Characterizing Shared Mobility Operator and User Behavior Using Big Data Analytics and Machine Learning

A

Thesis

Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

in partial fulfillment
of the requirements for the degree

Master of Science

by

Tina Tang

May 2021

APPROVAL SHEET

This

Thesis

is submitted in partial fulfillment of the requirements
for the degree of

Master of Science

Author: Tina Tang

This Thesis has been read and approved by the examing committee:

Advisor: Devin K. Harris

Advisor:

Committee Member: Brian L. Smith

Committee Member: Andrew Mondschein

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2021

Abstract

Towards more sustainable use of resources in cities, there is a rising trend in shared mobility for collaborative consumption. As a condition of working with cities, third party organizations managing shared vehicle fleets often have to provide public access to real-time data describing the location of vehicles. These datasets hold enormous value for monitoring and evaluating emerging transportation services; however, a major challenge for city planners and regulators remains extracting the value from streaming transportation data by leveraging analysis and visualization methods. E-scooters are an emerging shared mobility service that have been adopted in cities across the world, but, despite their popularity, cities are still searching for more effective monitoring methods in order to understand the benefits brought to their communities or lack thereof. Using real-time e-scooter data from Charlottesville, Virginia as a case study, this work aims to characterize operator and user behavior by using big data analytics and machine learning to gather important insights. Specifically, this work provides the following contributions via three analytical studies: (1) Study I demonstrates how e-scooter data can be harvested from streaming GPS traces and then aggregated and spatially joined with demographic, employment, and built environment data. A multiple regression analysis examining the relationships between these datasets revealed that e-scooter distribution was influenced by economic activity whereas e-scooter use was influenced by micro-transit need factors and built environment characteristics. (2) Study II presents data aggregation and visualization approaches for monitoring and evaluating e-scooter operator distribution decisions, showing that utilization is a suitable measure for planning and revealing that there is room for improvement for equitable fleet distribution. (3) Study III shows the efficacy of using Latent Dirichlet Allocation to characterize user trip behavior from an unstructured set of estimated e-scooter trips. Findings suggest that trip behavior differed significantly during periods with increased student population influxes. Charlottesville planners and regulators may use the results and methods presented in this work to make data-driven decisions for improving micro-mobility as a service for the community they serve.

Keywords: *big data analytics, machine learning, data visualization, spatial data fusion, multiple regression, Latent Dirichlet Allocation, shared mobility, micro-mobility*

Acknowledgements

I extend a heartfelt thank you to each person that has supported me as, without you, this work would not have been possible. The list runs long – too long to list every person by name, in fact, but I trust you know who you are. That said, I would like to thank:

- ◊ My advisor, Devin K. Harris, for his continuous support of my work, dedicated mentorship, and curious enthusiasm about e-scooters.
- ◊ My lab members – particularly, Mohamad Alipour for his mentorship in my first year and Tianshu Li for her friendship.
- ◊ Amanda Poncy, Charlottesville's Pedestrian and Bicycle Coordinator, for collaborating with me in this work and providing invaluable input from the city's perspective.
- ◊ My committee for their support and time dedicated to reviewing this work.
- ◊ My UVA community. I am incredibly grateful for my time at the university. During my undergrad years, I was lucky to find an amazing group of lifelong friends that I cannot imagine life without. I had a wonderful undergrad advisor, Teresa Culver, who introduced me to research then and encouraged me to return to UVA to pursue my interests in machine learning. A special thank you to Brian Smith and Dottie Gardner for their kindness and support which helped me through tough times. And, of course, thank you to my ESE grad student family as their support and friendships meant the world. Thanks in particular to Courtney Rogers for always making sure my world continued to spin regardless of what challenges emerged.
- ◊ My professors whose passion for teaching machine learning and data analytics inspire me to be a lifelong learner.
- ◊ My friends near and far that have encouraged me along the way.
- ◊ My family, especially my cousin Connie, my brother and my parents. Their love and support has carried me through life.
- ◊ Last, but certainly not least, my significant other, for being my biggest supporter. Thank you for always encouraging me to pursue my dreams.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Towards Smarter Cities and Smarter Mobility

A United Nations report estimated that 68% of the world population will live in urban areas by 2050 [1]. As urban areas grow, efficient allocation and sharing of resources becomes increasingly important. The term "smart city" often refers to urban areas that strive to leverage big data analysis as a means to improve mobility, safety, governance, and living standards for city residents [2]. The development of strategic policies and data analysis techniques is fundamental for building and managing smart cities [3], [4]. However, a major challenge for city planners and local governments is understanding how to extract value from large quantities of information gathered by scattered sources [5]. As such, leveraging appropriate data mining and machine learning techniques for analysis is critical [5].

Within smart city contexts, one key area of focus for data-driven innovation and improvement is mobility [6], [7]. Towards maximizing the utilization of mobility resources in cities and more sustainable consumption, there is a rising trend in disconnecting vehicle usage from ownership and, instead, moving to shared mobility services [8]. Examples of shared mobility include carsharing, ride-hailing, station-based bikesharing, and, most recently, dockless mobility sharing [8]. Dockless mobility refers to shared micro-mobility vehicles such as electric bikes and scooters that can be parked anywhere without station restrictions [9], [10]. The popularity of e-scooters, specifically, has been expansive with Americans taking 86 million scooter trips in 2019 [11]. Popularity aside, e-scooter share is still an emerging transportation mode for which policies must be written and operations must be monitored [12], [13]. To that end, cities are searching for effective ways to monitor, manage, and optimize this new transportation mode to ensure it is actually benefitting the communities they serve [12], [14].

As third party organizations to cities, shared mobility operators are often required to provide publicly accessible data detailing where vehicles in their fleet are located in real-time [12], [15]. This data is high resolution and does not contain personally identifiable information [16], making it an ideal resource for mobility analysis. However, the raw GPS trace data provides little value without further processing, visualization, and analysis. Using data from a real-time e-scooter data feed in Charlottesville, Virginia as a case study, this work aims to characterize operator and user behavior using big data analytics and machine learning to help city planners make data-driven decisions for improving micro-mobility as a service for the communities they serve.

1.2 Thesis Outline

This work begins with an overview of relevant literature on big data analytics and machine learning methods for mobility focused analysis followed by a detailed review of micro-mobility research. Then, Study I details an approach for extracting value from big data harvested from streaming real-time GPS trace data. By fusing aggregated e-scooter data spatially with demographic, employment, and built environment data, multiple regression is used to explain the factors that drive e-scooter distribution and ridership decisions. Study II dives deeper into characterizing operator behavior by aggregating and visualizing the collected data in a meaningful manner across four distinct periods with a focus on evaluation against equitable access policy. Finally, Study III characterizes e-scooter user behavior by using Latent Dirichlet Allocation to discover hidden trip themes from an unstructured collection of estimated trips as this sheds light on how e-scooters are currently being used. City planners may use the approaches and results detailed in this work to evaluate emerging transportation modes, working towards improved data-driven city management.

2 Literature Review

2.1 Big Data Analytics

New digitalization standards in the age of the Internet of Things (IoT) have led to the widespread usage of low-cost sensors [17] and real-time data streaming from those sensors [18]. The large amounts of data available from these IoT devices are examples of big data [19]. Big data in transportation offers an opportunity to gather detailed insight into mobility behavior if appropriate data collection, storage, fusion, processing and visualization approaches are utilized [18]–[21]. Determining the appropriate data analysis techniques is often challenging due to the volume, variety, and formatting variability in transportation data [21]. Additionally, current Intelligent Transportation Systems data monitoring and analysis functionality is widely acknowledged as limited, motivating a need for better approaches [22]. More specifically, the growing abundance of sensor GPS trace data and open data feeds necessitates analysis approaches for extracting value from these read-only JSON code heaps that a human analyst cannot readily understand [23]. Seeking to contribute to this area of research, this work aims to identify and showcase the efficacy of suitable data aggregation, fusion, and visualization techniques for extracting value from streaming real-time sensor data in order to better understand and manage emerging shared mobility services. Additionally, the data used throughout this work does not contain any individual user information, which aligns with smart city data privacy objectives [24].

2.2 Machine Learning

Artificial Intelligence is made possible by processes such as Machine Learning (ML) where systems learn tasks such as classification, clustering, and pattern recognition from input data [19]. Prior to choosing a particular ML technique it is important to consider the type of data available and the goal of analysis to obtain meaningful results. In Study I, the goal was to understand the effect of demographic, employment, and built environment factors on e-scooter placement and e-scooter utilization. As such, a suitable machine learning approach is multiple linear regression as it can determine the predictive power of multiple independent variables for each of the dependent variables and has the added benefit of outputting easily interpretable results. To support this method further, another study demonstrated using multiple linear regression successfully to examine the relationship between demographic data and car-sharing availability [25]. In Study III, the goal was to discover the main routes traversed by e-scooter users in an otherwise unstructured collection of trips. For knowledge discovery such as finding latent trip themes, unsupervised machine learning methods such as clustering are often utilized to classify data into discrete groups [5], [19], [26]. Due to the structure of the aggregated e-scooter trip records, a method called Latent Dirichlet Allocation (LDA) is fitting for reasons that are explained thoroughly section 5.2. Briefly, LDA is a generative mixture model that takes large collections of discrete data usually in the form of text documents as input and produces a distribution of topics that describe the entire input dataset [27].

2.3 Micro-mobility

The proliferation of e-scooters as a micro-transit option has offered residents and visitors in cities across the world a new way to navigate a city's landscape. The National Association of City Transportation Officials reported that Americans completed 38.5 million e-scooter trips in 2018 [28] and 86 million trips in 2019 [11], highlighting its substantial growth in adoption across cities in the United States. Despite the widespread popularity, little is known about the factors that contribute to increased e-scooter usage and it is unclear whether scooters are meeting the needs of those residents who could benefit the most from them.

Recent research into the potential of e-scooter adoption has found evidence that e-scooters could have a positive economic impact on communities [29]. In one study, researchers analyzed the comparative benefits of hypothetical e-scooter trips in Chicago over other modes of transportation [29]. The study found that e-scooters are a cost-effective alternative to personal vehicles for short trips between 0.5 and 2 miles and estimated that e-scooters could provide a 16% increase in the number of jobs accessible within a 30 minute commute compared to walking or public transit [29]. Importantly, this work also noted that e-scooter benefits can differ significantly between geographic areas based on access to transit lines and bus routes [29]. Further, a study evaluating the benefit of micro-mobility compared to ridesharing services such as Uber found that e-scooters could serve as a faster means of transport during typical commute windows when traffic congestion is heavy in urban areas such as Washington D.C. [13]. Another study showed that there is a net reduction in environmental impacts when short distance car trips are regularly replaced with a greener alternative such as e-scooters, further highlighting potential benefits [30].

Despite the potential for e-scooters to address short distance transportation needs, it remains unclear if e-scooters are meeting those needs in practice. Many cities have conducted dockless mobility pilot programs to test micro-mobility as a new service and published reports of their findings [31]–[33]. Survey responses indicated that a quarter of the e-scooter trips in Baltimore, Maryland and approximately half of the e-scooter trips in Charlottesville, Virginia are commute trips [32], [33]. In contrast, studies that have analyzed actual e-scooter usage patterns have not found evidence to suggest e-scooters are being used for commutes [34]–[39]. Several of these studies observed that the time of peak e-scooter usage is outside typical morning and evening commute windows suggesting that commutes at most make up a small fraction of e-scooter trips [34]–[39]. Additionally, in at least one study, e-scooter usage was substantially higher on weekends than weekdays, supporting the claim that e-scooters are primarily used for recreation, as opposed to a first or last-mile solution [37].

Interestingly, studies have found that bikeshare usage patterns do reflect commuting behavior [39]–[43]. However, this commute activity is only significant in docked bikeshare whereas dockless bikeshare behavior indicated casual use [9], [10]. A study specifically comparing docked bikeshare to e-scooter usage found similar results, adding that bikeshare users with memberships were most likely to use the service for regular commuting [39]. This finding suggests that the reliability of finding an available vehicle is important for commuters.

Studies conducted in Austin, Texas and Indianapolis, Indiana found that e-scooters are mainly used in city centers and university campuses based on where trips started and ended [35]–[37]. However, an e-scooter trip can only take place from a specified location if an e-scooter is present; thus, if e-scooters are most available in city centers and on university campuses, then it follows that more e-scooter trips will take place at these locations. Additionally, the absence of information on how e-scooter companies deploy and reposition their fleets has been cited as a major limiting factor for understanding e-scooter usage [34]. Furthermore, one study analyzed e-scooter utilization in terms of minutes a fleet of e-scooters were reserved per day, finding that only 15% of e-scooters were used for more than an hour per day [37]. This suggests that most e-scooters are parked and unused for the majority of the day, indicating that operational strategies have room for improvement to better serve communities. Thus, it is necessary to investigate where e-scooters are available and whether they are accessible in geographic areas where residents rely more on public transportation or walking as a primary mode of travel as these are the most likely converters to e-scooter trips.

Charlottesville is a small city with an area of 10.24 square miles, a population of 47,266, and is home to the University of Virginia, a comparatively large university employing almost 30,000 people and supporting a student body of nearly 24,000 [44], [45]. E-scooters were introduced in Charlottesville soon after the start of the city's Dockless Mobility Pilot Program in November 2018 [33]. A report to City Council from June 2019 summarized the new service as overall positively received and popular with over 115,000 trips total, averaging to about 700 trips per day [33]. Approximately 50% of rides ended near the University of Virginia, suggesting that e-scooters were disproportionately popular amongst students and staff of the university [33]. Figure 1 below shows examples of VeoRide e-scooters parked around Charlottesville.

Figure 1. VeoRide e-scooters near UVA

A new law passed in the state of Virginia tasked local authorities with regulating micro-mobility [46]. This means that in cities such as Charlottesville, local government officials are responsible for evaluating and regulating e-scooter operations, an emerging service for which new regulations must be written. One finding from the Pilot Program emphasized the need for active program management in order to ensure the e-scooter fleet is distributed equitably throughout Charlottesville. As such, city operators continue to seek better ways to understand e-scooter data in order to make informed recommendations.

Throughout this thesis, data analysis approaches to extract meaningful insight about shared micro-mobility operator and user behavior from streaming raw GPS trace data are presented and discussed. These insights about system behavior can help public officials regulating e-scooters and operators distributing e-scooters evaluate the level of service offered to the community.

3 Study I: E-Scooter Availability v. Utilization Insights

3.1 Motivation

While the potential benefit e-scooters offer to communities to address short distance transportation needs have been shown, evidence that those needs are being met in practice is mixed [34]–[39]. Thus, there is a need to investigate current e-scooter share system behavior and the extent to which it addresses transportation needs in cities. This work was motivated by the need to understand if the e-scooter trends observed in previous studies are skewed by how e-scooter fleets are distributed. A focus on the utilization rate of an operational e-scooter fleet allows researchers to analyze e-scooter usage conditioned on e-scooter availability. This work sheds light on the factors that drive e-scooter operator distribution strategies and illustrates how these factors align with e-scooter utilization in Charlottesville, Virginia using a geospatial multiple regression analysis.

3.2 Methods

The location and reservation status of each e-scooter in Charlottesville were collected from a real time data feed and stored in a database at high frequency intervals over a period of four months. Additional data about geospatial factors across Charlottesville, such as economic activity indicators, resident commute needs, and built environment descriptors were collected from other sources [47]–[50]. This data was fused with the e-scooter data by joining each data element to its bounding Census block group, as illustrated in Figure 2. The e-scooter data was further processed so that each Census block group had an aggregate measure of e-scooter utilization and availability.

Next, the relationships between e-scooter availability and the geospatial factors were investigated using multiple linear regression. The relationships between e-scooter utilization and the geospatial factors were examined similarly. Finally, results from the linear models were compared to understand how different factors drive e-scooter availability versus utilization. These methods are described in detail in the following paragraphs.

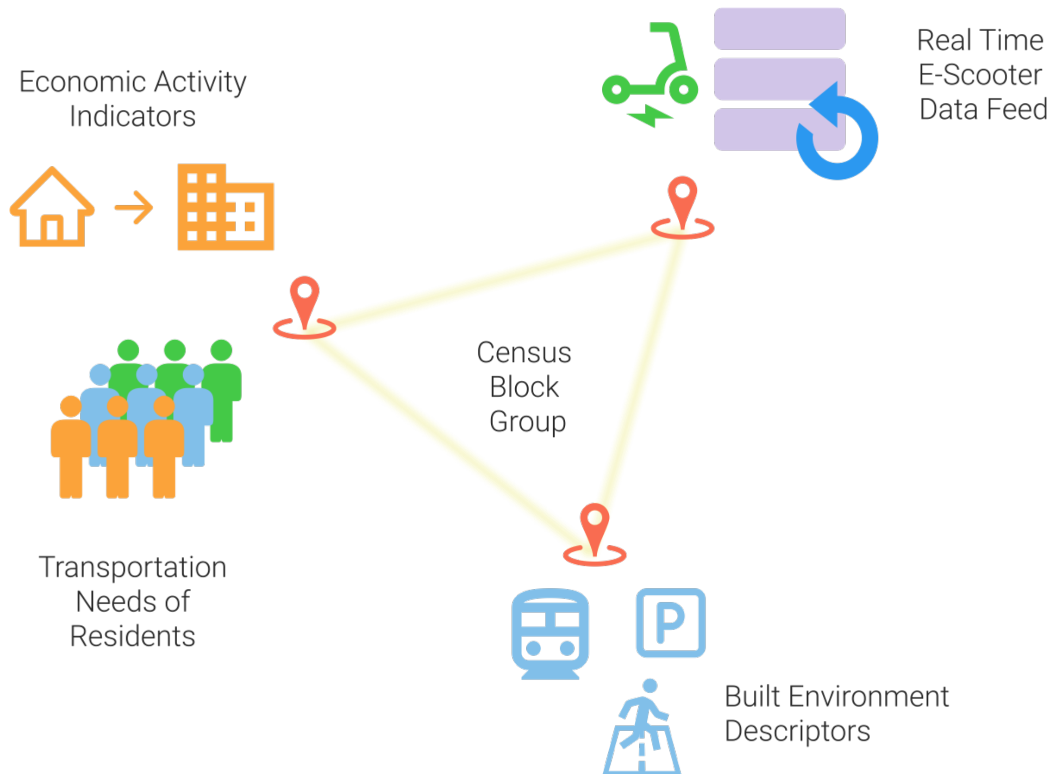


Figure 2. Data fusion by geolocation joined at Census block group resolution

Web scraping and data harvesting have become popular methods for building data sets and evaluating trends [12], [25], [51]. One example of this approach for geospatial analysis of shared micro-mobility in urban environments takes advantage of the General Bikeshare Feed Specification (GBFS), which is a standardized real-time, open data feed that is widely used for micro-mobility operations [12]. Many cities require operators to publish real-time fleet information in this format [12], [16]. This is a requirement for dockless mobility operators in the city of Charlottesville [15]. This data includes the real-time GPS location of every vehicle in the operator's fleet as well as indicators for whether each vehicle is reserved or disabled without sharing any user information. Thus, this data is an ideal resource for researchers seeking to analyze scooter operator and user behavior as it is easily and widely accessible and does not pose any direct user privacy concerns.

For this study, VeoRide's real-time GBFS data feed was continuously queried at a two-minute frequency, which equates to taking timed snapshots of each e-scooter's location and reservation status. The orange outline in Figure 3 below shows an example of the raw data made available by any GBFS feed. Per query, all the pulled records were stored along with an added timestamp outlined in blue in Figure 3. Over a four-month period from March 15, 2020 to July 15, 2020, the collected data was combined to build a rich set of ordered e-scooter observations.

bike_id	lat	lon	vehicle_type	is_reserved	is_disabled	last_updt_dtm
50106066	38.0346	-78.4987	Scooter	1	0	2021-01-17 22:46:04
50106066	38.0407	-78.4959	Scooter	1	0	2021-01-17 22:48:04
50106211	38.0410	-78.5009	Scooter	1	0	2021-01-18 02:10:04
50106211	38.0450	-78.4947	Scooter	1	0	2021-01-18 02:12:04
50106211	38.0499	-78.4965	Scooter	1	0	2021-01-18 02:14:04

Figure 3. E-scooter Raw Data

Although e-scooter trip characteristics were not the focus of this study, trips were pulled from the raw data to provide a high-level summary of e-scooter usage in Charlottesville during the study period. Additionally, it was important to visualize where trips started for comparison with e-scooter utilization. The trips were extracted by identifying consecutive records where the "is_reserved" indicator is set to 1, denoting an active reservation. An example is highlighted in green in Figure 3 above. This is similar to the methods implemented by Zou et al. for extracting e-scooter trips from a Washington D.C. operator's GBFS data feed [12]. Raw GPS data is often subject to tracking errors due to the sporadic unavailability of satellites [12]. Additionally, there was noise in the dataset from canceled trips or trips that were not ended properly. Thus, similarly to data cleaning implemented by Zou et al., trips shorter than 0.02 miles or greater than 90 minutes were excluded from analysis [12].

Over the four-month data collection period from March 15, 2020 through July 15, 2020, 10,170 trips were extracted from the raw e-scooter feed data for an average daily trip count of 87. The daily trip count over the period of study is illustrated in Figure 4. The large drop in ridership in mid-March is attributed to students leaving the University of Virginia as mandated by university leadership during the early stages of the COVID-19 pandemic. It should be noted that compared to the Charlottesville pilot study previously mentioned, the average number of trips observed was significantly smaller as a result of the particular period of study. The implications and impacts of this issue are discussed in a later section. Furthermore, there was a minor disruption in data collection due to a technical issue from May 21 through May 26, but the data is otherwise continuous. This gap explains the flattening line at the end of May marked by the dotted vertical lines in Figure 4.

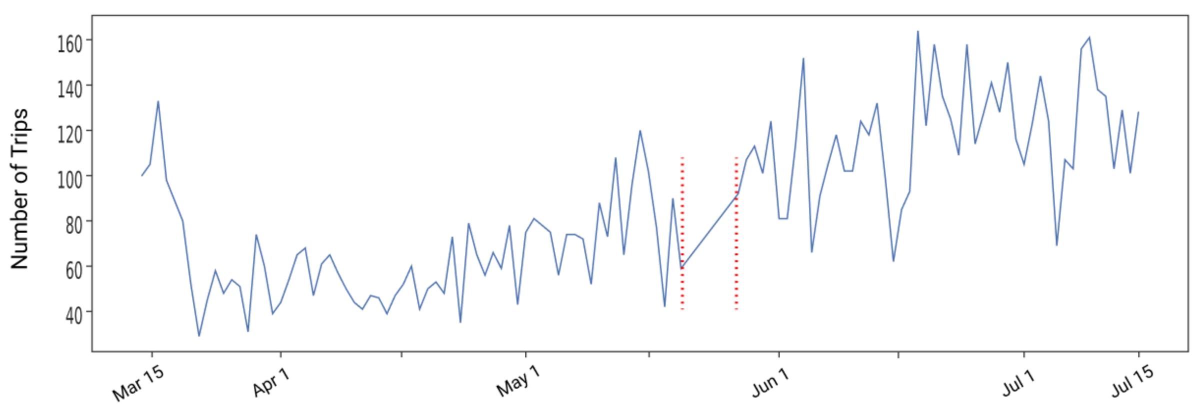


Figure 4. Daily e-scooter trip counts from March 15th through July 15th 2020

Figure 5 illustrates how trips were distributed across days of the week and hours of the day. The trip distribution heat map closely resembles the plot presented by Jiao et al. in a study focused on e-scooter travel behavior in Austin, Texas [36] and fits the description provided in other studies [37]–[39]. Similar to previously mentioned studies, the trip distribution shown in Figure 5 does not provide strong evidence that e-scooters are used for commute trips in Charlottesville. Importantly, as the largest employers in Charlottesville are the University of Virginia and the UVA hospital, commute times may not fit typical morning and evening windows [52]. The average trip displacement during this period was 0.59 miles and the average riding speed estimated from the trip observations was approximately 6.9 miles per hour.

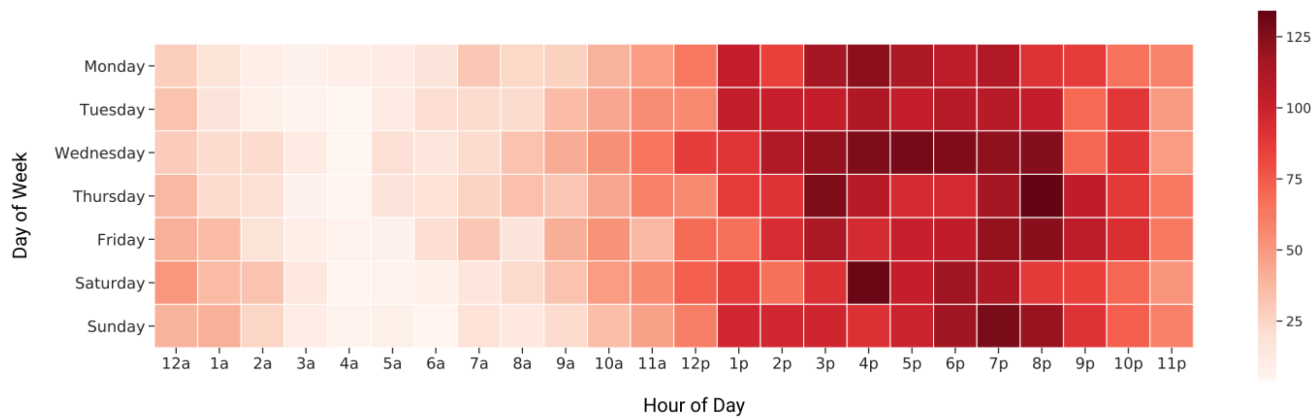


Figure 5. E-scooter trip count temporal distribution

An important piece of information from the GBFS feed data that has not been extensively studied to the author's knowledge is the unreserved e-scooters for each slice of e-scooter observations where the "is_reserved" indicator is 0. This data describes the locations of e-scooters that were not in use when the live data feed was queried. The location of available e-scooters is a critical factor to understanding how easily an e-scooter can be accessed and is a causal factor for whether a trip event can occur. This information provides insight into the e-scooter operator's decision-making. To quantify e-scooter availability, the mean number of vehicles available at each hourly time slice within a quarter mile per block group was calculated. This metric was used because it is considered within walking distance and thus is more interpretable [50], [53]. Essentially, this approach answers the question, “How many scooters could be found within walking distance of any point in each block group?”. It should be noted that e-scooter availability did not vary significantly from slice to slice; thus, there was little loss of information for a significantly improved computation resource trade off by including one slice of observations per hour. Next, the fleet utilization percentage was calculated by dividing the number of vehicles in use per block group by the total number of vehicles in operation per time slice. The mean of this value across all time slices describes the average fleet utilization rate per block group, enabling the analysis of the factors driving e-scooter placement versus use.

Next, the focus shifted to identifying areas where residents are most likely to benefit from commuting using an e-scooter. One study demonstrated the effectiveness of visualizing transit need using geospatial visualization techniques to highlight areas with transit demand followed by transit supply using demographic data from the US Census and transit data to visualize supply [54]. This study adopted a similar approach where the demand distribution for short-distance transportation alternatives was visualized based on Census statistics describing transportation use, highlighting areas where micro-mobility need is the most salient.

The need for a short-distance travel alternative in Charlottesville was characterized by focusing on factors driving alternative transportation needs, including individuals without personal vehicles and individuals that rely on public transportation or walking. High resolution estimates of this information is publicly available via the U.S. Census Bureau [55]. The U.S. Census Bureau conducts the American Community Survey (ACS) annually which provides social, economic, demographic, and housing characteristics estimates published at the block group level [48]. A block group is a nationally defined geographical unit smaller than a census tract, but larger than a census block, controlling for population counts between 600 and 3000 [56]. Block group location and area descriptions are made available via the U.S. Census Bureau [56]. The aggregate of survey data from new survey samples over a 5-year period is denoted by ACS-5, which provides a more reliable set of estimates [48]. ACS-5 estimates from 2013 to 2017 were used for this study as it was the most recent data set available via the Census Python API at the time of this work. Figure 6 illustrates the spatial distribution of e-scooter need factors in the city of Charlottesville including (a) the percentage of households without access to a vehicle, (b) the percentage of residents using public transit as a means to travel to work, and (c) the percentage of residents walking to work. From visual inspection, an observer can identify block groups where micro-mobility need is the most salient. It was noted that indicators of micro-mobility need was not observed as concentrated in the downtown area of Charlottesville outlined with a bold, orange line in Figure 5.

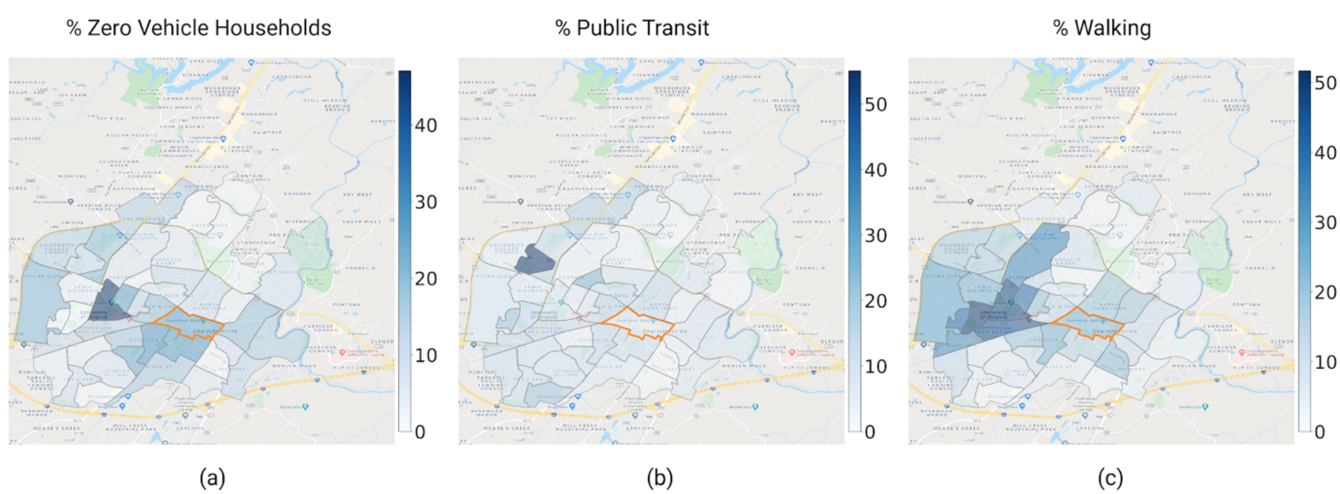


Figure 6. Geospatial distribution of micro-mobility transportation demand in Charlottesville

Another important data element in this study was the mobility of individuals from one block group to another as this sheds light on whether the benefits of emerging transportation modes are resident-focused or visitor-focused. The U.S. Census Bureau maintains a Longitudinal Employer-Household Dynamics program which partners with state labor market offices to produce rich sets of data illustrating employee-to-employer travel [47], [57]. These high-resolution data sets are free and publicly available [47], [57]. In this study, the origin-destination data within Albemarle County and Charlottesville city blocks were used to calculate the inflow and outflow of commute activity, creating an indicator for economic activity. The most recent origin-destination pair data sets from 2017 were used for this study; however, it should be noted that a small subset of the origin-destination pairs are from 2015 as this was the latest data available for federal jobs. The LEHD data were grouped at the block group level in order to directly join it with the aforementioned demographic variables from ACS-5. The influx of commuters into block groups was calculated by taking the sum of the number of jobs for which residents living in one block group travel into another block group for work. Similarly, the outflux of commuters was calculated by taking the sum of the number of individuals traveling to jobs that differ from their residence block group. Larger commute influx values in this study represent a measure for economic activity as more services are offered in busier areas. The spatial distribution of these metrics are visualized in Figure 7 where (a) shows the residential population per block group, P_r, before commuting activity, (b) shows number of commuters traveling into a block group they do not live in for work denoted by J_{ci}, (c) shows number of commuters leaving their residential block groups to travel into another block for work denoted by J_{co}, and finally (d) shows the redistributed population per block group, P_c, after accounting for commuting activity described in the following equation.

$$P_c = P_r + \sum J_{ci} - \sum J_{co}$$

These factors offer insight into whether economic activity is more important than residential demographics when operators make e-scooter placement decisions. From visual inspection of Figure 7, it is clear that jobs are very centralized in downtown Charlottesville outlined with a bold, orange line. This pattern of spatially condensed employment areas is commonly observed in cities in general [58], [59].

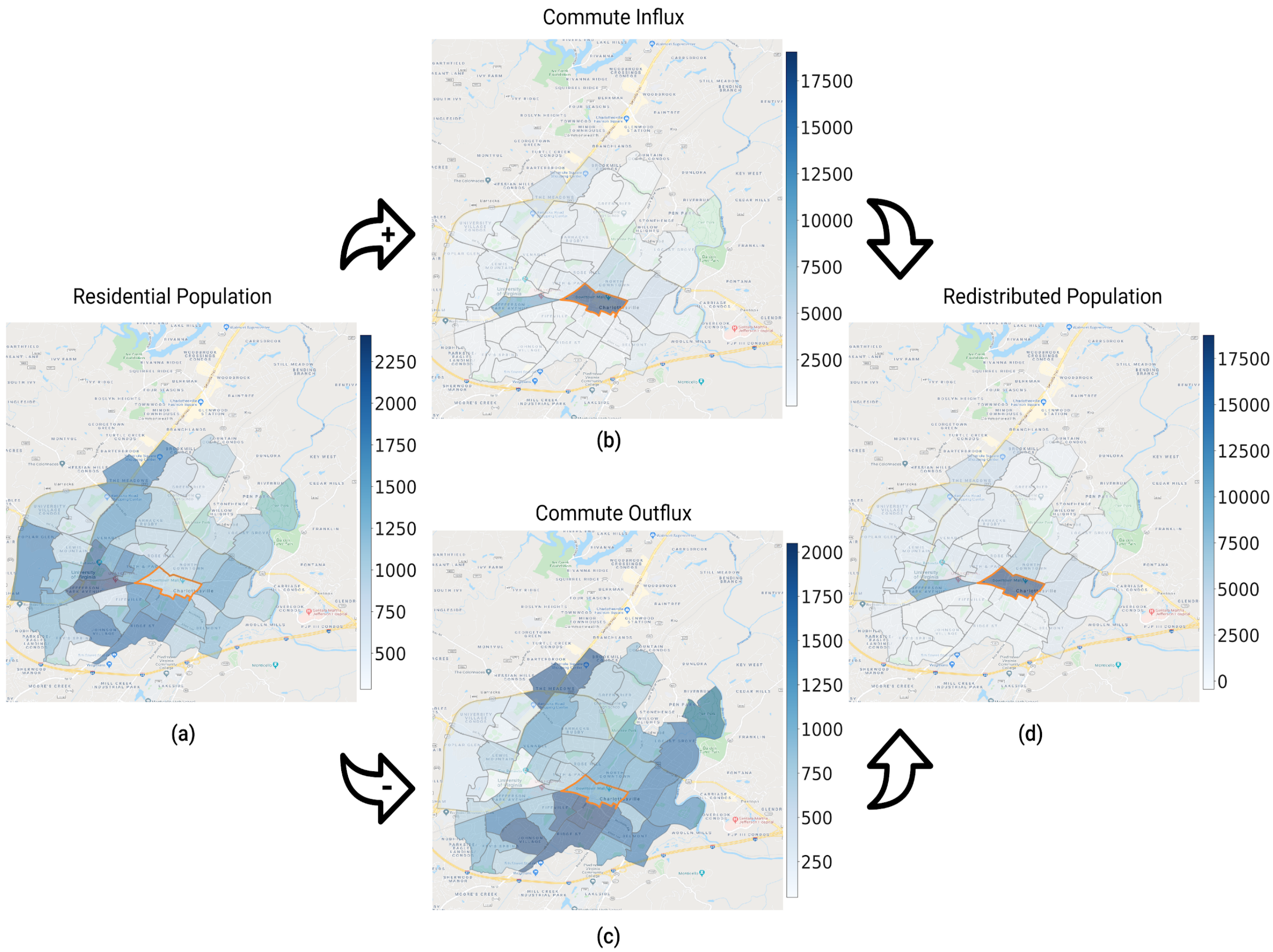


Figure 7. Effect of commute influx and outflux on population distribution in Charlottesville

Next, variables that characterize the built environment were collected by using Census block group centroid GPS coordinates and reverse geocoding them to obtain addresses as input for a Walk Score tool that extracts a walk score, transit score, and bike score per input. These scores could then be spatially joined with the previously discussed data. The Walk Score tool is free and publicly accessible [50]. Its methods for estimating walkability, bike-ability, and transit friendliness have been confirmed in research studies [50], [60], [61]. These measures are informative because they enable the evaluation of the importance of favorable walking conditions, transit conditions, and biking conditions. For example, walk scores are derived based on the number of amenities reachable via a 5 minute walk which equates to about a quarter mile distance [50]. Walk Score also provides categorical statuses based on the scores such as "Very Walkable" versus "Car Dependent". By their methods, a score from 90 to 100 indicates a status of "Walker's Paradise", which suggests that daily errands can be completed on foot. Similarly, "Biker's Paradise" suggests that daily errands can be completed on a bike and "Rider's Paradise" indicates world class public transportation such as the transit options offered in New York City or San Francisco. Figure 8 illustrates the built environment descriptor scores for each block group in this study. From visual inspection, Charlottesville's most walkable areas can be easily located as those shown in green and yellow. This visualization also suggests that transit options in Charlottesville are lacking, indicating that residents may be seeking an alternative if they currently rely on transit.

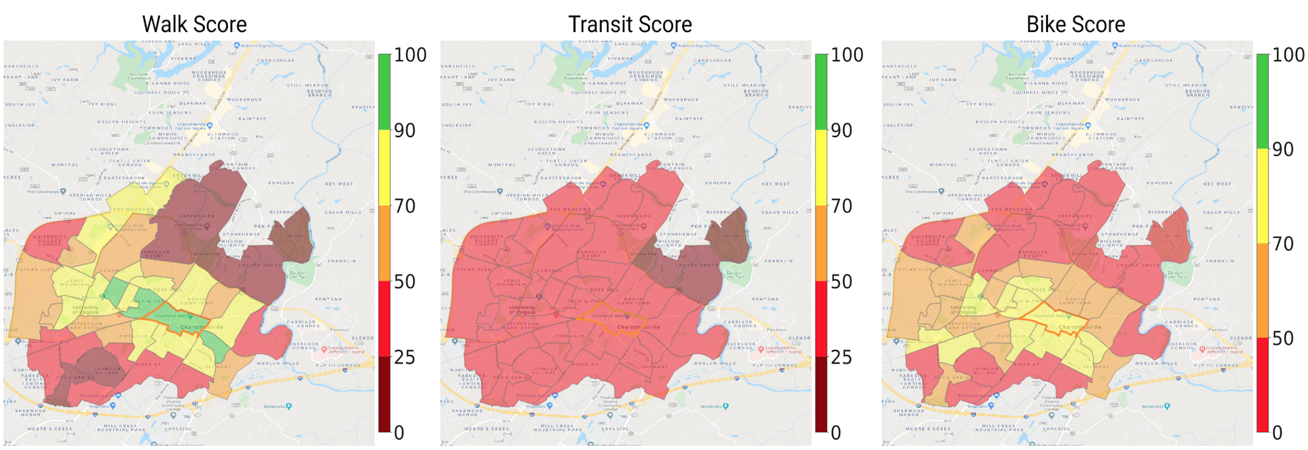


Figure 8. The walk scores, transit scores, and bike scores of each block group

Lastly, an additional measure used was the surface area of parking lots in a block group from Charlottesville's Open Data Portal [49]. This measure is important because evidence of a strong relationship between e-scooter placement or utilization could indicate a focus on visitors traveling into a block group and parking rather than residents. Table 1 summarizes all the variables investigated in this study where "DV" denotes the dependent variables of interest and "IV" denotes the independent variables. The primary objective of the study is then to understand which of the independent variables have predictive power on e-scooter availability and utilization, respectively. To do so, this study utilizes the Statsmodels package via Python to fit an Ordinary Least Squares (OLS) model to evaluate each of the dependent variables: e-scooter availability and e-scooter utilization [62].

It should be noted that a Census Tract that includes the University of Virginia is categorized as outside of Charlottesville city according to Census boundary lines. The block groups in tract 109 were concatenated with the block groups within Charlottesville city. As data from Charlottesville's dockless mobility pilot program indicated e-scooter behavior was heavily influenced by students [33], this study also evaluates how influential factors change when those block groups are included and excluded.

Table 1. OLS model variable descriptive statistics

Type	Description	Unit	Mean	Std Dev	Min	Max
DV	E-Scooter Availability	avg # e-scooters available	4.37	4.88	0.24	28.09
DV	E-Scooter Utilization	avg % of fleet reserved	0.01	0.01	0.00	0.06
IV	Population Density	# residents/mi^2	6044.82	4301.15	1158.14	21522.22
IV	Block Group Area	square miles	0.28	0.16	0.08	0.85
IV	Median Age	years	33.28	10.15	19.10	62.10
IV	% 0-Vehicle Households	% households	0.11	0.10	0.00	0.47
IV	% Public Transit	% residents	0.09	0.09	0.00	0.55
IV	% Walking	% residents	0.15	0.15	0.00	0.52
IV	Commute Influx	# people commuting into a block group	1639.26	3338.89	20.00	19048.00
IV	Commute Outflux	# of people commuting out of a block group	1062.19	534.46	48.00	2050.00
IV	Walk Score of Block Group Centroid	0 to 100	37.07	6.77	22.00	48.00
IV	Transit Score of Block Group Centroid	0 to 100	55.88	26.61	11.00	99.00
IV	Bike Score of Block Group Centroid	0 to 100	55.84	18.38	15.00	84.00
IV	Parking Lot Area	square miles	0.03	0.04	0.00	0.23

Using the data collected and aggregated into the variables defined in Table 1, a multiple regression analysis was completed and the modeling results are provided in the next section.

3.3 Results

Tables 2 and 3 present the results of the multiple regression models fit to analyze the relationship between the variables in Table 1. The following two tables compare how results differ when the blocks containing the University of Virginia are excluded and included in the analysis.

Table 2. OLS model results excluding UVA blocks

Model 1				**Model 2**			
Response Variable: E-scooter availability (excluding UVA blocks)				**Response Variable:** E-scooter utilization rate (excluding UVA blocks)			
Adjusted R^2: 0.534				**Adjusted R^2:** 0.609			
Explanatory Variable	**Coefficient**	**Std Error**	***p***	**Explanatory Variable**	**Coefficient**	**Std Error**	***p***
Population Density	0.00	0.00	.15	Population Density	0.00	0.00	.01**
Block Group Area	1.47	10.06	.89	Block Group Area	0.01	0.02	.53
Median Age	-0.03	0.11	.79	Median Age	0.00	0.00	.67
% 0-Vehicle Households	7.51	9.29	.43	% 0-Vehicle Households	-0.02	0.02	.29
% Walking	-9.15	8.38	.29	% Walking	-0.02	0.02	.20
% Public Transit	-0.67	14.18	0.96	% Public Transit	0.11	0.03	.001***
Commute Influx	0.00	0.00	.0000 ***	Commute Influx	0.00	0.00	.79
Commute Outflux	0.00	0.00	.68	Commute Outflux	0.00	0.00	.75
Walk Score of Block Group Centroid	0.05	0.05	.36	Walk Score of Block Group Centroid	0.00	0.00	.003**
Transit Score of Block Group Centroid	-0.10	0.19	.58	Transit Score of Block Group Centroid	0.00	0.00	.33
Bike Score of Block Group Centroid	0.05	0.07	.54	Bike Score of Block Group Centroid	0.00	0.00	.10
Parking Lot Area	-25.09	26.50	.35	Parking Lot Area	-0.12	0.05	.025*

The results of the first model in Table 2 reveal that commute influx is the only variable with significant explanatory power for e-scooter availability. Further, commute influx is significant at the .001 level, indicating that it is highly unlikely that this result was due to random noise. Although the coefficient is small, it indicates that for every increase of 1000 commuters into a block group, the number of e-scooters available per quarter mile increases by 1. In this study, commute influx is used as a metric for the level of economic activity evident in a block group where higher values of commute influx indicate a busier area during business hours.

The results of the second model in Table 2 reveal a completely different set of variables with significant explanatory power. The percentage of residents in a block group using public transportation as a means to travel to work proves to be most statistically significant at the .001 level. This makes sense given the earlier observation that transportation in Charlottesville did not receive high transit scores, indicating poorer transit conditions exist overall. The model suggests that a ten percent increase in transit commuters increases the e-scooter utilization rate by 1.105 percent. This is not a small effect given that the utilization rate range is very small, between 0 and 6 percent. By order of statistical significance, the next variables of importance significant at the .01 level are walk score and population density. Population density is descriptive of the residential population in a block group. Walk score indicates how favorable walking conditions are. Lastly, parking lot area was significant at the .05 level, but with a negative correlation. This suggests that a 1 square mile of parking decreases the utilization rate of e-scooters by a factor of 0.12.

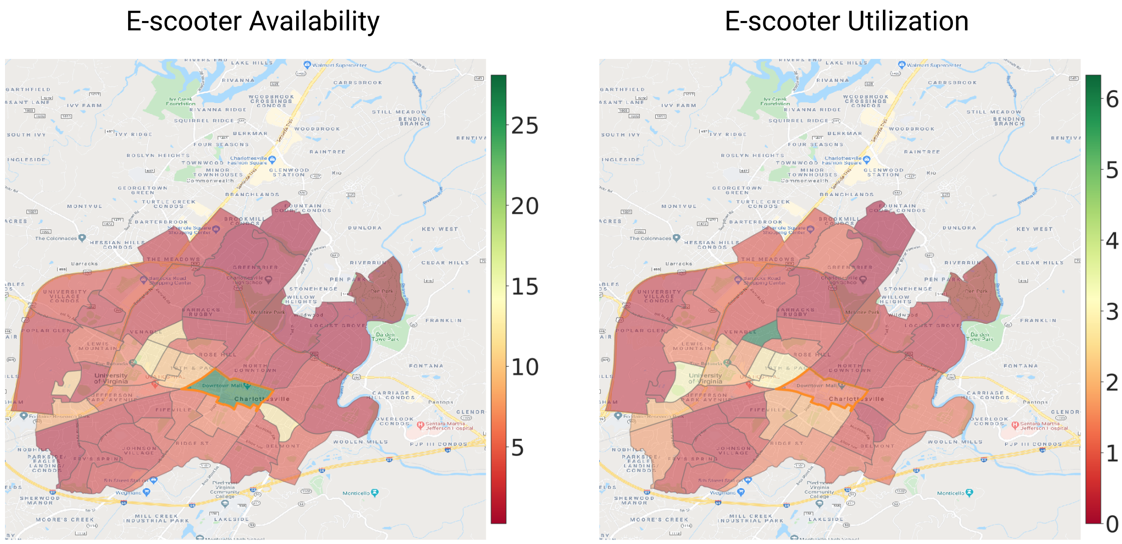
Table 3. OLS model results including UVA blocks

Model 3				**Model 4**			
Response Variable: E-scooter availability (including UVA blocks)				**Response Variable:** E-scooter utilization rate (including UVA blocks)			
Adjusted R^2: 0.563				**Adjusted R^2:** 0.344			
Explanatory Variable	**Coefficient**	**Std Error**	***p***	**Explanatory Variable**	**Coefficient**	**Std Error**	***p***
Population Density	0.00	0.00	.11	Population Density	0.00	0.00	.015*
Block Group Area	-2.09	6.23	.74	Block Group Area	-0.01	0.02	.66
Median Age	-0.04	0.10	.70	Median Age	0.00	0.00	.75
% 0-Vehicle Households	5.72	6.52	.39	% 0-Vehicle Households	-0.01	0.02	.73
% Public Transit	-11.15	7.32	.14	% Public Transit	-0.02	0.02	.35
% Walking	-2.27	7.78	.77	% Walking	0.02	0.02	.30
Commute Influx	0.00	0.00	.000***	Commute Influx	0.00	0.00	.97
Commute Outflux	0.00	0.00	.78	Commute Outflux	0.00	0.00	.53
Walk Score of Block Group Centroid	0.06	0.05	.25	Walk Score of Block Group Centroid	0.00	0.00	.029*
Transit Score of Block Group Centroid	-0.08	0.17	.62	Transit Score of Block Group Centroid	0.00	0.00	.21
Bike Score of Block Group Centroid	0.02	0.06	.70	Bike Score of Block Group Centroid	0.00	0.00	.28
Parking Lot Area	-21.71	20.09	.29	Parking Lot Area	-0.04	0.05	.43

The third and fourth models in Table 3 are fit to evaluate if the inclusion of UVA block groups changes results, indicating the impact of the student population. As previously noted, the results presented are reflective of the study period only, in which only a small percentage of the typical UVA student population is present due to the COVID-19 pandemic. The decrease in the Adjusted R2 of model 4 compared to model 2 indicates a weaker fit when UVA blocks are included. It is observed that only population density and walk score remain significant factors for e-scooter utilization when UVA blocks are factored in. The percentage of residents using transit to commute no longer becomes significant while it was previously significant at the .001 level when UVA blocks were excluded from the model. This suggests that e-scooter utilization in university areas are less driven by a transit need and that the inclusion of UVA blocks significantly changes results overall. The results are further discussed in the following section.

3.4 Discussion

Further examination of the results in this study reveal insights as to the underlying reasons behind the reported observations. Similar to previous studies, the e-scooter trip distribution in Charlottesville did not show evidence of regular commuting behavior. However, upon further investigation, geospatial visualizations reveal that available e-scooter vehicles are concentrated in areas that are heavily commuted into. This suggests that e-scooter fleet distribution at the time of this study was more likely to benefit visitors to a block than residents. As commuting requires a reliable, consistent way of travel, e-scooter availability is critical to e-scooter commute adoption. For example, as previously noted, studies evaluating docked vs dockless mobility found that docked mobility is used for commuting but dockless mobility is not [9], [10], [39]. Insights from this study indicate that increasing e-scooter availability in residential zones where micro-mobility need is greater could offer communities a more reliable commute alternative. Figure 9a illustrates how e-scooter fleets are distributed on average throughout Charlottesville while Figure 9b shows the proportion of those vehicles being used on average.



(a) (b)

Figure 9. Spatial distribution of e-scooter availability and utilization throughout Charlottesville

Then, Figure 10 illustrates the total number of e-scooter trips that started from each block from March 15, 2020 through July 15, 2020. From visual inspection of the figure, it can be seen that it resembles the distribution of e-scooter availability in Figure 9a more closely than the visualization of e-scooter utilization in Figure 9b. This suggests that e-scooter placement may be influenced by a systematically self-reinforcing relationship in which e-scooters are placed where the majority of trips start from, but the majority of trips inevitably start from where the vehicles are placed.

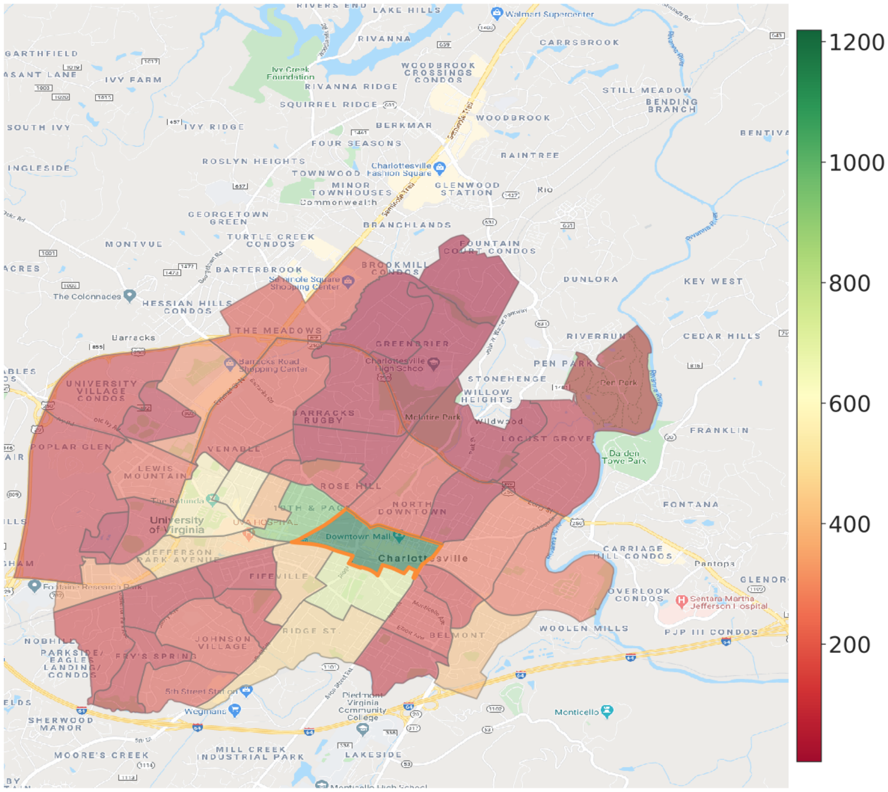


Figure 10. Total trip count per block group from March 15, 2020 through July 15, 2020

In previous studies, researchers found that office and institutional land use rather than residential area was positively correlated with higher ridership [35]. The results of this study suggest that this effect can be explained by how an e-scooter trip is dependent on an e-scooter being available, which is shown to be lower in more residential areas in this study. However, when e-scooter utilization is considered, findings in this study indicate that population density is important. Further, it is noted that built environment factors such as favorable walking conditions are important factors for e-scooter utilization.

For city regulators and e-scooter operators, results suggest that increasing e-scooter availability in residential areas where residents use transit as a major means of transportation may increase utilization when considering the Charlottesville areas excluding UVA. Generally, e-scooter operations may benefit from specifically considering fleet utilization rate data and resident transportation need factors when making fleet distribution decisions. From visual inspection of Figure 9, it can be noted that although e-scooters are the most available in Charlottesville's downtown area, the utilization rate is very low, suggesting a potential need to rebalance the fleet.

This work was accepted into the Transportation Research Board 100th Annual Conference proceedings and contributes to the micro-mobility literature by analyzing e-scooter availability versus utilization. Findings suggest that:

1. The factors driving e-scooter fleet distribution differ from the factors driving e-scooter utilization. E-scooter availability is significantly influenced by commute influx, an indicator of economic activity. E-scooter utilization, in contrast, is significantly influenced by residential characteristics such as the percentage of residents using public transportation to travel to work. It must be noted that these findings are specific to Charlottesville during the COVID-19 pandemic.
2. Findings suggest that e-scooter fleet distribution decisions may be informed by where trips start from, which may be a self-reinforcing relationship based on where e-scooters are placed
3. E-scooter utilization may be a more suitable metric for informing e-scooter redistribution decisions

4 Study II: Characterizing Operator Behavior with Data Visualization

4.1 Motivation

In Study I, trip summary data was used as a high level characterization of e-scooter ridership during the initial 4 month study period. After continuing to harvest e-scooter data from the real-time GBFS feed for one year from March 15, 2020 through March 21, 2021, the trips per day were extracted and visualized once more in Figure 11 to observe changes in the shared mobility system throughout this period.

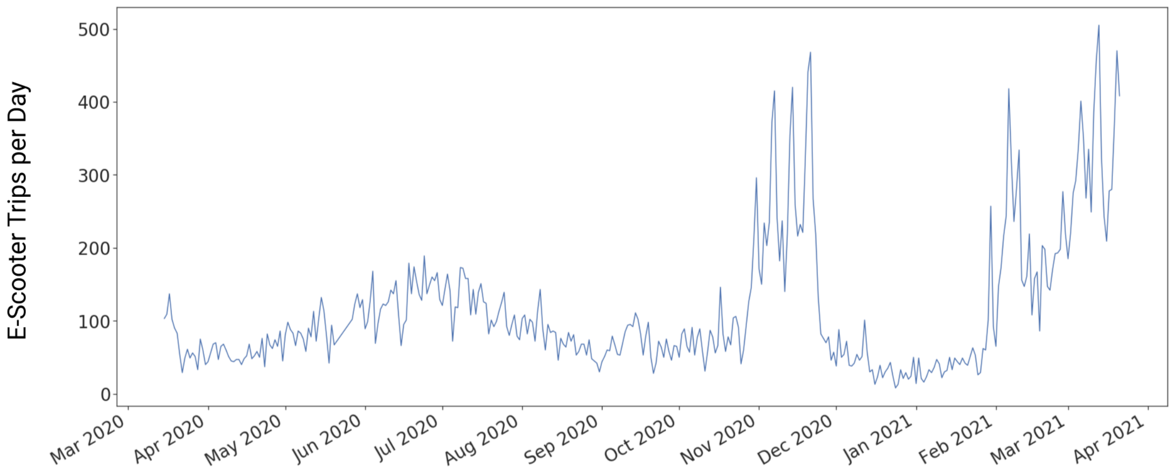


Figure 11. Daily e-scooter trip counts from March 15 2020 through March 21 2021

From visual inspection of Figure 11, it is evident that ridership increased sharply at the end of October. As this date is not marked by a significant student population influx or outflux due to academic calendar events, Study II aims to find evidence indicating changes in operator behavior to explain the spikes in ridership. Additionally, Study II takes the temporal aspect of the collected data into account by examining 4 periods of interest where the temporal groupings are determined by the peak and non-peak ridership periods shown in Figure 12.

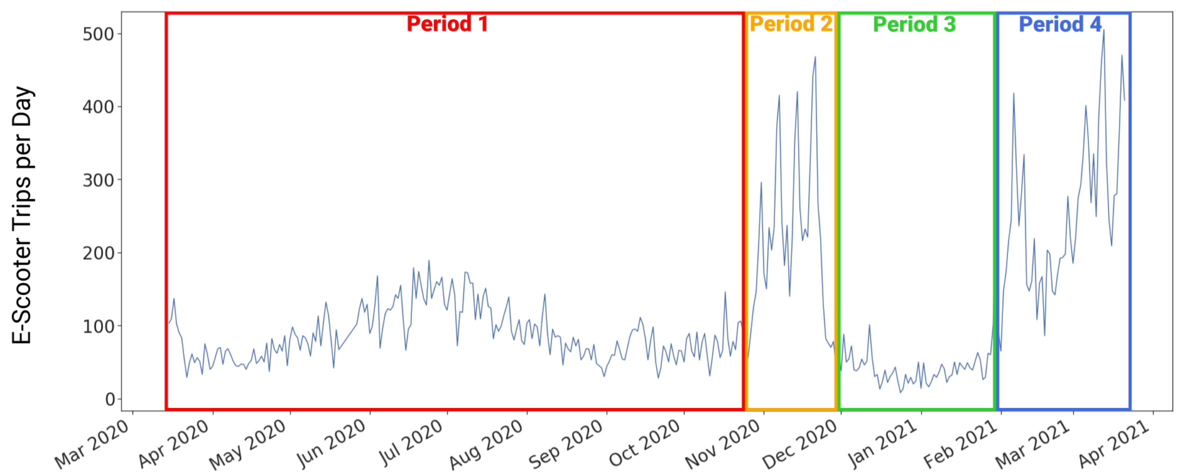


Figure 12. Study II Periods of Interest

Figure 12 shows a macro-level view of ridership trends, informing the reader of the periods of interest for which further analysis is required. Per findings in Study I, increasing e-scooter availability in optimized locations where micro-mobility need is evident is important for improving the service offered to the community. As such, it is hypothesized that the increase in ridership in Period 2 is explained by a significant shift in how operators distributed their e-scooter fleet during this time. This hypothesis is further investigated by diving from a macro-level view to a micro-level view, examining e-scooter availability and utilization for each of the four periods at Census block level resolution.

Additionally, findings from the Charlottesville Dockless Mobility Pilot Program indicated that improved monitoring practices were required to ensure micro-mobility vehicles were equitably distributed per regulation [33]. Charlottesville Dockless Mobility Regulations state that operators must distribute 10% of their fleet to designated equity zones [15]. Additionally, operators are required to implement community outreach to promote micro-mobility services in low-income communities. To this end, Study II presents methods for meaningful data aggregation and visualization to measure operator performance towards meeting the requirements as stated in regulations. City planners may use the methods and results presented in this work to evaluate micro-mobility distribution equity.

4.2 Methods

4.2.1 Analyzing Micro-mobility Operations at Census Block Resolution

To collect sufficient data for trend analysis, e-scooter GPS trace records were harvested from a real-time GBFS data feed for one year. The data collection and trip extraction methods are detailed in section 3.2. The entire study period spanned from March 15, 2020 until March 21, 2021. In total, the database storage size required for the raw GPS trace data was only about 3.1 gigabytes total, which is a small cost for the information value that can be extracted from such a dataset. This measure is also useful for city government officials and planners to account for projecting data storage allocation needs.

To increase the resolution at which e-scooter fleet utilization is examined compared to Study I, e-scooter availability is aggregated at the block level. Recall that in Study I, data was aggregated at the Census block group level. In comparison, the Census block is the smallest geographic unit used by the U.S. Census Bureau [56]. The data was split into the respective periods of interest as containers for further analysis. Then, measures of e-scooter availability and e-scooter utilization per block were calculated for each period.

To calculate e-scooter availability, a slice of all available e-scooters per hour per day is extracted from the raw data. Then, the number of scooters available per block is summed across each block in Charlottesville city and Albemarle county shown in equation (1) below. Next, the mean is taken across all the hours of each day and then again across all days in each period. To calculate e-scooter utilization, the number of trips taken per day per block is summed and then mean is taken across each hour of the day. Then, the mean is taken across all days in each period.

Lastly, the hourly trip average across each period is divided by e-scooter availability measures, summarized in equation (2) below.

$$(1)\ scooter\ availability\ per\ block = \sum scooter\ \forall_{block\ \epsilon\ blocks} \forall_{hours\ \epsilon\ day} \forall_{days\ \epsilon\ period}$$

$$(2)\ scooter\ utilization\ per\ block = \frac{\sum scooter\ trip\ \forall_{block\ \epsilon\ blocks} \forall_{hours\ \epsilon\ day} \forall_{days\ \epsilon\ period}}{\sum scooter\ \ \forall_{block\ \epsilon\ blocks} \forall_{hours\ \epsilon\ day} \forall_{days\ \epsilon\ period}}$$

Next, e-scooter availability is visualized per block per period to investigate if e-scooter fleet distribution changed significantly to explain the spike in e-scooter ridership towards the end of October. Further, e-scooter utilization per block per period is visualized. Recall that Study I results suggested utilization is a useful measure for improving micro-mobility fleet distribution. Using these visuals, observed operator behavior and its effect on e-scooter ridership trends are analyzed.

4.2.2 Evaluating Equitable Distribution and Access

To measure fleet distribution in designated equity zones, the measures of e-scooter availability per block per time slice detailed in section 4.2.1 were spatially joined with the bounding polygons indicating equity zones. Then, this value was divided by the max number of e-scooters available per time slice to calculate fleet distribution per block inside an equity zone. This value is then summed across blocks per day to arrive at a measure of e-scooter fleet distribution to equity zones per day, which is plotted for visualization. This value is also aggregated across each period of interest in order to visualize how e-scooters are spatially distributed across blocks inside equity zones. Lastly, the measures of e-scooter trips per block per time slice described in section 4.2.1 were summed across blocks to arrive at a measure of trips taken from equity zones per day. These values are then plotted against total e-scooter trips for comparison and discussion.

4.3 Results

By comparing the visualization of average e-scooter availability per block in Period 1 to Period 2 in Figure 13, it is clear that operators did make significant changes in e-scooter fleet distribution at the end of October. Per findings in Study I, operators had e-scooters largely concentrated in downtown Charlottesville in Period 1. It can be observed that, on average, there were 25 e-scooters packed into one small downtown block and about 12 in another block downtown. The rest of the fleet is scattered widely across blocks in Charlottesville and surrounding Albemarle blocks. Then, in Period 2, e-scooter availability around the University of Virginia increased significantly. E-scooters were also more contained within the VeoRide service area. The changes operators made in e-scooter placement explains the spike in ridership that occurred in Period 2.

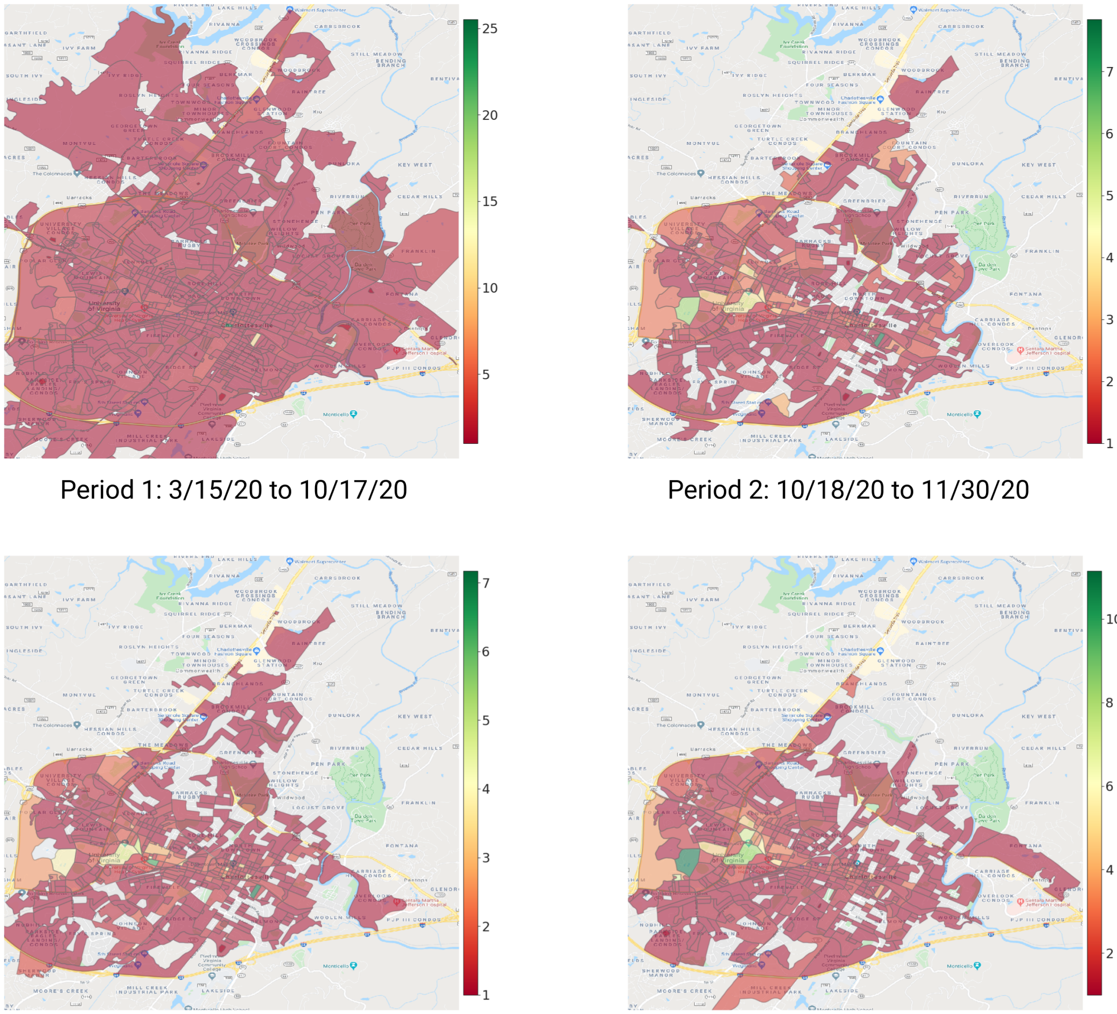
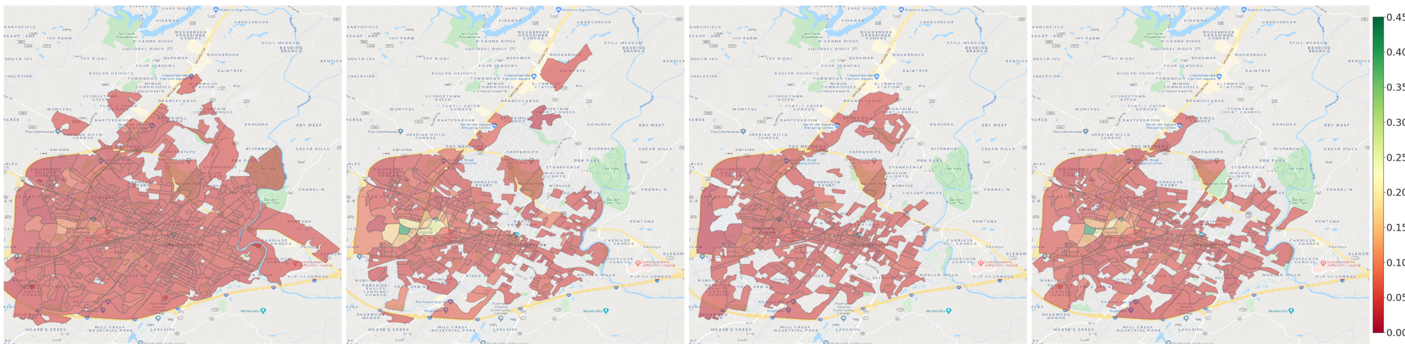


Figure 13. Average number of e-scooters available per block

Comparing e-scooter availability from Period 2 to Period 3, the distribution looks very similar and no significant changes are observed. However, the drop in ridership in Period 2 does align with when the student population departed from Charlottesville after the UVA Fall 2020 semester ended. Similarly, the ridership peak in Period 3 can be explained by the same reasoning as the start of the spike aligns with the start of UVA's Spring 2021 semester. Results indicate that e-scooter operators adjusted e-scooter placement to target the student population as e-scooters were under-utilized downtown. Ridership spikes indicate that the student population did benefit from this change. Next, e-scooter utilization changes across each period are visualized in Figure 14.



Period 1: 3/15/20 to 10/17/20 | Period 2: 10/18/20 to 11/30/20 | Period 3: 12/01/20 to 1/28/21 | Period 4: 1/29/21 to 3/21/21

Figure 14. E-scooter utilization per block

Study I findings suggested that utilization is an important metric for evaluating e-scooter placement. Scanning Period 1 from Figure 14, utilization measures appear poor on average across all blocks, indicating that the service could be improved by redistribution. In Period 2, e-scooter redistribution was observed and, as a result, e-scooters are better utilized in the blocks around the University of Virginia. This suggests that the change in fleet distribution benefited the student population specifically. After the outflux of students during the winter break of 2020 to 2021, utilization measures dropped again in Period 3, further suggesting that fleet distribution was targeting the student population. Then, when students returned for the Spring 2021 semester, utilization rates around the UVA blocks increased once more.

Figure 15 illustrates the percent of e-scooters distributed to equity zones over time while the horizontal line marks the 10% requirement enforced by regulations. Over the entire study period, the average e-scooter fleet distribution percentage in designated equity zones is 9.33%, which is slightly below the required 10% required. Importantly, the daily distribution shown in Figure 15 reveals that this distribution fluctuates significantly over time. Furthermore, Figure 15 shows that during the major shift in operations at the end of October 2020, this included a significant and consistent reduction in the percentage of the e-scooter fleet distributed to equity zones in Periods 2 through 4. While e-scooter fleet distribution to equity zones appears to be increasing from June 2020 to July 2020, an inspection of the max number of e-scooters available in the lower plot of Figure 15 confirms that this effect is accompanied by a trending reduction in the number of e-scooters available. Findings suggest that operators must balance prioritizing equitable distribution when making strategic shifts in distribution overall.

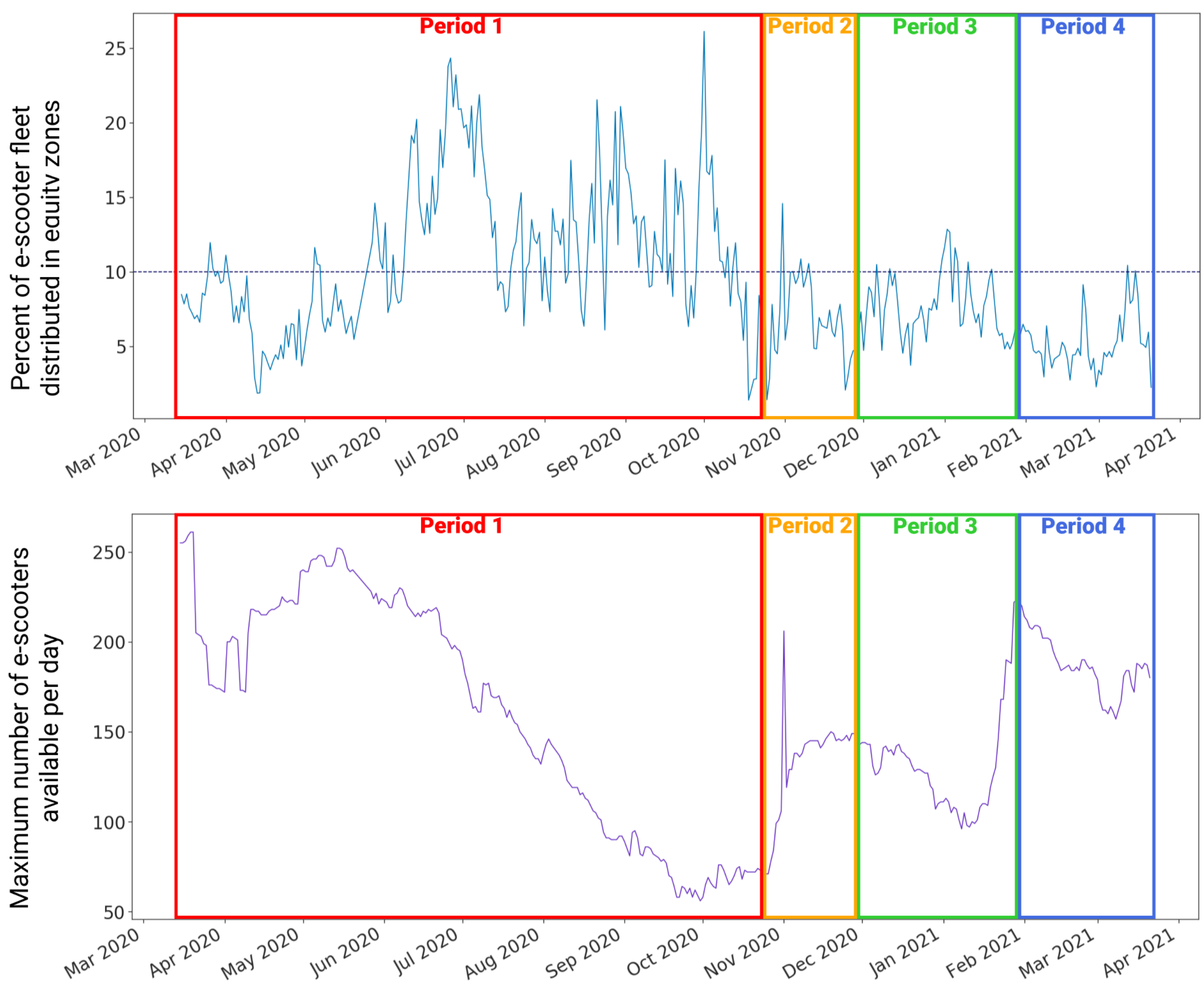


Figure 15. E-scooter fleet distribution in equity zones

Figure 16 below shows the average spatial distribution of e-scooters in equity zones for each of the four periods of interest. The equity zones, which are designated as Census Tracts, are outlined in bold purple lines in Figure 16.

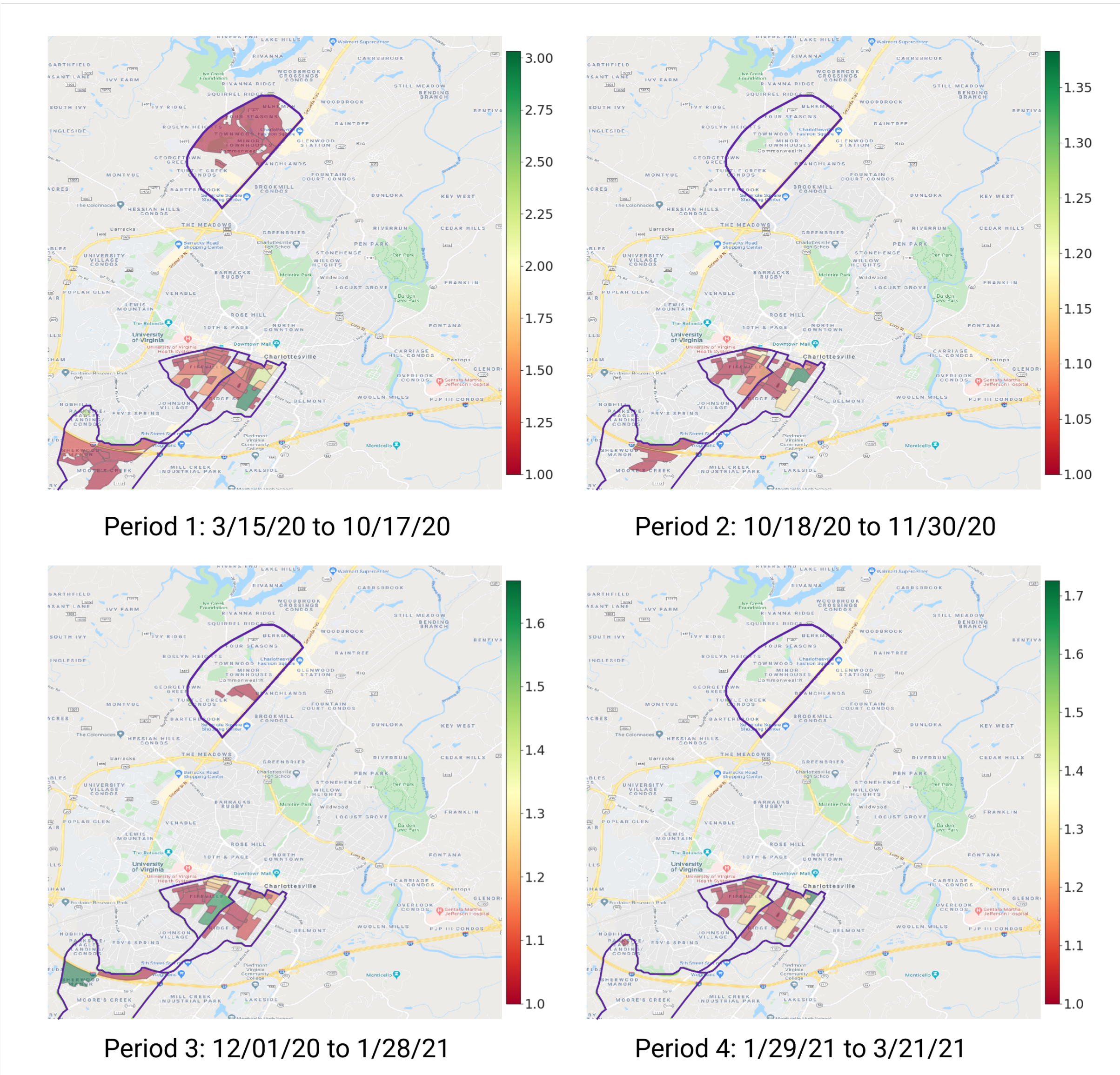


Figure 16. E-scooter fleet spatial distribution in equity zones

Lastly, to understand if additional community outreach is needed to promote ridership from designated equity zones, the number of trips from equity zones are visualized in Figure 17.

Visualized alongside the total number of e-scooter trips taken per day, it is clear that the spikes in ridership seen in Period 2 and Period 4 are not accompanied by spikes in ridership from equity zones, indicating that outreach and redistribution efforts are needed. These results are further discussed in the next section.

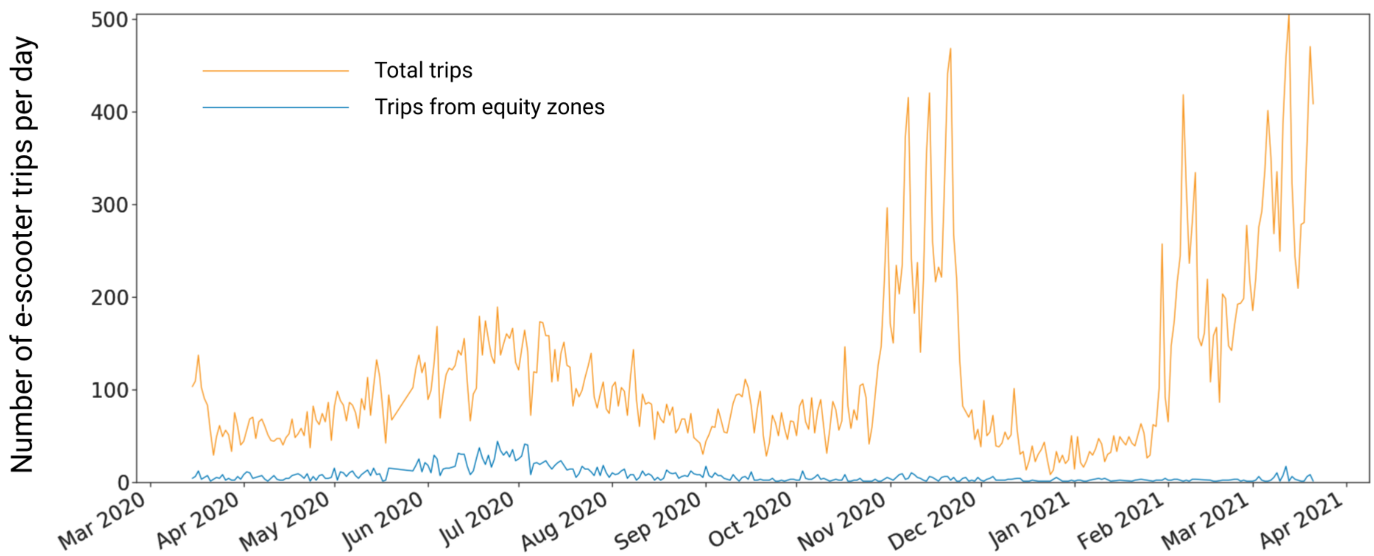


Figure 17. Daily e-scooter trips from equity zones v. total daily trips

4.4 Discussion

Previous studies indicated that the lack of existing research into how e-scooter operators deploy and reposition their fleets is a significant limiting factor [34]. Study II contributes to this research area by characterizing operator behavior using data aggregation and visualization techniques. The results from Study II demonstrate the importance of optimizing e-scooter operations while maintaining equitable distribution. Additionally, the importance of effective visualization efforts are emphasized for operations monitoring and evaluation. A discussion with Amanda Poncy, Charlottesville's Pedestrian and Bicycle Coordinator, confirmed that VeoRide hired a local manager at the end of October. The local manager's operations changes are clearly shown in Figure 13 and those changes resulted in an increase in e-scooter utilization in Periods 2 and 4 as shown in Figure 14. However, it is also clear that the updated distribution is skewed to benefit the student community as the increased benefit is not evident in Period 3 when students are largely not in town. By leveraging utilization as a guiding metric, results suggest it would be beneficial to distribute e-scooters in a manner that would better serve the local community, especially when students are not in town.

Moreover, visualizations from Study II indicate that operators must increase e-scooter fleet distribution in designated equity zones. By plotting average fleet distribution in equity zones daily, planners can track how this measure is trending against target goals written into policy. Figure 17 shows that operation changes in e-scooter placement at the end of October benefitted the UVA community, but the ridership increases were not evident from equity zones. This suggests that the current service is not as accessible for people in low-income communities, indicating a need to improve outreach efforts. Additionally, these results could motivate regulators to update policy to ensure operators improve equitable distribution and outreach as required.

Further research is needed to understand why e-scooters are overwhelmingly used by the UVA population compared to the local Charlottesville community. Perhaps outreach targeting the local Charlottesville community about micro-mobility and discount programs would help increase utilization across the entire service area. Additionally, further research is needed to identify optimal areas for e-scooter redistribution in real time.

5 Study III: Characterizing User Behavior using Latent Dirichlet Allocation

5.1 Motivation

Studies I and II focused on characterizing operator behavior in order to understand how operator decisions on e-scooter placement impacts the level of service offered to the community.

However, one major piece of the system remains under-explored – given where e-scooters are placed, where are users scooting to? Study III aims to characterize user behavior by diving deeper into the e-scooter trip data extracted from the harvested raw GPS trace data. By examining the types of routes e-scooter users traverse, researchers can better understand micro-mobility needs in Charlottesville and the surrounding Albemarle county areas.

When considering the added benefit of a new service, city planners often want to know how a new service complements existing services in the city. With respect to transportation services, city planners might ask if an added service allows the community to traverse new types of routes to access different areas around a city. They may want to know the top routes that users of a new service prefer. Study III aims to extract the overarching types of e-scooter trips taken by current users using Latent Dirichlet Allocation.

5.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative mixture model for collections of discrete data developed by David M. Blei, Andrew Y. Ng, and Michael I. Jordan [27]. As the model was created in the context of text analysis, it assumes that each document in an unstructured collection of documents is made up of a mixture of topics [27]. The documents are modeled using a hidden Dirichlet random variable and the output of the model is a probabilistic distribution on a latent, lower-dimensional topic space [27]. In other words, from an unstructured collection of documents, the model extracts relevant topics. LDA is frequently referred to simply as "topic modeling" as it is a method for discovering the hidden topics present in a collection of documents [63]. Beyond text data, topic modeling has been used to find patterns in image data, social network data, and even genetic data [63].

Recall that an extracted e-scooter trip is a set of discrete GPS points. Although it is not readily obvious, trips can resemble sentences where a GPS coordinate pair is treated like a word. A trip can then be viewed as type of document containing many GPS point words. Viewed in this way, when there exists a large collection of trips that are otherwise unstructured, topic modeling is a very fitting analysis technique. Topic modeling can be used to extract the hidden route patterns characterizing a collection of trips. This is an effective approach because GPS points that are popular across trips will recur across several different trips in the dataset. To further support this method, another study used LDA to reveal hidden trip pattens from a collection of taxi trips [64] and another found departure and arrival trends from bike-sharing trips in Paris [65]. Study III extends upon existing work by applying LDA to estimated trips harvested from a real-time micro-mobility feed. The following section details the methods for using LDA to extract latent e-scooter trip themes. Then, the extracted topics are frozen and investigated to determine how the trip topic distribution changes when the student population is present in Charlottesville compared to when it is not. This last step sheds light on which types of e-scooter trips are popular amongst the student population compared to the local Charlottesville population.

5.3 Methods

In order to perform topic modeling, the e-scooter trip data must be prepared the same way text data is prepared as an input to LDA. The raw trip data has GPS coordinates at very high resolution out to four decimal points. In order for success in this method, there must be GPS point co-occurrence across trips for meaningful topics to emerge. Thus, points that only have differences out to the 4th decimal point, but generally describe the same point in a route should be grouped together. This is similar to a pre-processing step in text analysis called "stemming," which groups words sharing the same root together [66]. For example, the words “work”, “worked”, and “working” would be grouped together because they have the same root word, “work”. In order to complete a similar pre-processing step with the e-scooter data, the first step is to round the longitude and latitude points to three decimal points, effectively grouping similar points to one root word. Every rounded point then should be considered as a spatial bin of 0.0001 degrees which equates to 36.4 feet in latitude and 28.8 feet in longitude [67]. Then, longitude and latitude fields are concatenated for each trip to create “words” for topic model analysis. Prior to rounding, there were 52, 879 unique longitude-latitude pairs. Even in traditional text topic modeling, there is a performance cost to large vocabularies and it is typical to reduce the vocabulary to the top 5,000 most frequently used words. In the aforementioned rounding step in this case, the vocabulary is then reduced to a manageable 3,312 unique longitude-latitude pairs.

To implement topic modeling on the e-scooter trip data, a popular machine learning library called scikit-learn is utilized [68]. The required input for scikit-learn’s LDA function is a document-term matrix. Figure 18 illustrates how to fit GPS trip data into this context by creating a trip-point matrix.

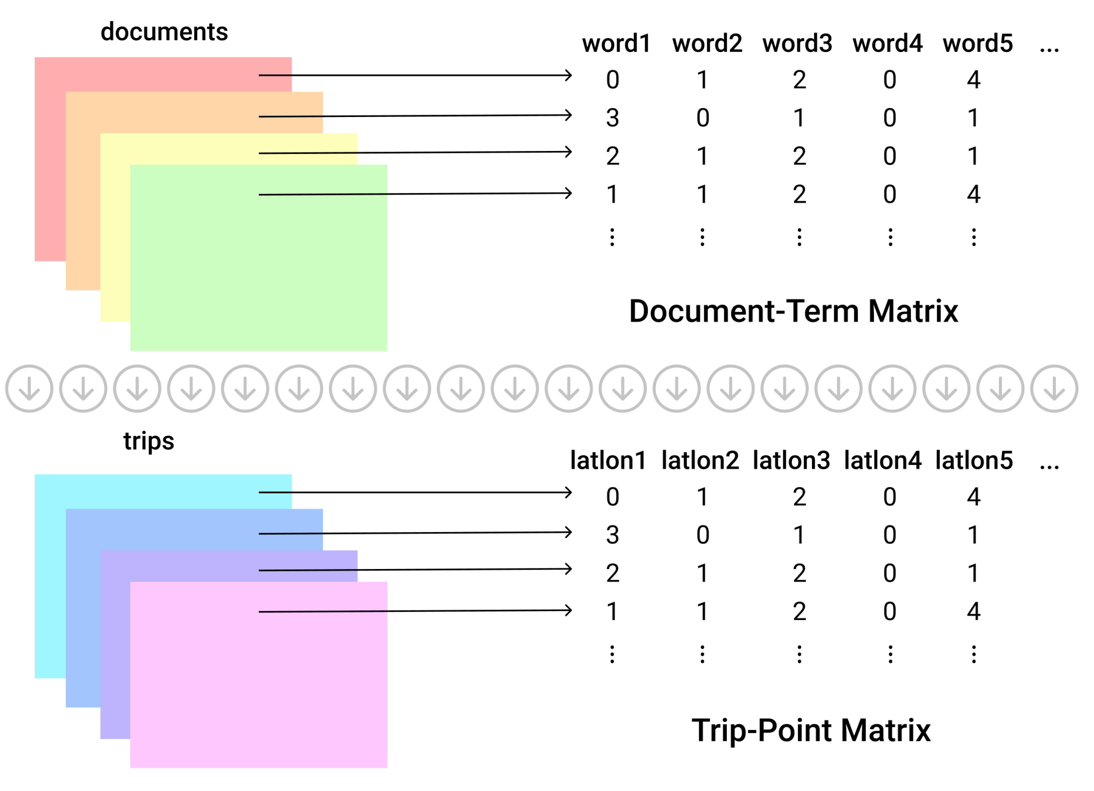


Figure 18. Document-Term Matrix to Trip-Point Matrix

Recall that each trip is treated as if it is a document. For 42,301 extracted trips, a trip-point matrix of size 42,301 rows by 3,312 columns (number of trips, number of unique GPS points) is created where the values indicate how many times a GPS point from the vocabulary occurs in each trip. Next, the trip-point matrix is fed as input into the LDA model and the number of components is set to fifteen. The output of the LDA model is then fifteen topics with a probabilistic distribution of GPS points that make up each topic. Additionally, the probabilistic distribution of topics each trip is made of is extracted from the model.

In topic modeling for text, topics are visualized by extracting the most frequent words per topic [63]. The analyst can then determine the latent topics in a corpus by examining these word groupings per topic. For this study, the top 50 GPS points per topic were plotted over a map of Charlottesville to visualize the extracted trip topics. To plot the GPS points, the top words were joined back to the rounded GPS coordinates prior to concatenation. Then, topics were manually labeled by visual inspection of the plot.

Lastly, three periods of interest from Study II are examined: Period 2 which includes the UVA Fall 2020 semester, Period 3 which spans the winter break between academic seasons, and Period 4 which includes the UVA Spring 2021 semester. Then, plots of the temporal e-scooter trip distribution across days of the week and hours of the day per period are compared. This allows the reader to examine how the trip topic distributions compare during and outside of academic seasons when the population in Charlottesville changes dramatically. The results of this study are presented in the following section.

5.4 Results

Figure 19 illustrates the distribution of the 15 topics extracted from the LDA model. While there is overlap across some of the topics, hidden trip themes clearly emerge when the top 50 GPS points of each topic are plotted. The legend on the top right shows the trip distribution per topic. Over the entire study period, the trip topic distribution is surprisingly even, with only a few trip topics being slightly more popular than the others.

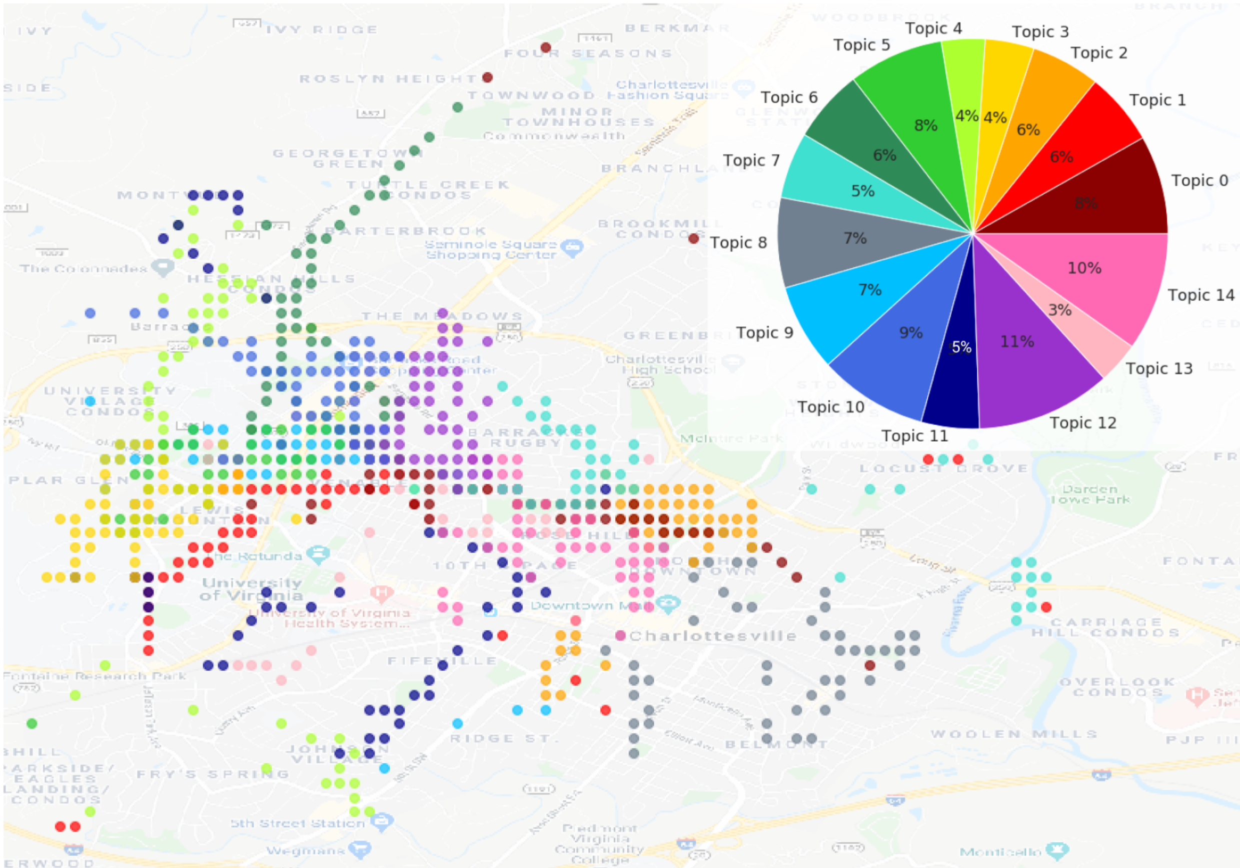


Figure 19. Trip Topic Distribution

Next, Figure 20 is a visualization of select topics where the probability distribution of all GPS points per topic is emphasized on a color scale. When a point occurs in a topic more frequently, the fill becomes a darker pink. The resulting visualization is a plot showing GPS points by their level of importance in each a topic. Then, by visual inspection of the map, each topic was manually labeled. Figures 20 and 21 also illustrate the diversity of trips traversed using e-scooters.

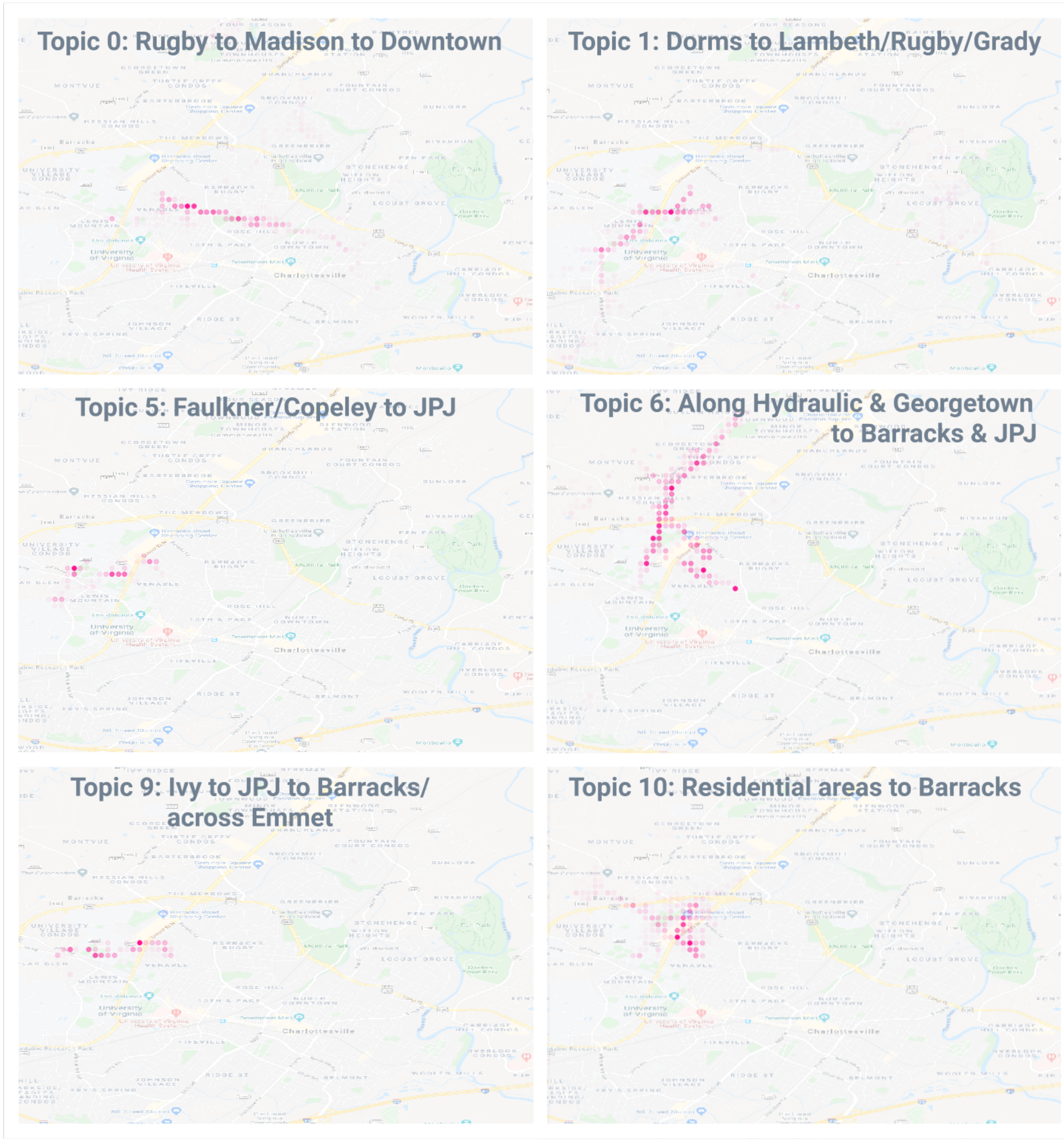


Figure 20. GPS Probabilistic Distribution per trip topic

Finally in Figure 21, the trip topic distributions were compared across three periods to examine how the types of e-scooter trips differ in and outside academic seasons. Topics 5 and 9 emerge as significantly more popular trip topics when students are in town. Topic 5 includes points connecting Faulkner and Copeley student housing to the JPJ parking area and stadium. Interestingly, Topic 9 is around the same area but also includes connections to the Barracks shopping center and surrounding residential areas. There is also a smaller increase in Topic 10 when students are in Charlottesville. Topic 10 seems to connect Barracks and surrounding residential areas.

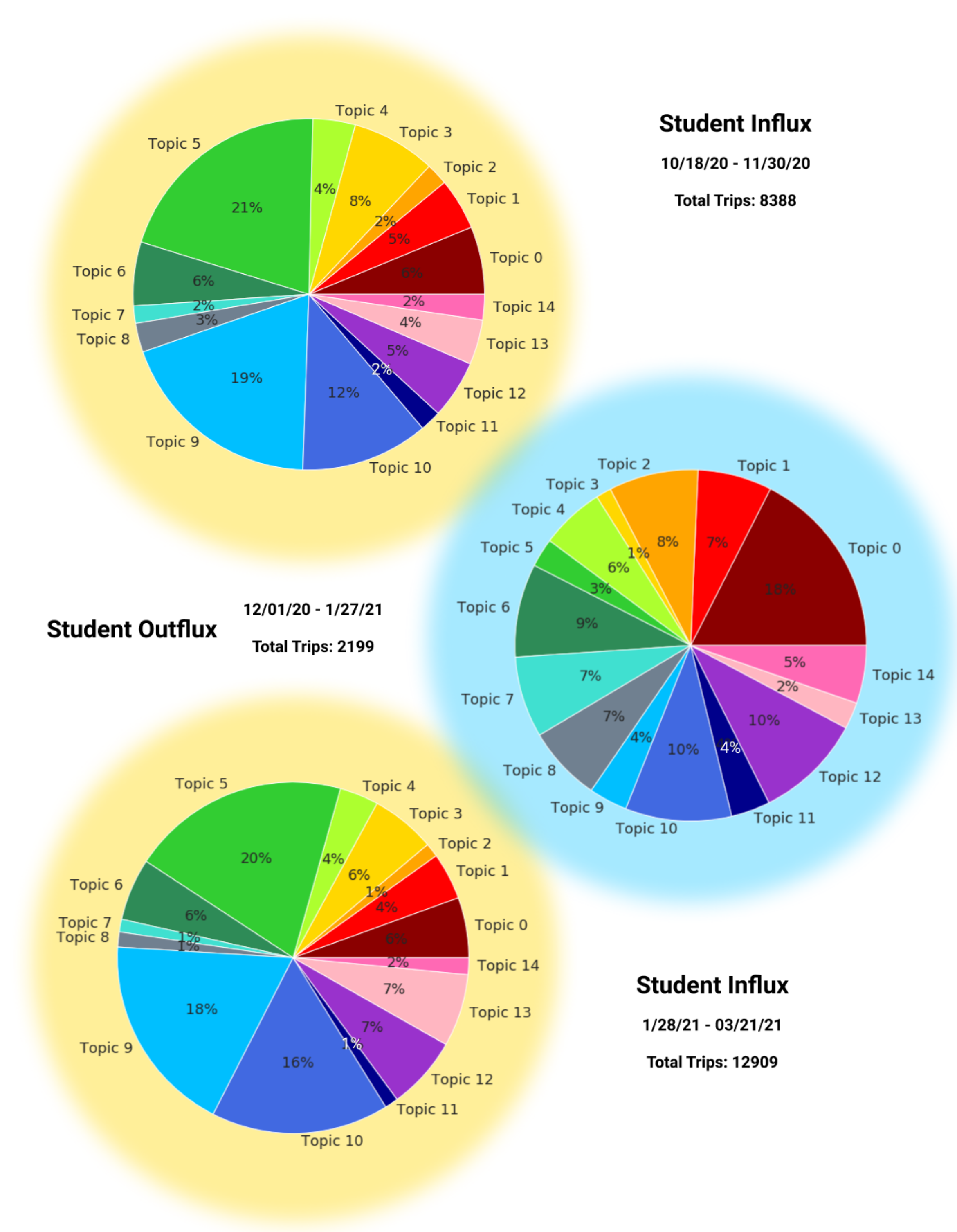


Figure 21. Trip topic distribution per period

The pie chart shown in the middle of Figure 21 provides a window for examining the local Charlottesville population when students have largely left town. Topic 0, which appears to connect residential areas around Rugby Road and Madison Avenue to Downtown Charlottesville surfaces as the most popular trip topic during this period. Lastly, Figure 22 shows how trip trends change between periods. The plots reveal that when there is an influx of students into the community, e-scooters are used mainly on the weekends, suggesting that trips lean towards recreational purposes.

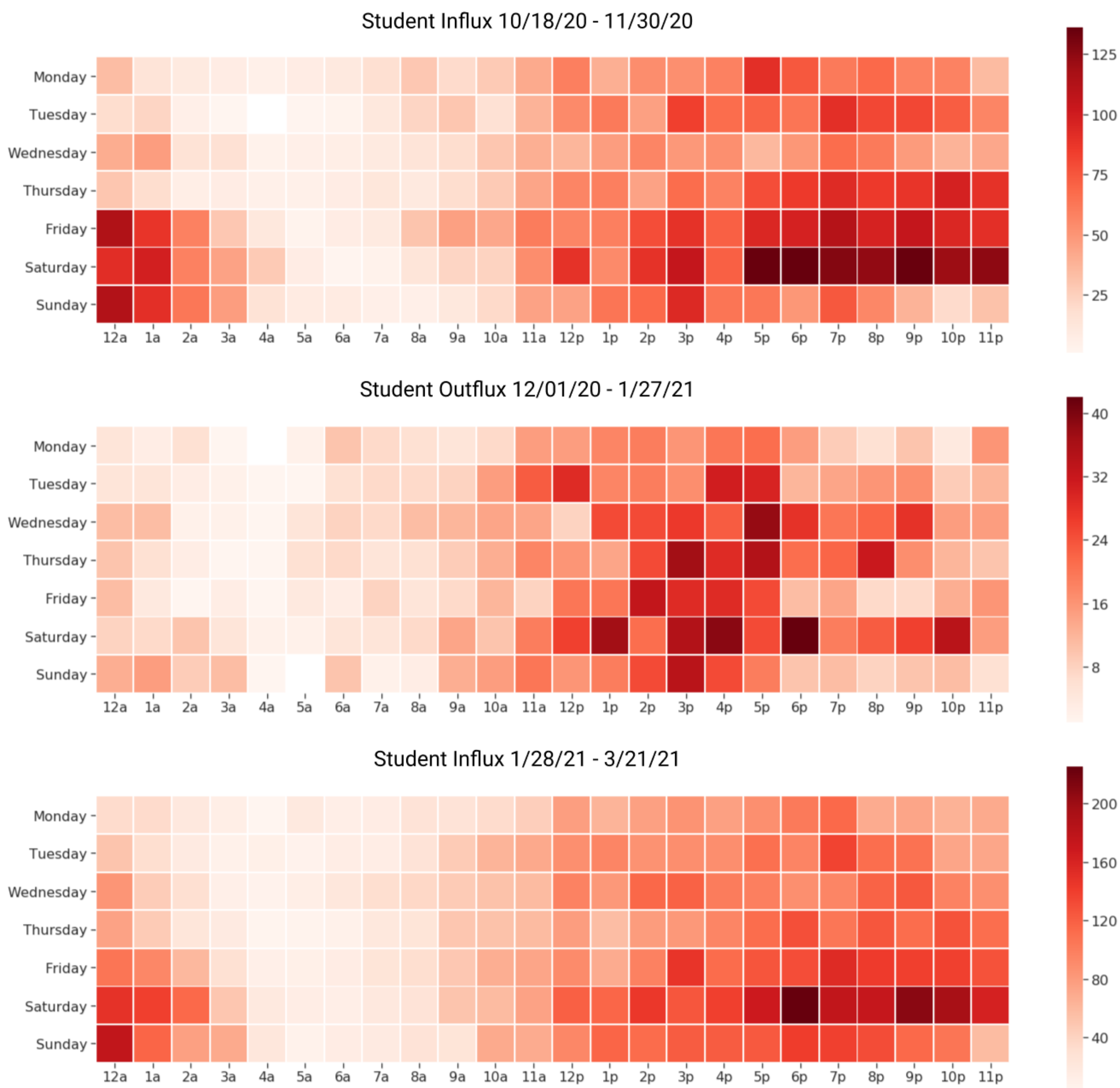


Figure 22. E-scooter temporal trip distribution per period

In comparison, trips are more evenly distributed across the days of the week outside of academic seasons. These results are discussed further in the following section.

5.5 Discussion

Study III demonstrates the efficacy of using Latent Dirichlet Allocation to characterize user behavior in shared micro-mobility systems using harvested data from an open real-time feed. Using LDA for trip topic discovery, latent trip themes that characterize all e-scooter trips taken from March 15, 2020 to March 21, 2021 were extracted and visualized. Findings suggest that e-scooters are used to traverse the streets of Charlottesville in a diverse manner overall as the extracted topics extended all across town. As expected based on findings from Study I and Study II, the most popular trip topics include areas where e-scooters are the most consistently available – downtown and around the University of Virginia. Additionally, the UVA student population seems to prefer trips around John Paul Jones Arena, nearby student housing, and the Barracks shopping center as these topics are significantly more popular during academic seasons. Importantly, it must be noted that the period of study is during the COVID-19 pandemic when students may prefer e-scooters over other forms of transportation such as public transit for public health reasons.

As the Barracks shopping center contains a Harris Teeter frequented by students, these trips could be related to weekly groceries access, but further research is needed to confirm this. Then, focusing on Period 3, Topic 0 connecting residential areas across town to Downtown Charlottesville emerges as the most popular trip topic. There were also increases in Topic 7 which connects a larger residential area to downtown and Topic 6 connecting northern residential areas to Barracks. This suggests that the local Charlottesville population uses e-scooters as a means to travel short distances across town. Further, the temporal distribution of trips during Period 3 indicates that the local Charlottesville population may use e-scooters for more practical purposes because trips occur more evenly across the days of the week. By examining topic distribution variance during periods of known population flux, researchers and city planners can learn about the main use cases given current e-scooter distribution strategies. Findings suggest that the use cases did change significantly depending on UVA student population flux. Visualizing trip topic distribution dynamically can be a useful method to evaluate the impact of redistribution decisions on the community.

6 Conclusion

6.1 Synthesizing Studies I, II, & III

To summarize, this work adds to the big data transportation literature by presenting data processing, aggregation, visualization, and machine learning techniques to holistically characterize shared mobility system behavior using streaming GPS trace data from an open feed, focusing on an emerging micro-mobility system in Charlottesville, Virginia as a case study. Previous micro-mobility studies cite a lack insight into operator distribution strategies as a limiting factor in the small, but growing research field [34]. Although e-scooters have the potential to fill small distance transportation needs including short commutes and connecting people from their homes to inconveniently far transit options, there is limited evidence showing these use cases in practice [13], [29]. Many studies suggest e-scooters are used more so for recreation [37]. Comparatively, commuting behavior is observed in docked bikesharing, suggesting that vehicle availability is critical for practical, routine trips [9]. Importantly, this work focuses on the conditional factor of e-scooter availability, hypothesizing that e-scooter use is significantly impacted by how fleets are distributed. In three studies, this work showcased data analytics approaches and shared insight on how e-scooter availability effects e-scooter use.

1. In Study I, a multiple regression analysis revealed that commute influx had the biggest effect on e-scooter availability as most scooters were concentrated in one block group in Downtown Charlottesville where many residents commute to. In comparison, the percent of residents using transit and walkability emerged as significant explanatory variables for e-scooter utilization. This indicates that although e-scooter placement decisions focused on buzzing economic factors, e-scooter ridership decision-making was based on micro-transit need and built environment factors. Visualizing utilization per block group, it was clear that an effective redistribution would improve the service for the community.
2. In Study II, data aggregation and visualization approaches were used to characterize and evaluate operator behavior. By visualizing utilization at high resolution across 4 distinct periods, a significant shift in e-scooter distribution was observed, which explained a significant rise in ridership. However, the utilization increase was only observed during periods when the city had a large influx of UVA students, indicating that the redistribution was only beneficial for the student population and had limited service improvements for the Charlottesville community as a whole. Additionally, visualizations revealed a need to improve equitable distribution efforts and community outreach.
3. In Study III, Latent Dirichlet Allocation was used to characterize the e-scooter trip behavior which describes micro-mobility use in Charlottesville during the entire study period. As expected, the most popular e-scooter trip themes included areas where e-scooters were the most available. The extracted trip topic distributions changed significantly when students left town. Seeing that utilization rates also dropped

sharply when students departed, it is clear that students have different trip preferences and needs compared to the local Charlottesville population. Additionally, results show that most trips taken during the academic seasons took place on weekend evenings, suggesting student trips leaned towards recreation. In Period 3, outside of academic seasons, trips were more evenly distributed across the days of the week, suggesting that the local Charlottesville population uses micro-transit for different purposes. As such, redistribution and outreach efforts including the local community are recommended to improve the reach and benefits of micro-mobility service in Charlottesville overall.

City planners and shared mobility service operators may leverage the methods presented in this work as a starting point to understand, monitor, and evaluate emerging transportation services that publish real time data to an open feed. Resulting data-driven decisions about vehicle distribution can align access with need, improving the level of service offered to communities.

6.2 Limitations

While this work revealed interesting insights about e-scooter operations and use in Charlottesville, it should be noted that the study period starts during the early stages of the COVID-19 pandemic, which caused a disruption to micro-mobility as a whole. The population in Charlottesville was unusually low due to the mass exodus of students in the second week of March 2020 as evidenced in Figure 4. Perceptions about travel were likely impacted by public health concerns about contracting the COVID-19 virus. For example, the finding in Study I showing increased utilization rate of e-scooters in more transit-dependent areas could be the result of anxiety surrounding using public transportation modes during this period. Similarly, trip topics extracted from Study III could be reflective of pandemic era trips to buy food or groceries as an alternative to public transportation.

Additionally, the small size of the case study area translates to a relatively small service area for e-scooter travels overall, limiting resulting conclusions to Charlottesville and perhaps other similar university towns. Finally, this work was built on the assumption of accurate GPS trace data. While obvious GPS errors were removed from the dataset in pre-processing steps, additional testing is needed to confirm GPS accuracy.

6.3 Future Work

This study sets the stage for future lines of research for big data analytics in shared mobility as the type of data analyzed is widely accessible in similar formats. Future studies can apply the proposed approaches in different cities and include additional transportation modes in their analysis to understand how emerging transportation modes complement existing ones. Further, researchers can work with e-scooter operators to test the effects of rebalancing e-scooter fleets based on utilization measures and the inferred transportation needs of the communities they serve. Finally, future research can be directed towards developing dynamic tools that apply the big data analytics approaches described throughout this work for producing real-time, dynamic displays to help city planners and regulators.

7 References

[1] United Nations, “World Urbanization Prospects: The 2018 Revision,” 2018.

[2] A. Abella, M. Ortiz-de-Urbina-Criado, and C. De-Pablos-Heredero, “A model for the analysis of data-driven innovation and value generation in smart cities’ ecosystems,” *Cities*, vol. 64, pp. 47–53, Apr. 2017.

[3] C. Lim, K. J. Kim, and P. P. Maglio, “Smart cities with big data: Reference models, challenges, and considerations,” *Cities*, vol. 82, 2018.

[4] A. R. Honarvar and A. Sami, “Towards Sustainable Smart City by Particulate Matter Prediction Using Urban Big Data, Excluding Expensive Air Pollution Infrastructures,” *Big Data Res.*, vol. 17, 2019.

[5] J. T. de Souza, A. C. de Francisco, C. M. Piekarski, and G. F. do Prado, “Data mining and machine learning to promote smart cities: A systematic review from 2000 to 2018,” *Sustainability (Switzerland)*, vol. 11, no. 4. 2019.

[6] R. Giffinger, “Smart cities Ranking of European medium-sized cities,” 2007.

[7] I. Zubizarreta, A. Seravalli, and S. Arrizabalaga, “Smart City Concept: What It Is and What It Should Be,” *J. Urban Plan. Dev.*, vol. 142, no. 1, 2016.

[8] C. A. S. Machado, N. P. M. de S. Hue, F. T. Berssaneti, and J. A. Quintanilha, “An overview of shared mobility,” *Sustainability (Switzerland)*, vol. 10, no. 12. 2018.

[9] G. McKenzie, “Docked vs. Dockless bike-sharing: Contrasting spatiotemporal patterns,” in *Leibniz International Proceedings in Informatics, LIPIcs*, 2018, vol. 114.

[10] H. Younes, Z. Zou, J. Wu, and G. Baiocchi, “Comparing the Temporal Determinants of Dockless Scooter-share and Station-based Bike-share in Washington, D.C.,” *Transp. Res. Part A Policy Pract.*, vol. 134, 2020.

[11] NACTO, “Shared Micromobility in the US: 2019,” 2019. [Online]. Available: https://nacto.org/shared-micromobility-2019/. [Accessed: 29-Mar-2021].

[12] Z. Zou, H. Younes, S. Erdoğan, and J. Wu, “Exploratory Analysis of Real-Time E-Scooter Trip Data in Washington, D.C.,” *Transp. Res. Rec.*, vol. 2674, no. 8, 2020.

[13] G. McKenzie, “Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services,” *Comput. Environ. Urban Syst.*, vol. 79, 2020.

[14] K. Button, H. Frye, and D. Reaves, “Economic regulation and E-scooter networks in the USA,” *Res. Transp. Econ.*, vol. 84, 2020.

[15] “Dockless Mobility Permit Program Regulations,” 2019. [Online]. Available: https://www.charlottesville.gov/DocumentCenter/View/1343/Dockless-Mobility-Permits-PDF. [Accessed: 29-Mar-2021].

[16] “Documentation for the General Bikeshare Feed Specification.” [Online]. Available: https://github.com/NABSA/gbfs. [Accessed: 29-Mar-2021].

[17] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, “Internet-of-Things-Based Smart Cities: Recent Advances and Challenges,” *IEEE Commun. Mag.*, vol. 55, no. 9, 2017.

[18] A. I. Torre-Bastida, J. Del Ser, I. Laña, M. Ilardia, M. N. Bilbao, and S. Campos-Cordobés, “Big Data for transportation and mobility: Recent advances, trends and challenges,” *IET Intelligent Transport Systems*, vol. 12, no. 8. 2018.

[19] F. Zantalis, G. Koulouras, S. Karabetsos, and D. Kandris, “A review of machine learning and IoT in smart transportation,” *Future Internet*, vol. 11, no. 4. 2019.

[20] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, “Visual analytics of mobility and transportation: State of the art and further research directions,” *IEEE Trans.*

Intell. Transp. Syst., vol. 18, no. 8, 2017.

[21] A. J. Khattak, "Integrating big data in metropolitan regions to understand driving volatility and implications for intelligent transportation systems," in *Advances in Intelligent Systems and Computing*, 2017, vol. 454.

[22] W. Suh, D. Henclewood, A. Guin, R. Guensler, M. Hunter, and R. Fujimoto, "Dynamic data driven transportation systems," *Multimed. Tools Appl.*, vol. 76, no. 23, 2017.

[23] S. Boubiche, D. E. Boubiche, A. Bilami, and H. Toral-Cruz, "Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks," *IEEE Access*, vol. 6, 2018.

[24] K. Zhang, J. Ni, K. Yang, X. Liang, J. Ren, and X. S. Shen, "Security and Privacy in Smart City Applications: Challenges and Solutions," *IEEE Commun. Mag.*, vol. 55, no. 1, 2017.

[25] J. Tyndall, "Where no cars go: Free-floating carshare and inequality of access," *Int. J. Sustain. Transp.*, vol. 11, no. 6, 2017.

[26] E. Toch, B. Lerner, E. Ben-Zion, and I. Ben-Gal, "Analyzing large-scale human mobility data: a survey of machine learning methods and applications," *Knowl. Inf. Syst.*, vol. 58, no. 3, pp. 501–523, Mar. 2019.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, 2003.

[28] NACTO, "Shared Micromobility in the US: 2018," 2018. [Online]. Available: https://nacto.org/shared-micromobility-2018/. [Accessed: 29-Mar-2021].

[29] C. S. Smith and P. J. Schwieterman, "E-Scooter Scenarios: Evaluating the Potential Mobility Benefits of Shared Dockless Scooters in Chicago," *Chaddick Inst. Policy Ser.*, no. 1, 2018.

[30] J. Hollingsworth, B. Copeland, and J. X. Johnson, "Are e-scooters polluters? the environmental impacts of shared dockless electric scooters," *Environ. Res. Lett.*, vol. 14, no. 8, 2019.

[31] "Alexandria Dockless Mobility Pilot Evaluation," 2019.

[32] BCDOT, "Baltimore City Dockless Vehicle Program Annual Evaluation Report," 2020.

[33] "Shared Dockless Mobility Devices: Report to City Council," 2019. [Online]. Available: https://www.charlottesville.gov/DocumentCenter/View/1344/City-Council-Bike-Share-Discussion-PDF. [Accessed: 29-Mar-2021].

[34] O. Caspi, M. J. Smart, and R. B. Noland, "Spatial associations of dockless shared e-scooter usage," *Transp. Res. Part D Transp. Environ.*, vol. 86, 2020.

[35] S. Bai and J. Jiao, "Dockless E-scooter usage patterns and urban built Environments: A comparison study of Austin, TX, and Minneapolis, MN," *Travel Behav. Soc.*, vol. 20, 2020.

[36] J. Jiao and S. Bai, "Understanding the shared e-scooter travels in Austin, TX," *ISPRS Int. J. Geo-Information*, vol. 9, no. 2, 2020.

[37] J. K. Mathew, M. Liu, H. Li, S. Seeder, and D. Bullock, "Analysis of E-Scooter Trips and Their Temporal Usage Patterns," *ITE J.*, vol. 89, no. 6, 2019.

[38] R. B. Noland, "Trip patterns and revenue of shared e-scooters in Louisville, Kentucky," *Transp. Find.*, 2019.

[39] G. McKenzie, "Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, D.C.," *J. Transp. Geogr.*, vol. 78, 2019.

[40] M. Ahillen, D. Mateo-Babiano, and J. Corcoran, "Dynamics of bike sharing in Washington, DC and Brisbane, Australia: Implications for policy and planning," *Int. J.*

Sustain. Transp., vol. 10, no. 5, 2016.

[41] W. El-Assi, M. Salah Mahmoud, and K. Nurul Habib, “Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto,” *Transportation (Amst).*, vol. 44, no. 3, 2017.

[42] K. Wang, G. Akar, and Y. J. Chen, “Bike sharing differences among Millennials, Gen Xers, and Baby Boomers: Lessons learnt from New York City’s bike share,” *Transp. Res. Part A Policy Pract.*, vol. 116, 2018.

[43] K. Kim, “Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations,” *J. Transp. Geogr.*, vol. 66, 2018.

[44] The University of Virginia, “Facts & Figures.” [Online]. Available: https://www.virginia.edu/facts. [Accessed: 29-Mar-2021].

[45] US Census Bureau, “QuickFacts: Charlottesville city.” [Online]. Available: https://www.census.gov/quickfacts/charlottesvillecityvirginiacounty. [Accessed: 29-Mar-2021].

[46] T. E. Pillion, “HB 2752 Motorized skateboards or scooters,” 2019. [Online]. Available: https://lis.virginia.gov/cgi-bin/legp604.exe?191+sum+HB2752. [Accessed: 31-Mar-2021].

[47] “Longitudinal Employer-Household Dynamics.” [Online]. Available: https://lehd.ces.census.gov/data/. [Accessed: 29-Mar-2021].

[48] “American Community Survey (ACS).” [Online]. Available: https://www.census.gov/programs-surveys/acs. [Accessed: 29-Mar-2021].

[49] “Vehicle Parking Area.” [Online]. Available: https://opendata.charlottesville.org/datasets/vehicle-parking-area?geometry=-78.597%2C38.016%2C-78.377%2C38.064. [Accessed: 29-Mar-2021].

[50] "Walk Score Methodology." [Online]. Available: https://www.walkscore.com/methodology.shtml. [Accessed: 29-Mar-2021].

[51] M. Alipour and D. K. Harris, "A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training," *J. Civ. Struct. Heal. Monit.*, vol. 10, no. 2, 2020.

[52] C-Ville Writers, "The big fifty: 50 largest employers in Charlottesville and nearby counties," 2018. [Online]. Available: https://www.c-ville.com/big-fifty-50-largest-employers-charlottesville-nearby-counties. [Accessed: 01-Apr-2021].

[53] K. J. Krizek, "Neighborhood services, trip purpose, and tour-based travel," *Transportation (Amst).*, vol. 30, no. 4, 2003.

[54] J. Jiao, "Identifying transit deserts in major Texas cities where the supplies missed the demands," *J. Transp. Land Use*, vol. 10, no. 1, 2017.

[55] "Commuting." [Online]. Available: https://www.census.gov/topics/employment/commuting/guidance/commuting.html. [Accessed: 29-Mar-2021].

[56] "Census Blocks and Block Groups."

[57] R. Manduca, "The US Census Longitudinal Employer-Household Dynamics Datasets," *REGION*, vol. 5, no. 3, 2018.

[58] J. Weitz and T. Crawford, "Where the jobs are going: Job sprawl in U.S. metropolitan regions, 2001-2006," *J. Am. Plan. Assoc.*, vol. 78, no. 1, 2012.

[59] R. Crane and D. G. Chatman, "As Jobs Sprawl, Whither the Commute?," *Univ. Calif. Transp. Cent.*, no. 1443, 2003.

[60] D. T. Duncan, J. Aldstadt, J. Whalen, S. J. Melly, and S. L. Gortmaker, "Validation of

Walk Score® for estimating neighborhood walkability: An analysis of four US metropolitan areas," *Int. J. Environ. Res. Public Health*, vol. 8, no. 11, 2011.

[61] L. J. Carr, S. I. Dunsiger, and B. H. Marcus, "Validation of Walk Score for estimating access to walkable amenities," *Br. J. Sports Med.*, vol. 45, no. 14, 2011.

[62] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python," in *Proceedings of the 9th Python in Science Conference*, 2010.

[63] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, 2012, vol. 55, no. 4, pp. 77–84.

[64] J. Bao, P. Liu, X. Qin, and H. Zhou, "Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data," *Accid. Anal. Prev.*, vol. 120, pp. 281–294, Nov. 2018.

[65] E. Côme, A. N. Randriamanamihaga, L. Oukhellou, and P. Aknin, "Spatio-temporal Analysis of Dynamic Origin-Destination Data using Latent Dirichlet Allocation: Application to the Vélib ' Bike Sharing System of Paris," *Transp. Res. Board 93rd Annu. Meet.*, 2014.

[66] Anjali, G. Jivani, and M. Anjali, "A Comparative Study of Stemming Algorithms," *October*, vol. 2, no. 2004, 2007.

[67] "How much distance does a degree, minute, and second cover on your maps?" [Online]. Available: https://www.usgs.gov/faqs/how-much-distance-does-a-degree-minute-and-second-cover-your-maps?qt-news_science_products=0#qt-news_science_products. [Accessed: 22-Apr-2021].

[68] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," 2011.