips are sampled, alongside the number of replacements, and these are applied to both simulations.

The simulations presented in Figure 5.7 experienced 16079 simulated trips over 169 hours; 9881 trips were satisfied in the simulation that uses the priority metric, while 9765 trips were satisfied by the simulation that sampled replacements from the data.

Summing the revenue from both simulations, the data replacements resulted in \$54356.05 of revenue over the week, while the priority replacements resulted in \$54897.29. The average revenue per trip was \$5.57 with the data replacements, and \$5.55 with the priority replacements.

Figure 5.7 shows the quantiles for the battery levels of all scooters in each simulation over

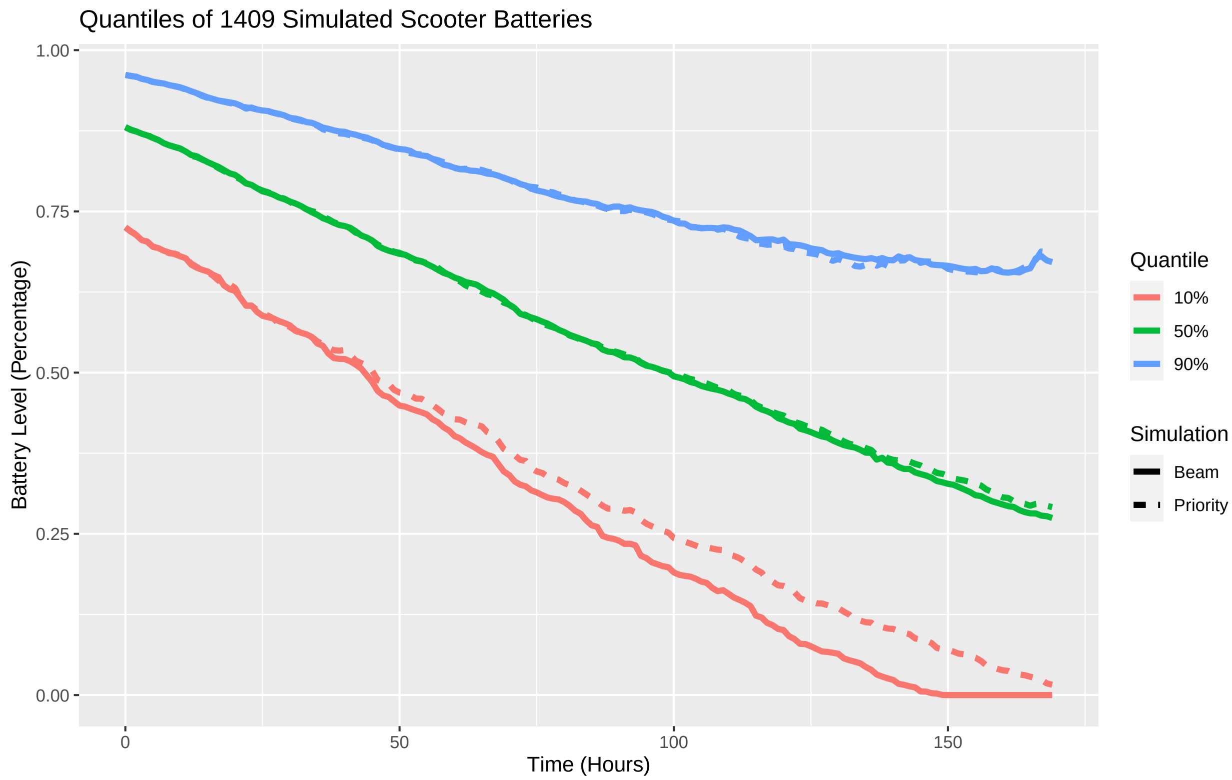


Fig. 5.7: Scooter batteries simulated over one week

the one week period. Both simulations continue from the one day simulation shown in Figure 5. While the battery levels of scooters in the 90% quantile are similar for both simulations, the 10% quantile drops much more quickly in the "Beam simulation", which attempted to simulate the decisions made by Beam employees, than in the "priority simulation", which uses the priority metric to decide where replacements occur. The priority simulation was able to satisfy over 100 more trips than the Beam simulation, most likely due to scooters not falling to low battery level quickly.

100 extra trips is comparatively small to the over 9000 trips lost over the entire system in both simulations, however it is important to note that trips are lost if there is no supply in that location, and that can occur if the scooters are low on battery, or if no scooters are there in the first place. Presented here is a simulation of an environment with no redistributions, so scooters that end in low demand zones are never moved to high demand zones, so there is potential that physical supply does not reach demand often. Hence, many of these trips lost

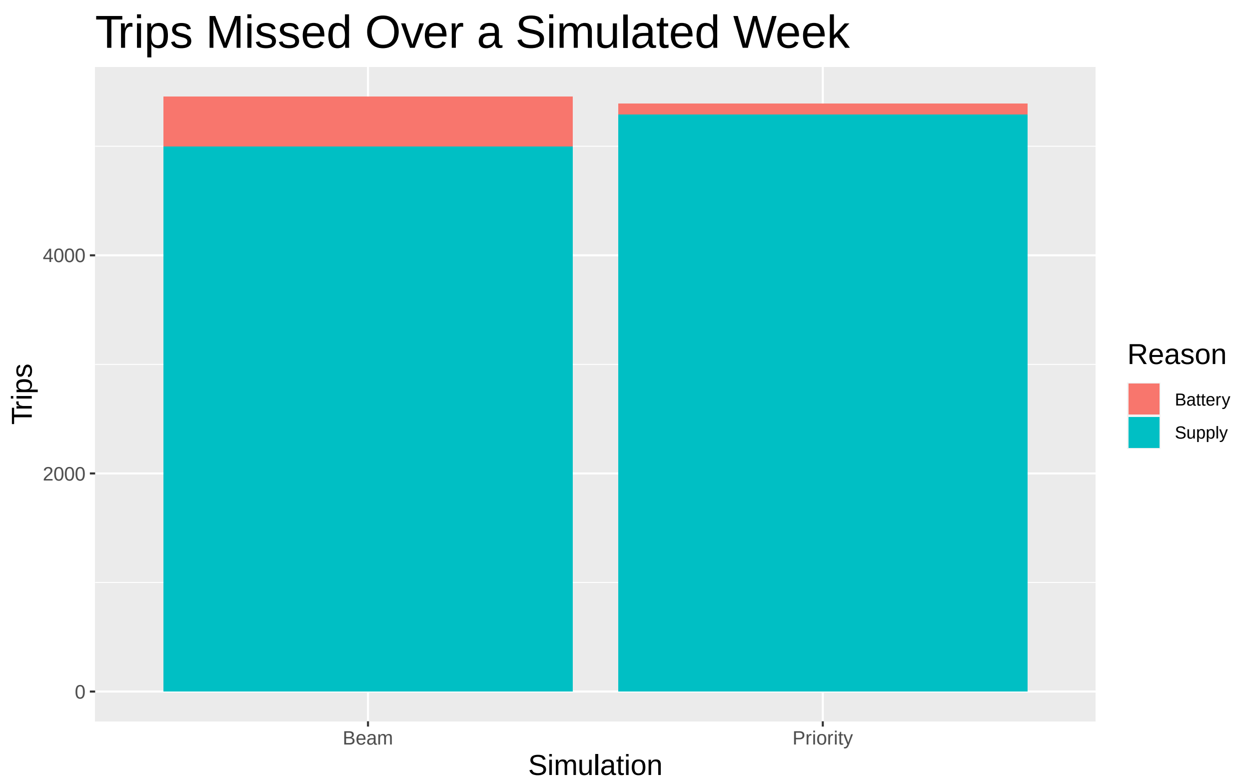


Fig. 5.8: Scooter batteries simulated over one week.

are the result of lack of supply and will occur in both simulations regardless of how batteries are replaced.

The results of another test are shown in Figure 5.8. Again, the total number of trips lost in the priority simulation are slightly less than those of the Beam simulation. However, the amount lost due to nearby scooters not having enough battery is significantly smaller in the priority simulation. Interestingly, the amount of trips lost due to lack of supply is higher in the priority simulation, potentially due to scooters being more often able to satisfy longer trips that take scooters out of high demand areas.

Redistributions are also the cause of another consideration with these results, and that is that the battery replacements sampled for the Beam simulation are based on a system that does have redistributions. Hence, the Beam simulation will focus on replacing batteries in high demand areas as it expects scooters to be redistributed into those areas. The Beam simulation does not take into account the battery level of scooters in the environment, and this may be why the priority simulation is more effective at preventing scooters from reaching low battery.

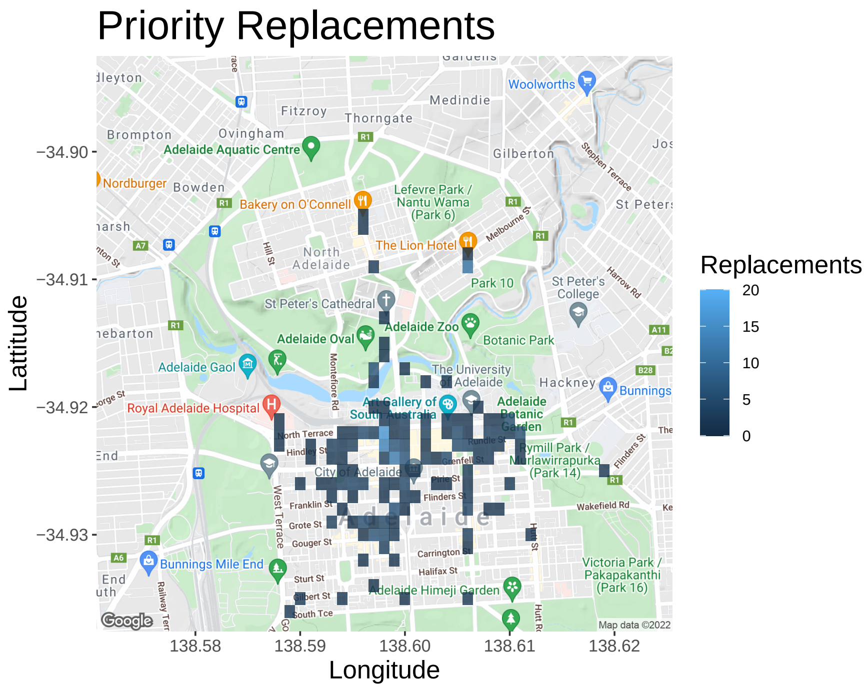
Also, neither simulation is able to effectively manage the demand of scooters in relation to battery level. As the average battery level of all scooters drops throughout both simulations, this means that at the end of the week, a large proportion of scooters are unavailable. It appears that scooter batteries are being depleted faster than they are being replaced. Assuming this does not occur in reality, it means that the number of recharges sampled is not high enough. This underestimate could be a result of the difficulty to ascertain the exact number of replacements from the data as a replacement can only be observed if that scooter is then used for a trip.

More simulations were run, and the priority simulation consistently had a higher resulting revenue and fewer trips missed. However, the difference is always very small. It should be expected that the Beam simulation will under-perform as redistributions are not occurring, so for these results to be in favour of the priority metric the difference would need to be far greater.

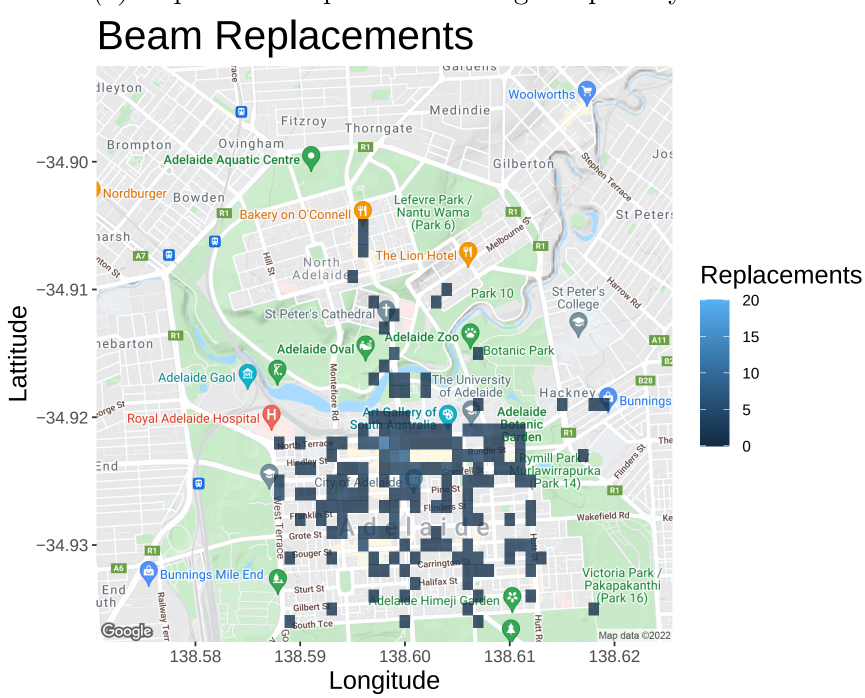
It is also possible to observe the areas in which replacements occur most frequently in either simulation. Figure 5.9 highlights the key difference between the two processes. The Beam simulation experiences the most replacements around the train station/mall area, again this is because the process operates under the assumption that scooters are being redistributed into high-demand areas.

The priority metric results in the most replacements in the southern area of the city. In fact almost 20 replacements occur there. This could likely be due to the fact that trips likely end in that location, resulting in more low-battery scooters. However in order for the metric to continuously suggest replacements in this location, scooters also need to be picked up there as well to make way for other scooters to be in need of replacement.

Considering that the focus of replacements is drastically different for the two methods, it is interesting that the difference in trips lost is so small. This implies either the location



(a) Replacements performed using the priority metric.



(b) Replacements performed using sampling.

Fig. 5.9: Battery replacements occurring over a week of simulation

of replacements has very little effect on the loss of trips, or that almost all of the trips lost are due to the lack of redistribution.

Observing the areas in which trips are missed in Figure 5.10, it can be seen that the distribution of lost trips is very similar in both simulations. The most trips are missed in high-demand locations, supporting the claim that without redistributions, many trips are lost due to lack of supply. It is important to note that replacements only occur in the beam simulation if there are scooters to recharge, so there had to have been supply when those batteries were replaced. What this implies is that focusing replacements in high-demand areas may not be as valuable as in other areas, as the priority simulation was able to satisfy not just the same amount of trips, but more, by focusing on area with significantly lower demand.

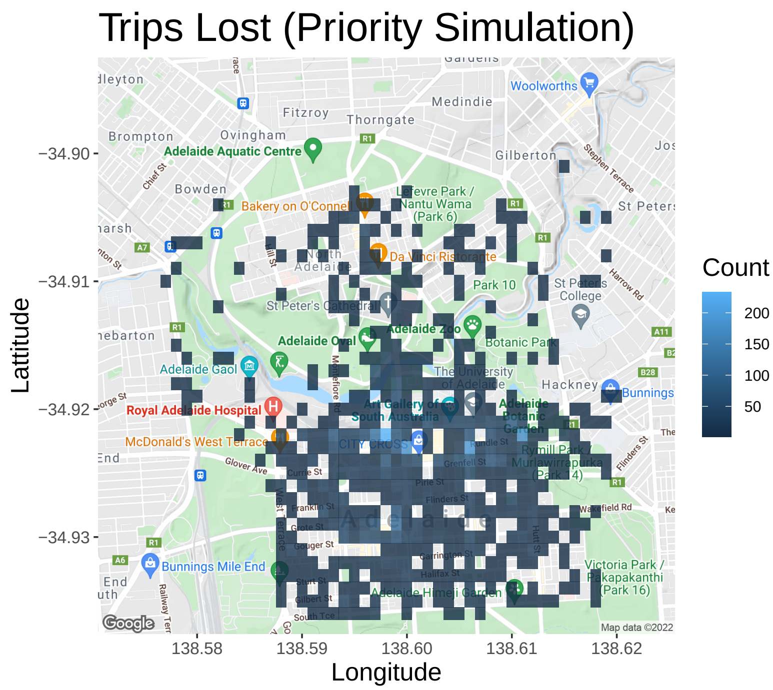
5.4.3 Kernel Density Estimated Method of Priority

While the results in Section 5.4.2 were in favour of the priority metric algorithm, the other metrics discussed in Chapter 4 are analysed here to see if their effect on revenue is greater. As mentioned in Chapter 3, calculations of priority using Kernel Density Filtering have fewer non-zero values. Since the priority metric is used as a comparison, this occurrence will only affect the results if there are a large number of replacements and very small estimated revenue increase. In this case, the algorithm will have to choose between a large number of zero values, which should not result in an informed decision. However, KDE filtering may be more accurate as it better represents the apparent randomness of battery loss rate during use by not assuming the relation between loss and time is linear.

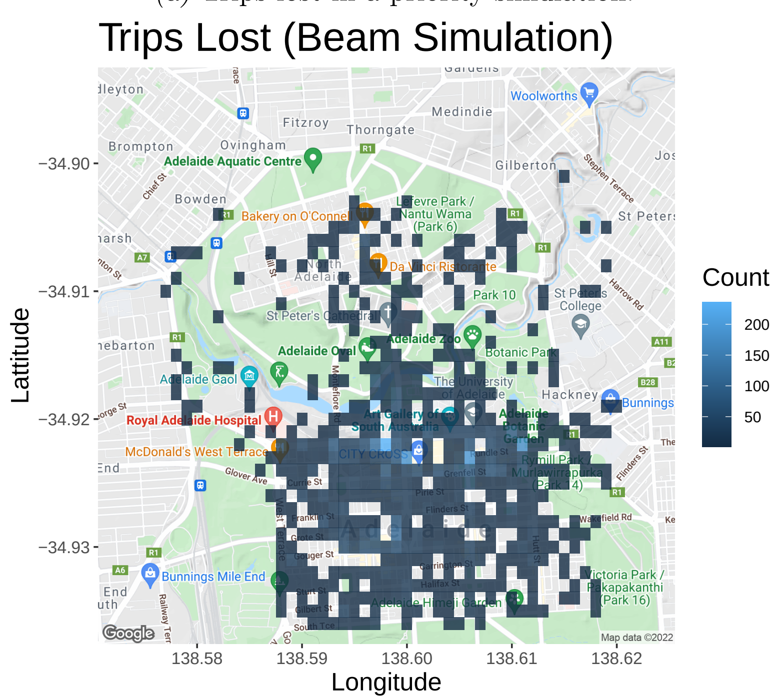
Priority is calculated with the following equation from Chapter 4:

$$I_{i,j}^{K}(b) = \sum_{z=1}^{m_i} P_{i,j}(z) \frac{\sum_{k=0}^{b-1} C_{z-b+k}^{y_i-b}}{C_z^{y_i}} \sum_{l=0}^{b-1} \left[\int_0^\infty x f_{i,j}(x)dx - \int_0^\infty x f_{i,j,y_i-l}(x)dx \right]. \tag{5.2}$$

where $\int_0^\infty x f_{i,j,l}(x)dx$ is the expected revenue from trips that use less battery than scooter



(a) Trips lost in a priority simulation.



(b) Trips lost in a beam simulation.

Fig. 5.10: Total trips lost during a week of simulation.

$y_i - l$ currently has.

Presented in Figure 5.11 is a comparison of the values of $I_{i,j}(1)$ when each location has one scooter at 1% battery, at any time of day. There are a few gaps, implying locations where priority is 0, due to a lack of data or incredibly low demand in those locations. The areas where priority is non-zero are almost identical for both methods, however the values of priority are generally lower when obtained using KDE filtering.

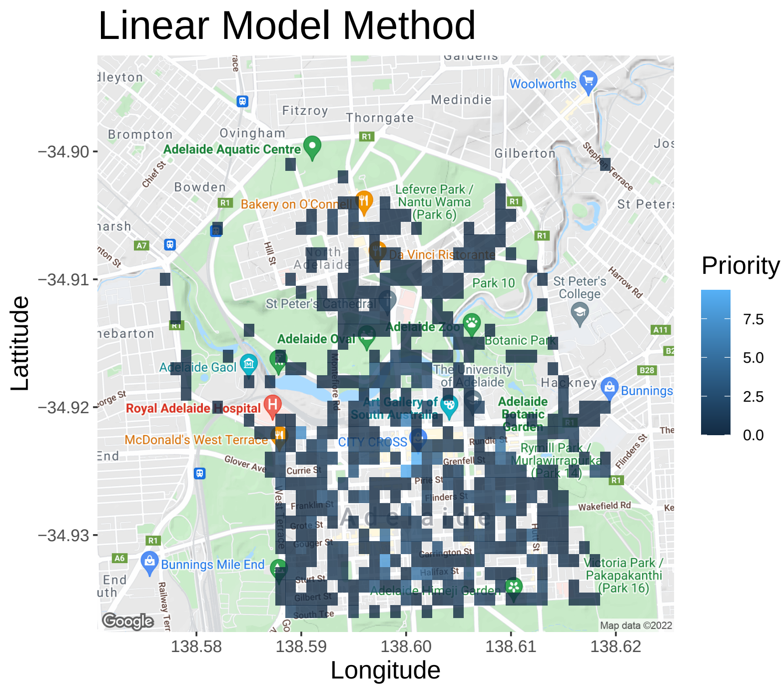
Importantly, the areas of high priority follow those with high demand. These areas being the Rundle Mall area (in the very centre of the maps) and the business sector at the bottom, to the left of the centre. Interestingly, both methods favour slightly to the North of the Mall (138.600,-34.920), however only the linear model method favours the entrance to the Mall (138.600,-34.924), which is usually where demand in that area is focused. Neither model appears to give high priority to the train station, however, which has some of the highest demand.

5.5 Journey Simulation

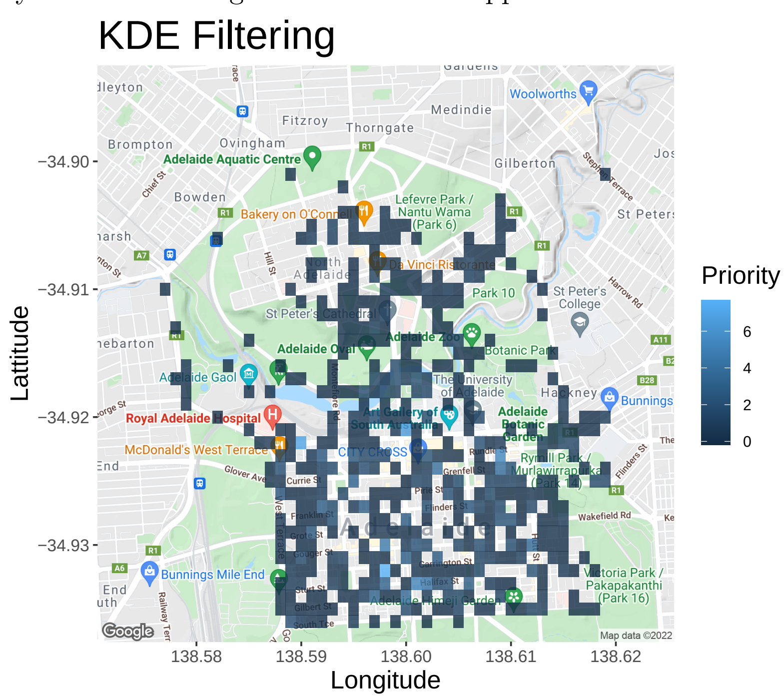
A number of findings from previous sections of this chapter motivate the use of 'journeys' to represent potential revenue gain over trips. Recall from Chapter 4 that a 'journey' refers to observing how a scooter is used beyond one trip, taking into account its drop-off location and subsequent trips by summing the battery lost and revenue gained 24 hours after the current trip finishes.

The motivation stems from the fact that the pattern of revenue from trips is relatively consistent throughout the environment, the mean revenue and likelihood of above average revenue is very similar, resulting in a priority metric that focuses primarily on the demand of locations. If the density of revenue is identical in each zone, it may as well not be considered, however it is desirable for a metric to take into account the possibility that an area may have low demand, but high revenue.

The trip densities imply that this is not the case, however they do not cover the po-



(a) Priority calculated using the linear model approximation of revenue increase.



(b) Priority calculated using KDE filtering.

Fig. 5.11: Priority when replacing one battery in a location with one scooter at 1% battery.

tentiality that certain locations may lead to trips that end in high demand areas, meaning that the scooter is more likely to experience multiple trips in a short time, generating more revenue. Considering that demand has been shown to be different depending on location, it should be expected that journey revenue densities can differ, as long as the distribution of dropoff locations changes depending on where the trip starts.

Recall that both priority metrics are in the form:

$$I_{i,j}(b) = \sum_{z=1}^{m} P_{i,j}(z) Q_i(z,b) R_i(b).$$

Since journeys observe the pattern after a trip starts, the demand, $P_{i,j}$, and the probability of being chosen $Q_i(z,b)$ are unchanged when journeys are analysed. It is only the revenue density, $f_{i,j}(x)$, that is used to calculate the expected revenue increase $R_i(b)$, that changes, as the revenue data $\mathbf{x}$ used to estimate $f_{i,j}$ has been modified. $R_i(b)$ will be larger, resulting in larger $I_{i,j}$ for non-zero $P_{i,j}$, however since I is used for comparison the increase in magnitude will have no effect on the results.

Simulations were run using journey revenue to calculate priority, and compared against a simulation were replacements were sampled from the data. Figure 5.12 shows a plot of the quantiles of all batteries over the one week period. As with previous tests, the results for both are very similar, however it can be observed that the lower quantile (10%) has a slightly more negative slope, as batteries with low percentage are not being replaced as often.

15183 trips were sampled in the simulation presented in Figure 5.12, 9214 of which were performed in the priority test. This is a rate of 60.6% of trips satisfied, which is actually slightly smaller than the 61.5% of trips satisfied of the priority simulation obtained in Section 5.4.2. This very small change implies that considering revenue in terms of journeys has minimal effect on the priority.

Figure 5.13 compares the four possible priority methods (linear model, KDE, and journey revenue), against the method assumed to be used by Beam by observing the distribution of

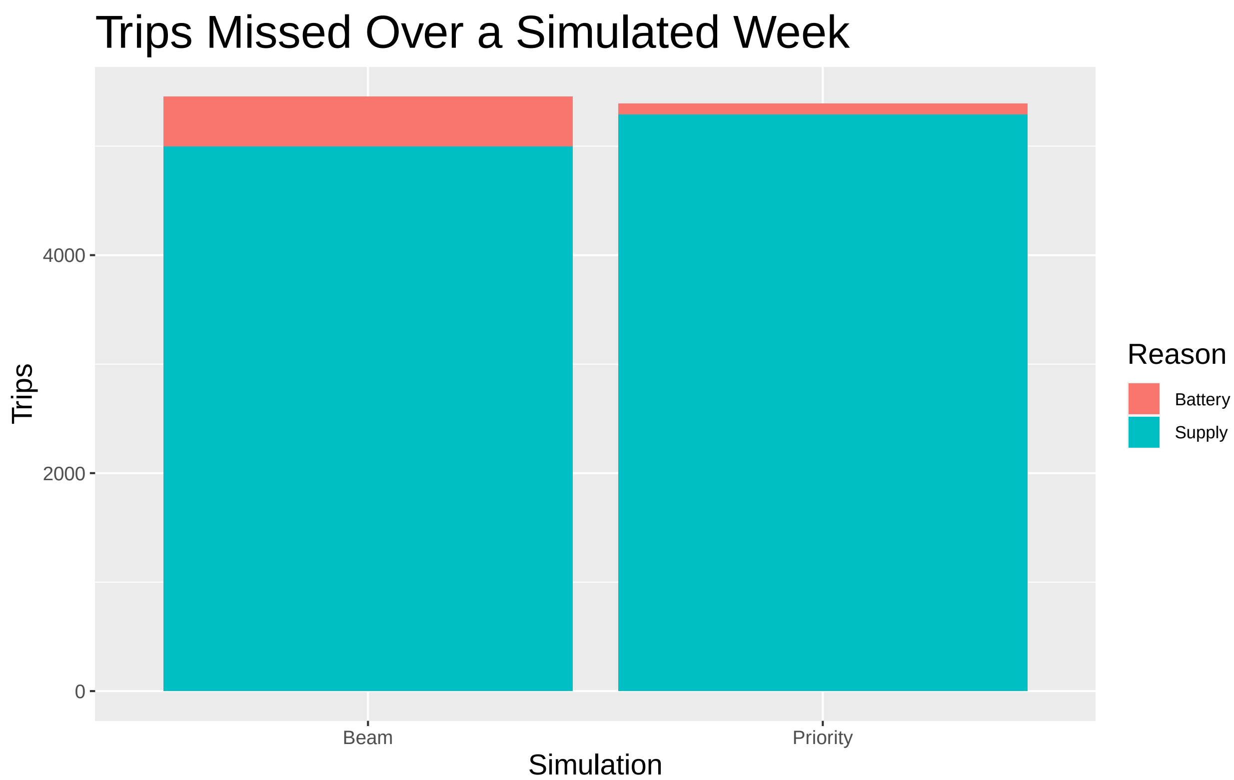


Fig. 5.12: A bar chart representing trips lost in simulations using Beam informed replacements and priority metric replacements, using the same samples. While the difference between total amount of trips lost is small, the priority metric saw far less trips lost as a result of insufficient battery. An explanation for why more trips were lost as a result of lack of supply could be that more trips satisfied, more scooters are taken to low demand areas.

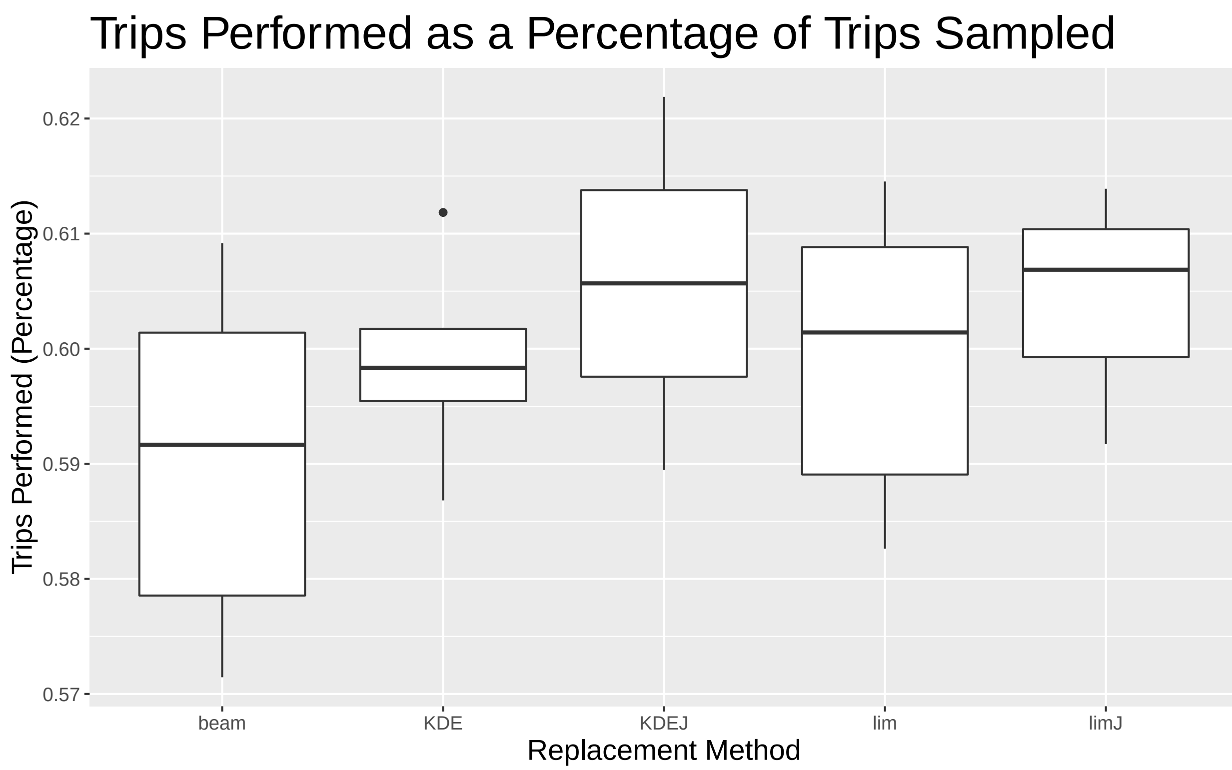


Fig. 5.13: Boxplot of trips performed as a percentage of trips sampled over a number of simulations.

percentages of trips satisfied over a number of simulations. Four simulations were run for each priority method, and a Beam simulation is obtained for each of these, which may be why the beam results have a higher deviation. It can be seen that the average percentage of trips performed is higher for every priority metric than replacement sampling. However, the average rate of trips performed is always within 59-61% no matter the method, which is a very small difference.

The methods that use journeys to calculate revenue (suffix J), do seem to be slightly more effective than the methods that do not. However again this difference is small, but unlike when comparing with Beam simulations, all of the priority methods operate on the same assumptions and this slight difference could imply that the likelihood of a trip resulting in another trip soon after could depend on the location of the scooter.

In summary, any method of priority calculation appears to result in a higher average rate of trip satisfaction than a method that replicates the current decision-making process, within these conditions. However, considering that the difference is small, around 1%, and that the current process is heavily reliant on redistributions, which are not part of these conditions, the results are not conclusive that a priority metric results in more efficient replacements.

5.6 Optimising the Number of Replacements

As explained in Section 5.4.2, the simulations never appear to maintain a system of scooters that are sufficiently charged. It should be expected that the average battery level of all scooters stays considerably above 0% such that there is a sufficient amount of supply to meet demand. However, sampling the number of replacements from what is seen historically in the data seems to result in a number of replacements per hour that does not allow this to occur. Hence, the number of replacements becomes another unknown variable.

Another way to compare methods, that stems from the number of replacements being unknown, is to observe the increase in revenue as the number of replacements increases. There are a number of resources used when a battery is replaced, from the battery itself, the energy in the battery, and the time and energy used by the employee to travel to the scooter. Hence, it can be assumed that there is a cost to replacement.

As each replacement may increase revenue, there is a balance between the revenue gained and cost of replacement, and depending on the behaviour of both and their relationship to each other, there may be an optimal number of replacements that results in the most profit. This may also change depending on the method used, and the method with the highest optimal number of replacements would be most preferable.

The method of calculating the revenue per replacement requires an initial state that represents a good example of where scooters may be located and how charged they are at any given time. This is done by sampling scooter locations from a location dataset similar to in Section 5.3.1, and sampling scooter battery levels from a dataset made from collecting the battery information of scooters at the start and end of trips. With the data provided, this is the only way to initialise scooter batteries in a way that represents patterns that exist

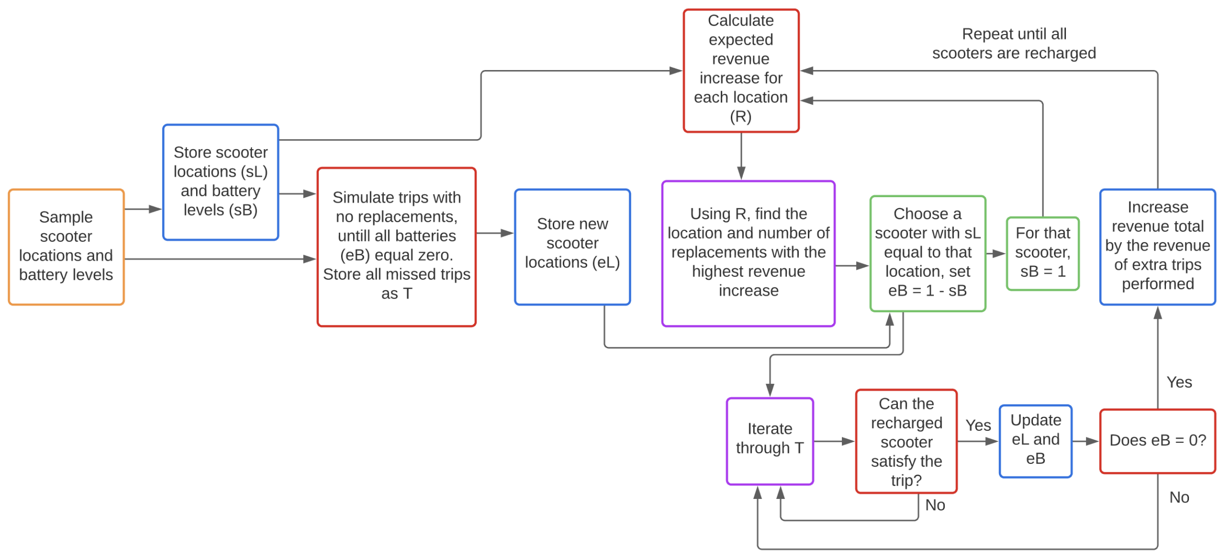


Fig. 5.14: Process used to calculate revenue gain per replacement.

in reality.

Once scooters have been initialised, a simulation is run with zero battery replacements until all scooters run out of charge. The total revenue gained and information on trips missed (T) is stored. Then, using the current method, a scooter is selected to have its battery replaced, using the conditions at the very start of the simulation, the initial location sL and initial battery level sB. Using the scooter's location when it ran out of battery (eL), and extra charge it would have from having its battery replaced before the simulation was ran (1 - sB), it can be observed if there were trips missed that the scooter can now satisfy. The list of missed trips is iterated through, and when one could have occurred in the location that the recharged scooter is currently at, this trip is compared against the scooter's new battery, eB, and if this trip can now occur the scooter is updated accordingly and the revenue is increased by the amount generated. This is done until the scooter runs out of battery again, and another scooter is chosen to have its battery replaced.

Figure 5.15 shows the results of 5 different tests using this method. It is important to note the the tests terminate at different amounts of battery replacements. This is due to the fact that the tests do not perform battery replacements if there is no estimated increase in

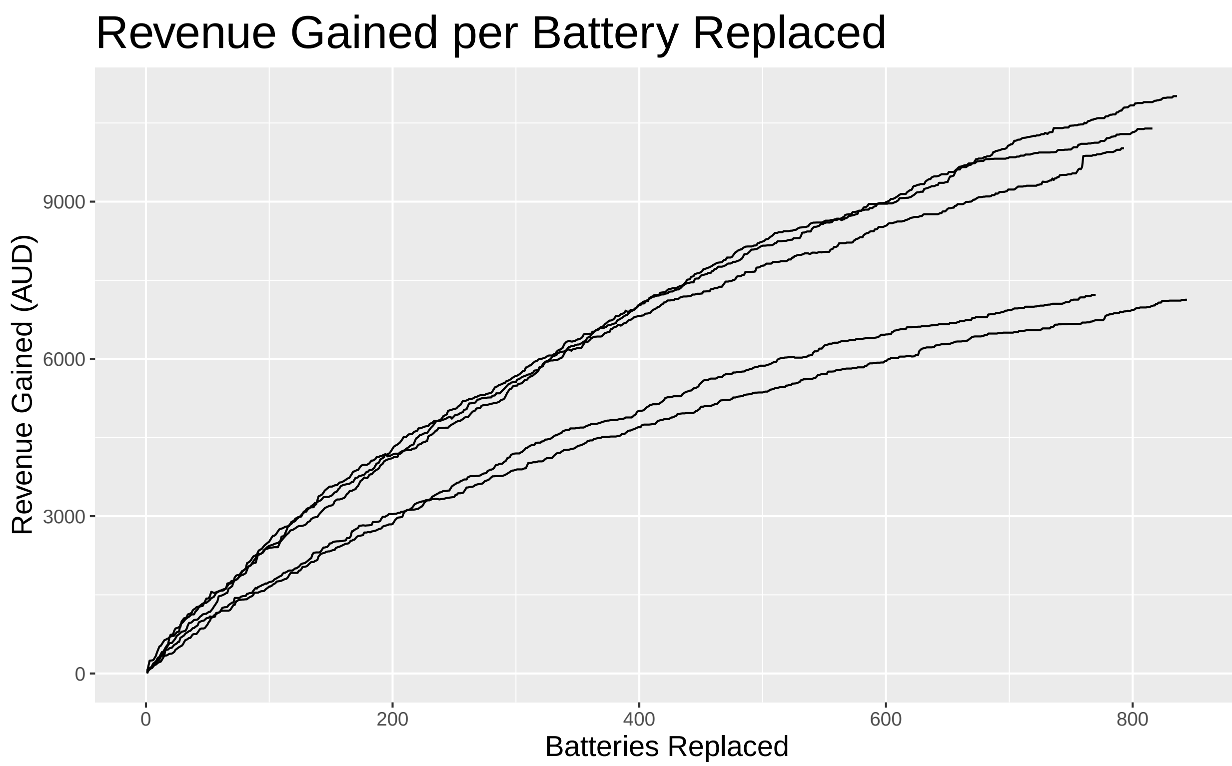


Fig. 5.15: Revenue gain per battery replacement for 5 tests. Each test is run until as many scooters as possible are recharged.

expected revenue when any more battery replacements were to occur, which depends on the initial battery levels of all scooters, and the demand of the areas the scooters end in.

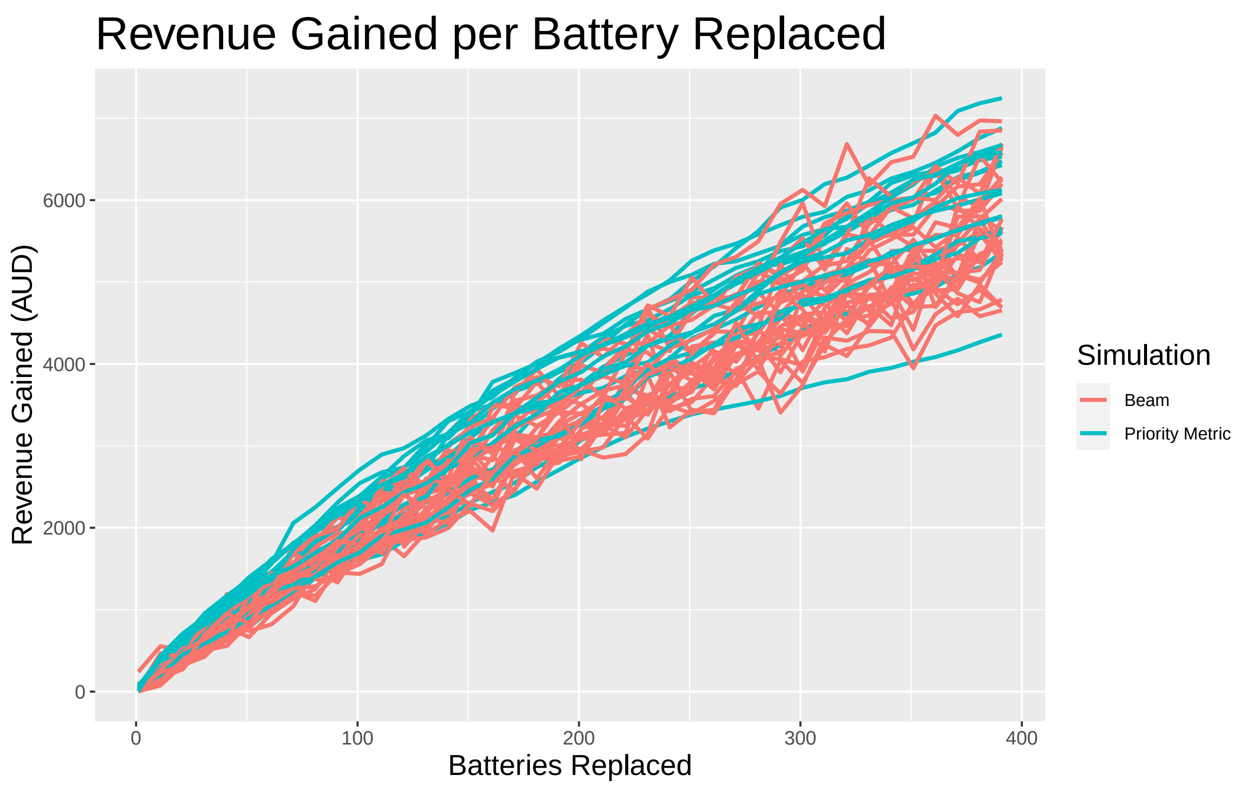
In all tests, a curve can be observed as the revenue gained per replacement tends to decrease with each replacement. Not every replacement results in an increase in revenue, which also contributes to this curve. The curve is a useful finding, as it implies that there is most likely an optimal number of replacements that results in the highest amount of profit.

There are a number of factors that contribute to the costs of battery replacement. Aside from the energy costs of recharging the batteries that are replaced, there is also the cost of travelling to each scooter and the wages of the employees involved, as more employees may need to be hired if more replacements are to occur in a short time. As such, if all the costs were known, it would be possible to calculate costs as a function of number of replacements.

Approximating the revenue per replacement using the tests developed here, and using the function of cost per replacement, the profit per replacement could be calculated. Optimal profit is assumed to exist because there are three possible cases for the relationship between the revenue and cost functions, with one being most likely:

1. The cost per replacement line lies entirely above the revenue per replacement line
2. The cost per replacement line intersects the revenue per replacement line
3. The cost per replacement line lies entirely below the revenue replacement line

In case 1, there is absolutely no profit. This is not assumed to occur in reality as otherwise e-scooters would not be a profitable business. In case 3, profit is unbounded and continues to increase as the number of replacements is increased. This is not assumed to occur, as energy costs of recharging mean that the cost per replacement is at least linear, and since the revenue curve has decreasing slope, it will intercept any linear cost function as replacements approach infinity. Hence, unless e-scooters are not profitable, case 2 must be true, and there must be some intercept between cost and revenue.



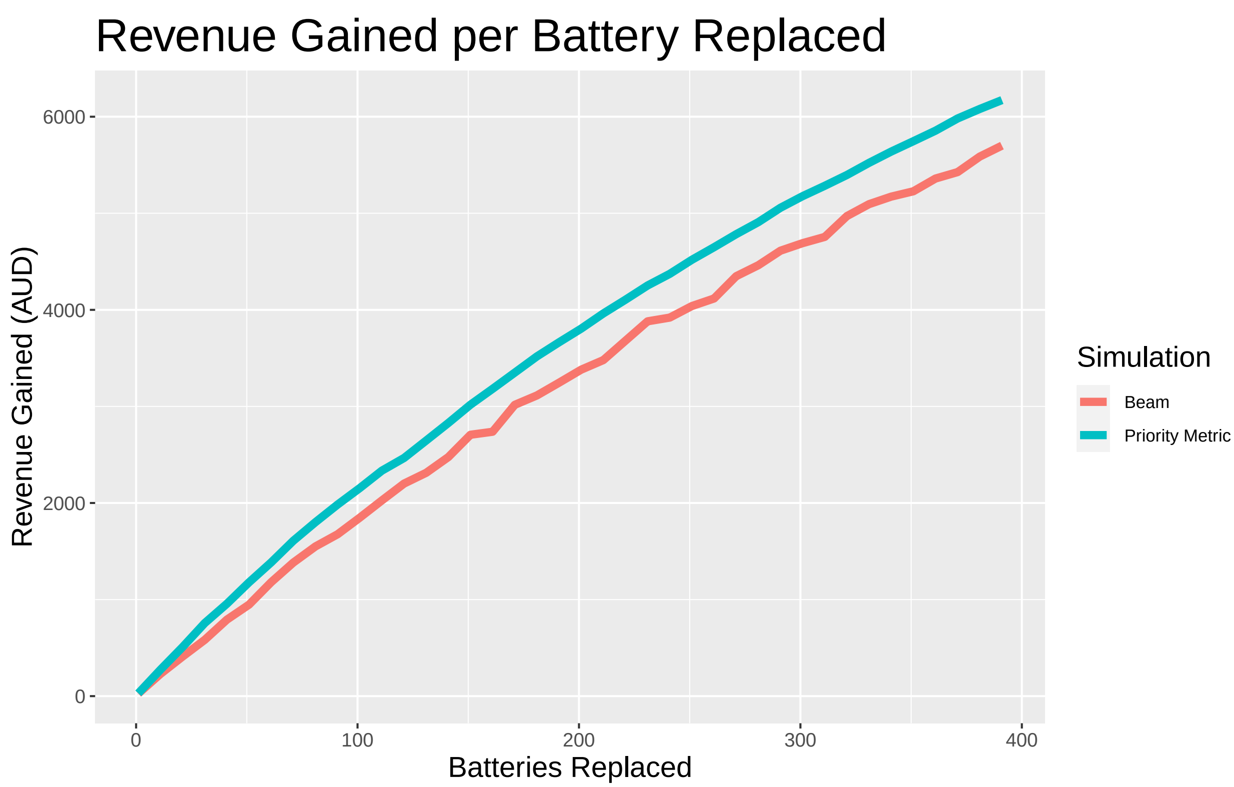


Fig. 5.16: Revenue gain per battery replacement over 30 tests each for priority metric replacement (blue) and Beam replacements (red). In order to run so many simulations, the amount of replacements was limited to 400, and the revenue gained was observed at increments of 10 replacements. Neither method consistently outperforms the other, however the revenue increase using a priority metric is more consistent and the mean revenue is higher.

Again, as the revenue function has decreasing slope, an intercept between revenue and cost must occur as revenue changes from being higher than the cost, to lower, meaning profit is bounded by some number of replacements, and there is a number of replacements between this bound, and zero, that results in the highest revenue.

As the costs are not included in the data provided, the exact number of replacements is impossible to calculate. However, the curve is sufficient to imply that an optimal number exists, and can be calculated if the costs are known.

5.7 Discussion

There is, however, the potential for redistribution to be of greater importance than replacement. For example, if a simulation is run for long enough, the entire supply may end up in low demand areas and never move. In this case, battery level will have no effect on the loss of revenue, as it will all be a result of lack of supply. In this case, future work will be needed for a reliable comparison of priority metric.

It is hence also possible to include other time-based changes in demand such as the day of the week or the month of the year. Day of week has also been shown in Chapter 2 to affect demand, as trips are much more popular on weekends, especially in areas where leisure trips are more common. However, when the dataset is filtered by both day of week and hour of day, the amount of trips to randomly chosen from becomes considerably smaller.

Lack of data is also a problem with attempting to simulate changes in demand over months. With just over a year's worth of trip data, it is unlikely that the simulation could take into account these patterns with only one occurrence of the majority of months. It is known that demand patterns will change over the year, especially in the Adelaide dataset as events like the Adelaide festival Fringe, which occur earlier in February, result in much higher demand for electric scooters. This also highlights a potential issue with the results obtained from this simulation. Scooters have been observed to increase in popularity over time when introduced to new areas. In the past two years since Beam was introduced, it has taken

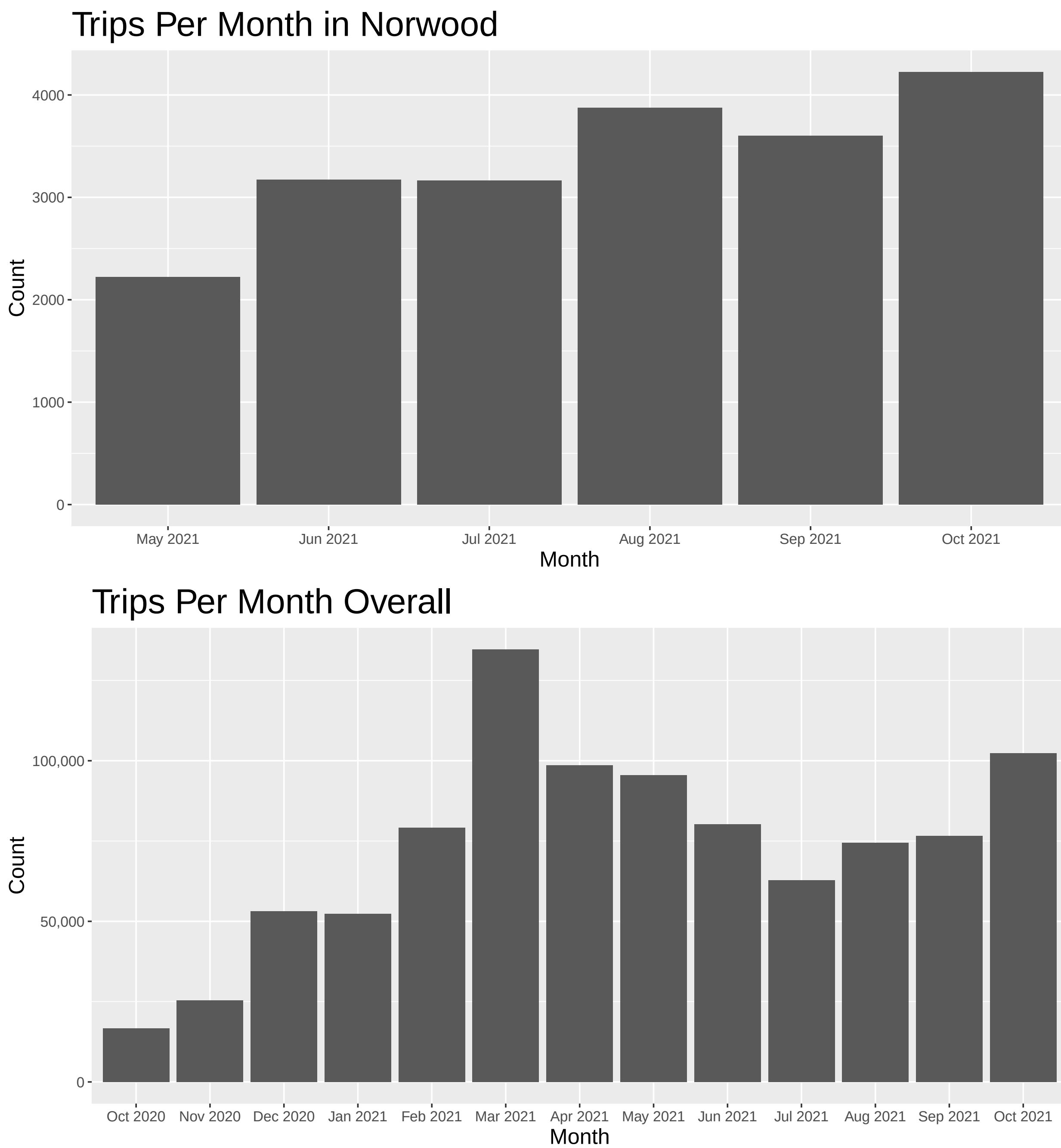


Fig. 5.17: Number of trips recorded per month in the Norwood area (top) and overall (bottom) since introduction. An increase in popularity can be observed initially. It is important to note that e-scooters were introduced to Norwood mid-May 2021, so the increase in popularity is only around 1000 trips per month [5].

time for users to acknowledge and utilise the scooters, and this can be seen with the recent introduction of scooters in the Norwood area as well (Figure 5.17). It is highly likely that the patterns of demand and trip location distribution could change over time, and results gained from simulations on this data may not be relevant for the future of e-scooter culture in the observed area, especially as Beam continues to introduce scooters to new areas.

For a short simulation however, that observes potential battery use patterns over a couple of days, month and day information is not entirely necessary for the accuracy of the comparisons. For the calculation of priority of scooters in certain locations, it only needs to be effective in a short period of time. For observations over weeks and months, much more data would be required.

In terms of the implementation of these findings. There is evidence to support that a more informed method of replacement selection will result in less trips missed due to a lack of battery charge. However, this method would need to take into account the number of employees replacing at a given time, their locations, and how quickly they can travel to high priority zones. A business that supplies e-scooters would benefit from the knowledge that there is an optimal average number of replacements before the costs outweigh the revenue obtained from replacing. Using the methods performed in this chapter, alongside their own functions of costs, it should be possible to determine this number.

6. SUMMARY

The simulations in Chapter 5 provide evidence that a method of battery replacement that prioritises low battery scooters in areas that historically experience trips with longer duration results in an increase in revenue over at least a week. At the very least, these results indicate that implementing a more informed method of battery selection will allow for more trips to be executed by users.

The 4 key topics in this research: location classification, battery usage estimation, priority calculation and simulation, are all deep areas of research and could each have more complex solutions. Hence, future work is possible in almost all parts of this thesis.

There are also some possible improvements to the aforementioned results. For instance, all simulations assumed that travelling to each scooter to perform a battery replacement was insignificant. However for the goal of reducing costs, fewer employees would be preferred and hence the path taken to replace batteries would be relevant. A more complicated optimisation algorithm could select scooters based on an optimal order of locations as well as revenue. Without this, the results obtained relate to areas of focus for replacements. In Chapter 5 it was found that Beam employees have a tendency to prioritise key areas of the Adelaide CBD, and this thesis provides evidence that these areas change depending on time and availability.

The classification methods compared in Chapter 2 were unable to outperform a simple gridding method in the tests performed. However this is potentially not always the case. For example, the suburban areas that became available for scooter trips during 2021 have different topology to a central city area. For instance, hotspots are fewer and spread apart, and individual users are less likely to start and end trips in similar locations as they may be riding to their houses. We expect that a classification method that allows for zones of varying sizes could be more effective.

The methods used to approximate trip duration given battery level in Chapter 3 and demand in Chapter 4 were relatively simple. These could be replaced with more informed approximations. The relationship between battery usage and trip duration was varied. It was assumed here that the relationship was either linear or discrete, and Figure 3.1 shows that the approximation of remaining duration can be incorrect by up to 40 minutes.

Throughout most parts of this research, a common restriction on possible work was a lack of data. There is much future work that could be done upon the provision of more data, both in terms of a larger dataset and one that includes more variables. For instance, more recordings on trips would provide more insight into the new areas opened to the east and south of the Adelaide CBD. With years worth of trips, monthly patterns could be observed and taken into account for replacement prioritisation. The expectation that areas for employees to focus replacements on change depending on certain variables could be supported by seasonal data.

It is important to note the amount of available data was unavoidable, as Beam has provided information on all trips since August of 2020, when they started to consistently provide scooters to residents of Adelaide. However, the lack of any information on battery replacements made it necessary to estimagte the location and time of any replacements from data. In Chapter 5, it was determined the number of replacements approximated was an underestimate, as all simulations were unable to satisfy the number of trips that have been seen to occur in reality. As such, we could only make findings that were regardless of the number of batteries replaced. This led to the finding at the end of Chapter 5, that an optimal number of replacements existed, but it is unknown whether this is currently reached by any e-scooter providers.

BIBLIOGRAPHY

[1] .id. *Workers' place of residence — City of Adelaide — Community profile*. 2015. URL: https://profile.id.com.au/adelaide/workers.

[2] Alessandro Ciociola et al. "E-scooter sharing: Leveraging open data for system design". In: *2020 IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE. 2020, pp. 1–8.

[3] Hao Da Dong. "Analysis and Optimization of Servicing Logistics for Self-Driving E-Scooters". 2021. URL: http://hdl.handle.net/1903/27509.

[4] Jutta Degele et al. "Identifying E-scooter sharing customer segments using clustering". In: *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE. 2018, pp. 1–8.

[5] "E-scooter trial to kick start in NPSP". In: *City of Norwood Payneham and St Peters* (May 2021). URL: https://www.npsp.sa.gov.au/article/view/1678.

[6] Eric Gonzales et al. "Modeling Taxi Demand with GPS Data from Taxis and Transit". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* {CA}-{MNTRC}-14-1141 (2014), p. 83. ISSN: 03029743. DOI: 10.1007/978-3-642-31205-2_4.

[7] Nanjing Jian et al. "Simulation optimization for a large-scale bike-sharing system". eng. In: *2016 Winter Simulation Conference (WSC)*. Vol. 0. IEEE, 2016, pp. 602–613. ISBN: 9781509044863.

[8] Hoang Thanh Lam et al. *Taxi Destination and Trip Time Prediction from Partial Trajectories*. Sept. 2015. arXiv: arXiv:1509.05257v1.

[9] Luis Martínez, José Viegas, and Elisabete Silva. “A traffic analysis zone definition: a new methodology and algorithm”. eng. In: *Transportation* 36.5 (2009), pp. 581–599. ISSN: 0049-4488.

[10] Grant McKenzie. “Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, DC”. In: *Journal of transport geography* 78 (2019), pp. 19–28.

[11] Raktim Mitra and Paul M Hess. “Who are the potential users of shared e-scooters? An examination of socio-demographic, attitudinal and environmental factors”. In: *Travel behaviour and society* 23 (2021), pp. 100–107.

[12] Neuron Mobility. *About Us*. Dec. 2021. URL: https://www.rideneuron.com/about-us/.

[13] ABC News. *Electric scooters are coming to Adelaide, if the State Government allows them*. Jan. 2019. URL: https://www.abc.net.au/news/2019-01-30/electric-%20s-to-be-trialled-during-adelaide-fringe/10763314.

[14] Gustavo Niemeyer. Feb. 2008. URL: http://geohash.org/.

[15] Robert B Noland. “Trip patterns and revenue of shared e-scooters in Louisville, Kentucky”. In: *Transport Findings* (2019), p. 7747.

[16] Jamol Pender, Shuang Tao, and Anders Wikum. “A stochastic model for electric scooter systems”. In: *Available at SSRN 3582320* (2020).

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[18] Daniel Schellong et al. "The promise and pitfalls of e-scooter sharing". In: *Boston Consulting Group* 12 (2019), p. 15.

[19] Liying Song et al. "Addressing the Last Mile Problem: Transport Impacts of Collection and Delivery Points". eng. In: *Transportation Research Record: Journal of the Transportation Research Board* 2097.1 (2009), pp. 9–18. ISSN: 0361-1981.

[20] Earth Policy Update. *Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide*. Apr. 2015. URL: http://www.earth-policy.org/plan_b_updates/2013/update112.

[21] J. W. Yoon, F. Pinelli, and F. Calabrese. "Cityride: A Predictive Bike Sharing Journey Advisor". In: *2012 IEEE 13th International Conference on Mobile Data Management*. 2012, pp. 306–311.