AN ABSTRACT OF THE DISSERTATION OF

Sylvan Hoover for the degree of Doctor of Philosophy in Industrial Engineering presented on June 2, 2021.

Title: Incidental Sensor Networks for Human Mobility Detection

Abstract approved: ____

J. David Porter

Transportation systems need to get better by moving ever more people while consuming ever fewer resources. To build better transportation systems, planners need an accurate understanding of how people exercise mobility and tools to apply that understanding to the transportation system. Such an understanding can come through the development of existing sources of implicit and explicit mobility data and tools suitable for planners to apply the results. Transportation organizations may struggle to produce the necessary tools internally, leaving external bodies, both public and private, to pursue development.

In this research, three frameworks were developed that employ data already being collected to facilitate analysis of human mobility and improve the utilization of that analysis in its application to transportation systems. First, two new metrics as potential objectives for finding solutions to a type of Urban Transit Routing Problem (UTRP) are proposed and applied. The metrics assess the social experience of transit users and can be used to produce transit routes that may improve a rider's transit experience. In the presented case study, the improved routes increased the social metrics by 242% and 119% compared to current baseline routes. Next, the UTRP construct is again adapted to produce solutions that allow transit planners to balance the need to reduce the susceptibility of disease transmission in their transit vehicles while maintaining transit network utility for potential riders. In the presented case study, a Pareto front is produced of solutions from which a transit planner could choose what best suits their community's needs. Both the UTRP-type frameworks use a novel source of mobility data to simulate the solutions' impacts in a real-world environment. Finally, further exploring new uses of mobility data, an anomaly detection framework that leverages redundancies in sampling populations that will arise as additional sources of data are identified is developed. The anomaly detection framework provides increased quality assurance to planners as new sources of data are developed. In the presented case study, a previously unacknowledged anomaly in traffic data is successfully identified.

The three frameworks demonstrate the potential of advancing the use of additional data in transportation planning. Future work requires additional resources to support data-driven transportation planning and adapting proven practices from elsewhere to the specific US transportation needs.

Incidental Sensor Networks for Human Mobility Detection

by

Sylvan Hoover

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 2, 2021
Commencement June 2021

Doctor of Philosophy dissertation of Sylvan Hoover presented on June 2, 2021.

APPROVED:

Major Professor, representing Industrial Engineering

Head of the School of Mechanical, Industrial, and Manufacturing Engineering

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Sylvan Hoover, Author

ACKNOWLEDGEMENTS

I extended my sincere appreciation to my advisor, Dr. J. David Porter, for all his guidance, assistance, and tolerance he has provided during our time working together. My greatest thanks to my wife, Amy, for her exceptional amount of patience while listening near endlessly to many of the matters contained in this document. Thanks to my mother, Harriet, for supporting my interests in all things transportation and technical. Finally, my gratitude to my committee members, Dr. Bechir Hamdaoui, Dr. Hector Vergara, Dr. Drew Gerkey, and Dr. Wade Marcum, and many more faculty and staff of Oregon State University for without their many contributions, this would not have been possible.

CONTRIBUTION OF AUTHORS

Chapter 3: Dr. Claudio Fuentes assisted in identification of an objective.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

LIST OF TABLES

Chapter 1: Introduction

1.1 Introduction

A good transportation system can move many people while consuming minimum resources (i.e., energy and money). Realizing transportation systems that meet these requirements requires building knowledge of how people exercise mobility and understanding how this knowledge can affect the transportation system so that it best serves the needs of its users.

A central challenge for transit planners has been to develop information about mobility patterns [1]. In addressing this challenge, both active and passive mobility data collection methods (MDCMs) have been used. Active MDCMs include surveys and pneumatic road tubes, which require the deployment of specialized resources to complete the data collection and thus are more costly. Passive MDCMs take advantage of data generated for other purposes (e.g., call detail records and electricity demand) to derive the desired information. Fewer resources are demanded of passive MDCMs, but in exchange, the data that can be extracted may not directly indicate the desired information. Passive MDCMs provide an opportunity to propose and research new means of understanding mobility.

Once mobility information is available, the next stage is to employ it to improve a transportation system. Multiple factors affect decisions about what mode a person may use to get from point A to point B. To paraphrase a saying in transportation planning: "*few people are car people or bike people; most are A to B people.*" Mobility information presents an opportunity to make a person's mode decision both the easy choice and the right choice to benefit the transportation system's effectiveness for all.

This dissertation aims to utilize passive MDCMs and the mobility information that can be derived from the data to develop methodologies to design transportation systems. New approaches established for passive data sources include both extracting the information and validating it. Mobility information is applied to address objectives not previously employed in transportation planning to facilitate a transportation system to serve its users better. Neither can serve as definite solutions to problems facing transportation systems, but may better inform those charged with that task.

1.2 Background

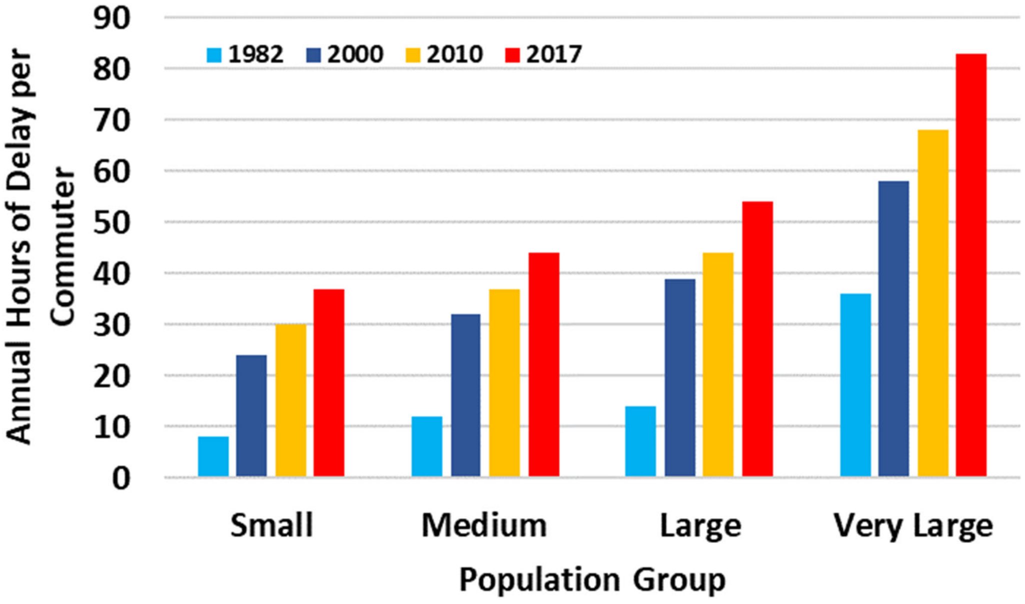
In the United States (US), traffic congestion is rising [2], average vehicle miles travelled are increasing [3], and roads are increasingly unsafe for cyclists and pedestrians [4]. These issues could be alleviated by shifting travel modes away from personal vehicles. The challenge is that the US has built its transportation system around the personal vehicle, making the personal vehicle not just a convenient choice but an easy choice.

One way to make alternative transportation modes more attractive is to understand better how people currently use the transportation system and tailor the system to their needs. However, developing this understanding presents obstacles. Methods to generate mobility information such as surveys, models, and inferences from other data sources have their shortcomings. What is needed are empirical options to capture mobility. Due to their prevalence in today's society, personal wireless devices (PWDs) could facilitate data collection to infer mobility.

PWDs are rapidly proliferating, with 96% of adults in the United States owning a cellphone [5]. PWDs contain components that transmit and receive using various wireless standards (e.g., Bluetooth and WiFi). Capturing wireless communications generated from the presence of PWDs presents an opportunity to collect a large sample of mobility.

Using PWDs to understand mobility has spanned many modes and many means of collection. The advantages of using PWDs are the increasing ubiquity of equipment capable of passively collecting data about mobility and the minimal capital costs of collecting this data when using existing equipment. The main disadvantage of PWDs that can result in data quality problems is that they were not intended to generate data about passive mobility, and thus additional processing is required. Data quality problems exist in active MDCMs as well [6–11], but the problems associated with passive MDCMs only compound such issues [12–16]. Developing additional methods to improve passive MDCMs quality will only serve to better analyses using the data.

To employ mobility data to design a better transportation system presents its own questions. What do people seek from transportation? Least cost? Fastest? Most comfortable? What is communicated frequently about the transportation system in the



Small = less than 500,000
Medium = 500,000 to 1 million
Large = 1 million to 3 million
Very Large = more than 3 million

Figure 1.1: Congestion Growth Trend - Hours of Delay per Auto Commuter [2]

US is its failings [17]. Figure 1.1 shows the increasing delay faced by vehicle commuters. Options exist outside of a world of vehicles stuck in traffic, and mobility data can serve as a pathway to better support those options.

Moving people in public transit is an example of using the transportation system more efficiently. However, making public transit efficient for peoples' lives requires designing a transportation system that fits their needs. For example, Figure 1.2 shows that transit ridership continues to decline in much of the US, especially bus ridership [18]. Therefore, more needs to be considered as to what entices or deters people from using public transit. Mobility data shows potential in contributing to this understanding, especially regarding how the design of routes and their constitute components (e.g., stops, schedule, and vehicle capacity) influences users' transportation choices.

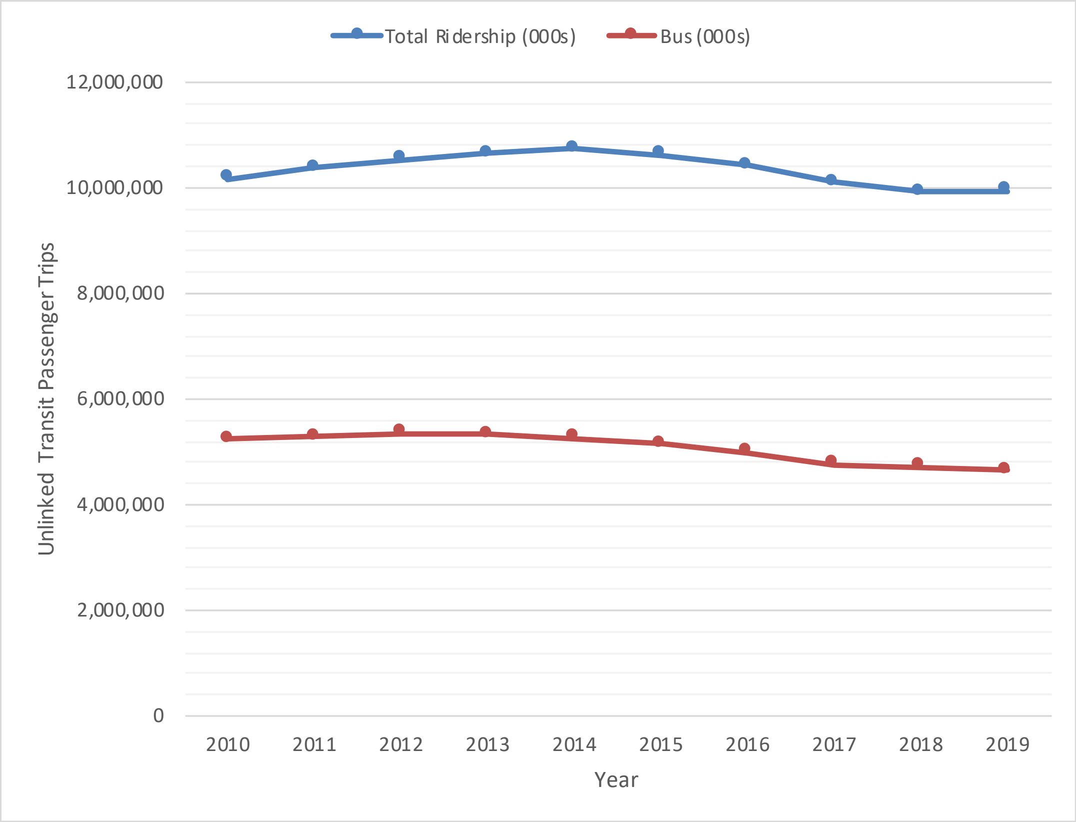


Figure 1.2: Public Transit Ridership [18]

1.3 Problem Statement

The demands on transportation systems continue to increase as populations grow and become wealthier. To accommodate the increasing demand, transportation system planning must improve to meet that demand efficiently.

Internal to transportation system planning, there is a gap in the data needed to improve transportation efficiency and the tools necessary to utilize the data to improve the transportation system. Thus, to enable improved transportation planning, both the deficiency in the data handling and the integration of the data into transportation planning must be addressed.

Passive MDCMs exist, but they must be developed and cultivated for uses outside of the initial intent for which they were collected. Potential partnerships between organizations and people not necessarily accustomed to working with each other need to be identified. Humans, unlike freight, consider more than just efficiency measures when making transportation decisions. A recognition of some of the additional factors does not currently occur in transportation system planning, but must if planning goals are to be effectively implemented.

Transportation lags behind other areas in fielding technologies. In addition, government transportation bodies struggle to attract the needed talent to develop improved approaches using state of the art technologies. The research presented in this work aims to give transportation system planners new tools to develop data sources and apply data to their work that may not have been possible to develop internally.

1.4 Dissertation Organization

The three subsequent chapters address issues identified in the problem statement with the development of implementable frameworks. Chapter 2 introduces two new metrics focusing on the social experience of riders as potential objectives for finding solutions to a type of Urban Transit Routing Problem (UTRP). Chapter 3 responds to current global events and presents a framework to produce UTRP solutions that allow transit planners to balance the need of reducing the susceptibility of disease transmission in their transit vehicles while maintaining transit network utility for potential riders. Chapter 4 develops an anomaly detection framework that leverages redundancies in data collection that will arise as sources of transportation data become better understood. Each chapter utilizes data either not traditionally used by transportation planners or combines existing sources in a novel way to provide results that transportation system planners can use to improve their work. Finally, Chapter 5 summarizes the major contributions of this research and suggests opportunities for future work that may lie at the intersection of the fields of engineering and transportation policy.

Chapter 2: Building a Socially-Aware Solution to the Urban Transit Routing Problem[1]

2.1 Introduction

Public transit agencies face a challenge many organizations encounter: do more with less. Public transit ridership in the United States has been decreasing, which is especially true for bus ridership [19]. As public transit ridership continues to decline, agencies must overcome the consequential reduction in fare revenues. Some factors drive the decrease in public transit ridership, such as fuel prices, housing density, and employment levels, but these are factors well beyond a transit agency's control. So, where can public transit agencies make impacts? Increases in service frequency have a positive impact on ridership, but not to an extent where the added cost is recouped [20]. What can public transit agencies do to get more riders with minimal capital investment costs?

A common reason riders cite for not taking public transit is a concern for personal well-being. Regardless that transit users are less likely to be positively screened for depression [21] and those living in transit-deprived neighborhoods are more likely to experience individual depression [22], unease about using transit remains. For example, feeling unsafe on public transit was a reason cited by a majority of respondents in Portland, Oregon, as to why they chose to drive [17]. The physical separation provided by a vehicle is not available in public transit, and with perceptions of safety contributing to declining ridership, agencies have limited resources available to address such concerns. One potential solution is to optimize a public transit network to increase rider comfort. Such is certainly a nice thought, but are there tangible benefits? The literature shows that, per mile traveled, a bus is safer than a personal vehicle [23]. Therefore, one may argue that feeling unsafe on public transit is more of a perception than an issue. The perception of safety through one's comfort in a public space is affected by the social ties that exist, explicitly or not, in that space [24]. Optimizing routes to facilitate passenger

[1] Under review with *Transportation Research Part E: Logistics and Transportation Review*

comfort thus mitigates one reason why a person would take a single-occupant vehicle trip.

In 1972, Stanley Milgram published "The Familiar Stranger: An Aspect of Urban Anonymity." This work was the first to raise the notion of individuals that one encounters in daily life, yet hold no additional role. Milgram's concept of the familiar stranger has expanded over the years and has been shown to influence a person's comfort in a public space. By leveraging performance metrics centered on the familiar stranger (or the social networks resulting thereof), transit routes can be optimized to maximize the comfort experienced by riders while using transit. This increase in rider comfort, and subsequent retention or increase in ridership numbers, can be accomplished with no additional resources other than those associated with a one-time route change.

2.2 Literature Review

This research aims at giving transit agencies a way to improve the rider experience. Ideally, this goal should translate into increases in ridership without requiring the significant expense demanded by capital intensive approaches such as increasing service frequency. To achieve this goal, an *urban transit routing problem* (UTRP) will be formulated to improve the social experience of riders.

In support of this goal, relevant previous literature in several areas was reviewed. Section 2.2.1 synthesizes different optimization approaches to the urban transit routing problem. Section 2.2.2 introduces the concept of the familiar stranger (FS), which is one of the metrics employed to improve rider experience. Finally, Section 2.2.3 elucidates the value of social network metrics, like clustering, when applied to social environments.

2.2.1 Urban Transit Routing Problem

Fundamentally, the UTRP aims at optimizing transit routes for a defined objective or a set of objectives. However, the UTRP manifests in various forms. For example, in addition to routes, some optimizations include stops, schedule, and vehicle outputs, among other associated decisions. A UTRP can be either single- or multi-objective, with metrics contributing to the perspective of riders, operators, or both.

In 2008, Guihaire and Hao [25] published a review paper that synthesized prior work addressing the UTRP dating back to 1925. In this work, the authors attribute the first use of evolutionary optimization (e.g., a genetic algorithm) to Xiong and Schneider [26]. Approaches have evolved since Xiong and Schneider's work, as shown in this section, but the fundamental aim of optimizing a transit network for economic efficiency has remained consistent.

The breadth of research that has been conducted on the UTRP creates challenges when drawing comparisons between one research study and another. Mumford [27] faced this challenge and thus published their source data to allow future comparisons of work whose objective was to optimize similar decision variables for a common metric. Mahmoudzadeh and Wang [28] developed a cluster-based scheduling approach to improve the alignment of a university shuttle service with travel behaviors and class times. Researchers claimed that implementation of their methodology could increase system efficiency (i.e., cost per passenger) by up to 25%. Using automated passenger counter and automatic vehicle location data from a campus shuttle service, they selected months during which travel patterns appeared consistent (e.g., middle of a term at the university), and used both supervised and unsupervised clustering methods to assess different approaches to schedule departure times to improve efficiency. Mahmoudzadeh and Wang's entire work achieved its goal without a single change to the university shuttle's physical network.

The UTRP can be adapted to model the characteristics of a specific real-world problem. A variant of the UTRP known as the School Bus Routing Problem (SBRP) manifests similar characteristics to the problem addressed with the social optimization framework proposed in this research. More specifically, both problems have riders with known origins and destinations, time constraints for when those riders are to be transported, a fixed number of available vehicles, and flexibility in stop location assignment. Recognizing the similarities, the SBRP is useful to inform the problem faced by this research.

2.2.1.1 School Bus Routing Problem

An application of the SBRP that gathered significant coverage in the general media involved the Boston Public Schools' bus network [29]. The authors introduced another variation of the UTRP called bio-objective routing decomposition (BiRD) to model and obtain feasible solutions to the bus time selection problem. BiRD breaks the initial optimization problem into numerous sub-problems and combines the solutions to solve the greater optimization. The assignment of students to stops was an optimization to minimize the number of stops and walking distance for students with beneficial additional constraints. Using the identified stops, BiRD builds routes for schools using a randomized greedy heuristic. Breaking optimization problems into sub-problems to be optimized, recognizing the different issues faced at each stage, is not unique to BiRD.

Lemos, Joshi, D'souza, *et al.* [30] explored the effect of different heuristic optimization methods (i.e., ant colony, honey bee, and greedy randomized optimizations) when finding feasible solutions to the SBRP. The results showed no heuristic optimization approach to be dominant, which suggests that situational interests (e.g., number of buses, distance traveled) drive which approach is best suited. Although this outcome does not lead to any specific heuristic to be selected for this framework, it suggests that multiple heuristics can provide solutions.

In another study, Leksakul, Smutkupt, Jintawiwat, *et al.* [31] used machine learning and evolutionary optimization to plan bus stops and routes for a factory shuttle, which is, effectively, an SBRP. The authors used six machine learning methods, including Maximin, K-means, Fuzzy C-means, Competitive Learning, and two hybrids of those four to identify stop locations based on employees' home addresses. Once stops were identified, routes were determined using an ant colony heuristic optimization algorithm. The proposed approach reduced employees' walking distance to bus stops by 79%. Again, breaking a problem of stop identification and route determination into sub-problems to employ appropriate optimization methods (be they heuristic, a machine learning method, integer optimization, etc.) proves an effective approach to solving the greater problem.

The breadth of approaches to find optimal solutions to the UTRP results in research for various areas of focus within transportation. Liu, Liu, Yuan, *et al.* [32] employed data from taxi and bus trips in Beijing, China, in a three-phase approach to predict bus route demand and optimize route planning to suit rider demands. The three phases included transportation mode choice modeling to predict mode for origin-destination (OD) pairs, optimizing bus routes to maximize the number of OD pairs likely to select the bus, and validating the two prior phases with empirical data. A weighted logistic regression was used for mode probabilities between OD pairs, whereas branch-and-bound was used to optimize the bus routes. Liu, Liu, Yuan, *et al.*'s use of OD data is similar to how OD data is used in an SBRP but has riders that can choose their mode, with the objective being to maximize the number of riders that choose mass transit. Moving closer to the social framework's objective, this research further supports the variety of methods that need to be adapted for the specific problem to be addressed.

All optimizations discussed so far lack the treatment of human passengers not as a commodity to be moved efficiently, but as social beings that may make transportation decisions beyond what is timely. Starting to account for other factors that can influence mode selection will be necessary if significant mode-shift is to be achieved.

2.2.1.2 Taxonomy of UTRP

The taxonomy of the UTRP is well laid out by Iliopoulou, Kepaptsoglou, and Vlahogianni [33]. The first challenge in identifying UTRP's taxonomy is identifying what research addresses the problem. Iliopoulou, Kepaptsoglou, and Vlahogianni primarily used the moniker Transit Route Network Design Problem (TRNDP) in their research, but noted that the problem is also commonly referred to as UTRP, Transit Network Design Problem, and Public Transport Network Design Problem. Iliopoulou, Kepaptsoglou, and Vlahogianni reference Kepaptsoglou and Karlaftis [34]'s definition of UTRP where "[t]he TRNDP is described by the objectives of the public transportation network service to be achieved, the operational characteristics and environment under which the network will operate, and the methodological approach for obtaining the optimal network design." As such, UTRP research is classified by its objectives, parameters, and methodology. Objectives and parameters are unique to their environment, but generally hold the interest of either the transit user or the transit operator (e.g., travel time for users is a user objective and the number of available vehicles is an operator parameter). Methodologies are classified as belonging to one of four groups: heuristic, analytical, mathematical programming, and metaheuristic [33, 34].

The aim of this research to optimize the social experience of riders fits well into the taxonomy as outlined by Iliopoulou, Kepaptsoglou, and Vlahogianni. The objective function of the resulting UTRP is to maximize a social metric using stop locations and routes (and resulting frequencies) as parameters. Feasible solutions to the UTRP are generated through the application of metaheuristic evolutionary algorithms. Other research has utilized the same parameters and methodologies, but the social objectives are unique to this work.

2.2.2 Familiar Stranger

Stanley Milgram first wrote about the idea of the *familiar stranger* (FS) [35] in 1972. "To become a familiar stranger a person has to be observed repeatedly for a certain time period, and without any interaction." Such a person sounds much like one that may be encountered in a transit mode that is taken regularly. Milgram speaks of a study of commuter stations performed around New York by students of the City University of New York and the experiences of those on the platform:

> They have a fantasy relationship to familiar strangers that may never eventuate in action. But it is a real relationship, in which both parties have agreed to mutually ignore each other, without any implication of hostility.

Milgram recounts a story of a woman collapsing on the street in Brooklyn. Another woman rushed to her aide; this woman had seen the now collapsed woman for years in her neighborhood, but the two had never spoken. The assisting woman "said later that she had felt a special responsibility for the woman, because, they had seen each other for years, even if they had never spoken." It is recognized that the relationships with people that can be called FSs have meaning and can impact our lives and the choices we make.

Following up on Milgram's work, Paulos and Goodman [24] updated and expanded on the concept of an FS. They conducted two studies, one a near replication of Milgram's, and the other, a walking interview with a subject to get a deeper understanding of how the physical and social environments impact the experience. The first study mostly found similar results to Milgram's, although with a lower recognized FS average (i.e., 3.1 vs. 4.0). The second study provided additional insight into the experience, including: "[p]eople most valued the number of familiar people nearby." The researchers further explored the possibility of creating a digital device to collect data on FS interactions but recognized the complexity of deployment. By utilizing existing data for analysis, new work could enable the additional analysis they foreshadowed.

Liang, Li, and Zhang [36] used existing data to propose a method to classify interactions from encounters into one of four categories: familiar stranger, in-role, friend, and stranger. Using three spatiotemporal datasets of human mobility, characteristics of encounters in the constructed encounter network were analyzed, including encounter frequency, inter-contact time, and node degrees. With their classifier, the researchers propose that such insights about encounter networks could be used for epidemiological studies.

FS analysis options are only realized when datasets are available. With public transportation being an environment where people find each other in close quarters with people outside of their typical social circles and the ongoing digitization of transit management, researchers are beginning to take advantage of this potential.

2.2.2.1 Familiar Strangers in Transit

In public transit, riders using a smart card (SC) to pay for their fare generate sets of uniquely identifiable records, which vary by system. In some systems, riders tap on and off, giving both origin and destination, while others may have the rider tap for their first ride of the day. These and other variations that occur with SC use impact the analyses that are possible. Research using SC data is currently the most developed expression of an FS-like concept in transportation.

Much as a person can become an FS through repeated visual exposure, diseases can also be transmitted by exposure. The spread of disease through proximity and the availability of spatiotemporal datasets for transit users has induced research generating encounter networks in transit to be focused on the epidemiological effects. Liu, Yin, Ma, *et al.* [37] analyzed the interaction of transit riders in Shenzhen, China, by creating an encounter network from inferring specific trains ridden and stations visited using SC data. The authors demonstrated their encounter network's application by modeling the spread of infectious diseases in a transit system. The study did not look further into applying network analyses to infer additional information other than assessing up to second-order infections in their application demonstration. Similarly, Sun, Axhausen, Lee, *et al.* [38] used SC data from Singapore to investigate how such data could be used to assess the spread of infectious diseases. As their research focused on a transmittable contagion, they identified those with interactions resulting in a high-degree node in the social network as "super-spreaders" of infectious diseases. They found that by well defining the "super-spreaders," assessing their movement "provides longer and more reliable lead-time."

Using encounter networks to assess the societal impacts of transit due to disease transmission is only one side of what such data can reveal. Moving away from solely the biological and back into the sociological, these transit-derived encounter networks can shed light on how people interact with each other. Asatani, Toriumi, Mori, *et al.* [39] used SC data from the Kansai area of Japan to study spatiotemporal co-occurrence in a transit system. The authors concluded that most of the co-occurrences were due to interpersonal relationships rather than the FS effect. One application can be seen with assessing potential social isolation of retired men and confirming the likely isolation through their lack of co-occurrences with others. Transit, an app developer of the same-named app, used SC data from Montreal to analyze the frequency of commuter encounters [40]. They focused specifically on the varying encounter patterns between routes (e.g., commuter vs. late-night). Their analysis's message is that transit can be social, and use evidence of commuters regularly encountering each other as a suggested basis for developing positive social connections. Both these analyses show the social benefit potential through co-occurrences in transit, and perhaps a simple co-occurrence measure is sufficient.

Zhang, Jin, Ge, *et al.* [41] took measuring co-occurrence one step further using SC data from Beijing, China, to produce a hidden friend metric to discover "hidden friend relations." The hidden friend metric involves two components: temporal stability and spatial stability. The weighted sum of the two sigmoid functions is referred to as the 'Stability Degree of a Hidden Friend Relation,' and it is used to build an undirected weighted graph of the discovered relationships. The researchers took their relationship groups further by assessing the points of interest near the origins and destinations of groups' trips to build a group profile. What is not established in Zhang, Jin, Ge, *et al.*'s research is the benefits of the additional analysis beyond a simple co-occurrence measure. The results produced by their study appear useful, but without comparison showing it superior to more straightforward methods developed by others.

In an attempt to connect the increasing understanding of SC data into greater societal meaning, Sun, Axhausen, Lee, *et al.* [42] used SC data, census data, and survey data to analyze the resulting encounter network of the Singapore transit system, explicitly recognizing the emergence of FS occurrences. Sun, Axhausen, Lee, *et al.* [38]'s later work looked at the transmittable contagion potential, but at this stage of their analysis, the more general societal meaning was discussed. For establishing an encounter network, the time between encounters of identifiers was used to show patterns with peaks at multiples of 24 hours. However, "more than 95% of the 494,323,272 encounters happened only once during the week," leading to the speculation that people would not be aware of the familiar strangers in their commute. Sun, Axhausen, Lee, *et al.* present questions in the earlier work [42] that are still faced in the later work [38], i.e., how to measure the familiarity in the passive familiar strangers networks, and how to define the threshold of familiarity on social diffusion processes. These questions are ones that can be explored as a means of building a more social transit network.

2.2.3 Social Network Clustering

The task of identifying social interactions in transit can be approached in numerous ways. For example, each encounter can be seen as an isolated incident contributing to a web of encounters, but not with an on-going meaning. Conversely, each encounter can be seen as a contribution to a perennial social experience and analyzed in a dynamic, evolving manner.

Watts and Strogatz [43] describe a ’small world’ network model whose topology possesses qualities like being "highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs." They note that systems with small world properties "display enhanced signal-propagation speed, computational power, and synchronizability." The properties of a small world network are described by its characteristic path length and clustering coefficient. By optimizing a transit network to produce a small world FS network, a small world network’s benefits could be realized in a social context. As an example, Agarwal, Liu, Murthy, *et al.* [44] analyzed two datasets of real-world social networks and compared the local clustering coefficient (as defined by Watts and Strogatz [43]) to those of two networks with the same node randomly generated by the Erdős–Rényi model. The results showed that real-world social networks are orders of magnitude higher than those randomly generated (0.51 and 0.69 vs. 0.001).

2.2.4 Literature Review Summary

In this literature review, both the need and possible solution methods for socially-aware route planning were shown. First, the UTRP is a well-researched problem examining how to optimize route sets. UTRP research thus informs this research in constructing a framework around which varying objectives can be pursued. Next, the FS and social network clustering provide a background on which a socially-aware objective can be built. Combining the parameters and methods of past UTRP research with the objectives built from FS and social network clustering can create a new socially-optimized route framework. Generating solutions to the UTRP in the form of routes optimized for riders' social experience serves only to make public transit more attractive to those when making a mode choice.

2.3 Methodology

Figure 2.1 illustrates the four phases of the framework developed in this research to improve the social experience of riders by developing a route plan for a transit system.

In phase I, spatiotemporal data about human mobility are processed to facilitate later information inferences. The spatiotemporal human mobility data could be real-world or synthetic but should represent the mobility needs likely to be demanded of the transit system.

In phase II, the mobility dataset obtained in phase I is used to complete network generation and trip inference tasks. The traversable paths available in the mobility dataset's spatial expanse are abstracted to a network representation to facilitate heuristic optimization. The mobility data within the dataset is processed to infer trips occurring between locations that could be served by a transit system.

Phase III involves the calculation of an objective metric, and phase IV consists of a routing heuristic optimization. Phase III and IV occur iteratively until the termination criteria are met.

2.3.1 Phase I - Mobility Data Preparation

Two data sources are used as the main inputs for the social experience improvement framework. The first data source consists of WiFi user sessions covering the area of interest and is used to extract mobility. The second data source is a network representation of navigable paths. The network scope is defined once the scope of mobility in the WiFi user session data is known. Therefore, the WiFi user session data must be prepared first before any subsequent steps.

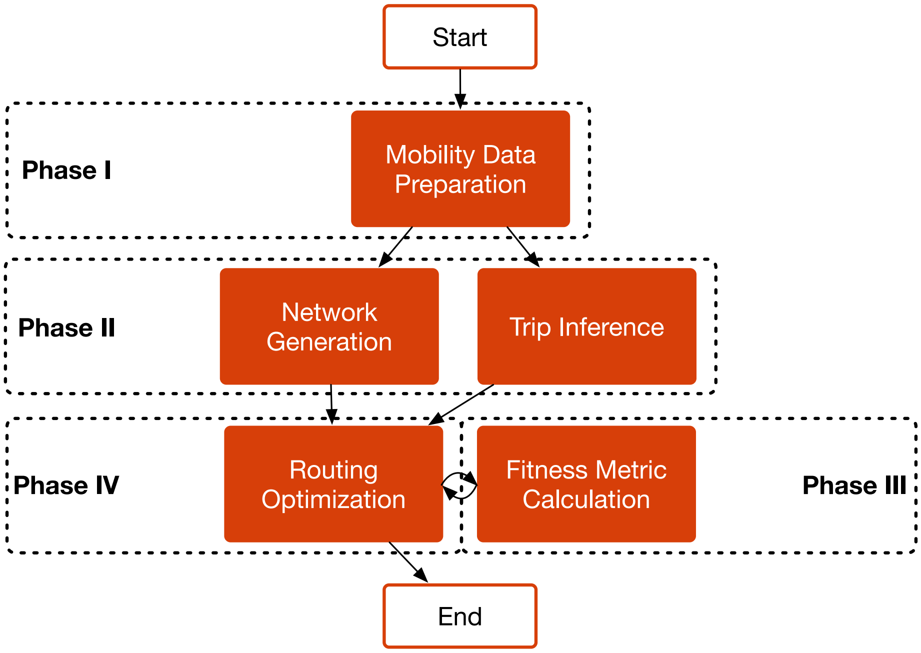


Figure 2.1: Main Phases of the Framework to Improve Rider Social Experience

2.3.1.1 WiFi User Session Data

The WiFi user session data is comprised of 17.6 million records that were logged over 238 days through the WiFi network deployed on the main campus of Oregon State University (OSU). Table 2.1 shows the information contained in those logs that is relevant for this research.

The spatial location of access points (APs) on the OSU campus was set initially based only on the buildings in which these units were located. To establish more precise WiFi user session locations, the AP location data required improvement. To improve AP location data, a mobile device running the WiGLE framework [45] was used to collect the radio MAC address, signal strength, and location data from as many APs as possible. As these AP measurements were collected, the relative GPS location of the mobile device was also recorded. The WiGLE framework uses trilateration to generate approximate locations of the APs using the collected data. A relationship was then established between the known wired MAC addresses of APs, used for WiFi user session logging, and the radio MAC addresses collected with the WiGLE framework. The specific steps followed

Field Name	Data Type	Description
MacPIN	String	Anonymized representation of the MAC address of a personal wireless device (PWD).
SessionStart_Epoch	Unix timestamp	Date and time a PWD initiated a connection with an access point (AP).
SessionEnd_Epoch	Unix timestamp	Date and time a PWD terminated a connection with an AP.
AP_Mac	String	MAC address of AP network back haul.
BuildingCode	String	Building abbreviation where AP is located on campus.

Table 2.1: Relevant fields of the WiFi user session data

to establish more precise AP locations were as follows:

1. Group list of unique, system-internal AP identifier strings from WiFi user session logs.

2. Join list of system-internal AP identifier strings to associated Ethernet (i.e., wired) MAC address.

3. Calculate possible *basic service set identifiers* (BSSIDs) for each AP based on Ethernet MAC address and join to list of system-internal AP identifier strings.

4. Query WiGLE database for trilaterated locations of BSSIDs associated with APs in WiFi user session logs.

    - If more than one BSSID for an AP is found in the WiGLE database, the average of the latitudes and longitudes is calculated as the AP's location.

5. Update WiFi user session logs with WiGLE-derived AP locations.

For APs that were not localized using the WiGLE framework, OpenStreetMap data was used to find the centroids of campus buildings, and these points were used as the approximate location for the APs.

2.3.2 Phase II - Network Generation and Trip Inference

2.3.2.1 Network Generation

Road network data was generated from OpenStreetMap data using the OSMnx package [46]. A spatial envelope slightly larger than the area that contained the locations of WiFi user sessions on the OSU campus was used to limit the road network's size. The Beaver Bus provides fixed-route transit service internal to the OSU campus. To best represent the roads available to the Beaver Bus service, the OSMnx network designation of 'drive_service' was used, providing all drivable streets as edges with all intersections as nodes. An overlay of the drivable streets network for the OSU campus is shown in Figure 2.2.

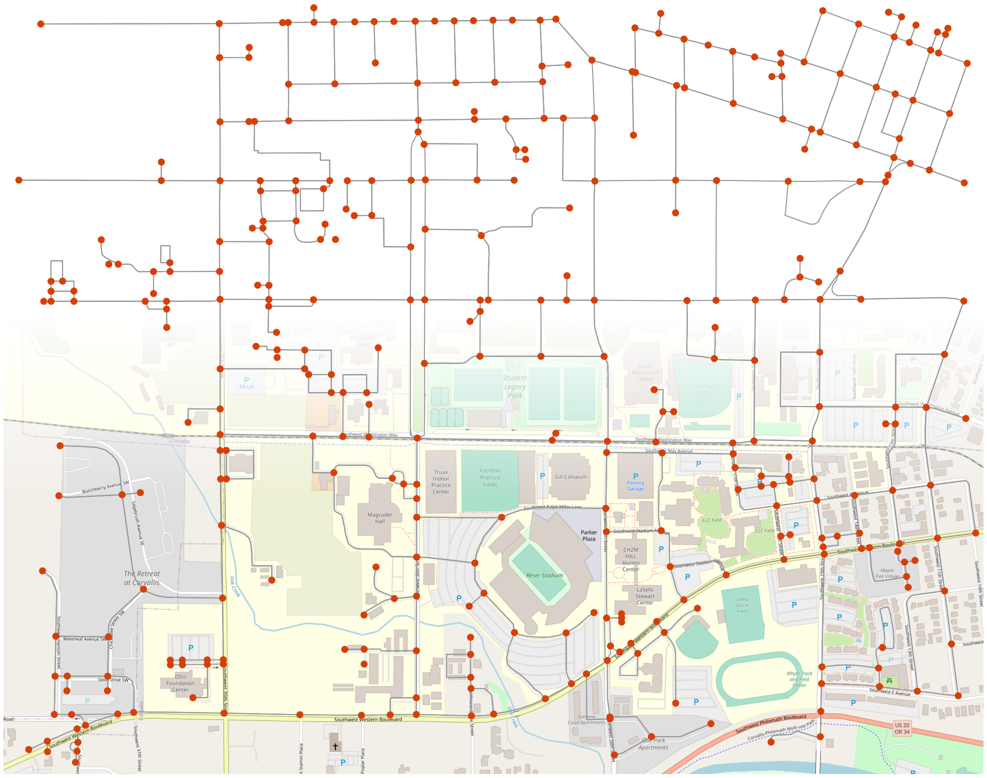


Figure 2.2: Drivable Network on OSU Campus

2.3.2.2 Trip Inference

Trip inference occurs by identifying when PWDs end WiFi user sessions in one building and start WiFi user sessions in another. The following step-by-step process was used to infer trips on the OSU campus using the WiFi user session data:

Step 1 - Remove WiFi user sessions shorter than assigned minimum duration. A minimum session length of 10 minutes was used to prevent the inclusion of WiFi user sessions where a PWD is just passing by a building or inside an adjacent building, which may cause a WiFi user session to occur without that building being occupied by the PWD. Two properties of the data are considered when executing this step:

- APs in the WiFi user session dataset demonstrate a 10-minute polling interval with controllers that log sessions. Not all WiFi user sessions are logged as starting and ending at these intervals, but spikes every 10-minutes of WiFi user session events are attributed to this polling behavior.
- PWDs roam between APs within a building, and occasionally connect briefly to APs in neighboring buildings. As trips within a building are not of interest to transit route planning, those are not considered. For brief trips to neighboring buildings, regardless of whether they actually occur or are merely an inter-building link, the constraint that any bus trip takes less time than walking controls their impact.

The WiFi user sessions that remained following this filtering were considered as origins/destinations of the PWDs.

Step 2.1 - Partition WiFi user sessions by PWD identifier and date. Partitioning isolates WiFi user sessions that are relevant to each other when inferring trips. Partitioning by PWD identifier occurs because, for trip inference, no inter-PWD relationships are considered. Partitioning by date occurs because as transit is not offered overnight, no trip is generated from the end of the last WiFi user session of one calendar date to the start of the first WiFi user session on the next calendar date (which is determined by timestamps associated with a WiFi user session).

Step 2.2 - Sort within WiFi user session partitions by time. To identify the ordering of origins/destinations, WiFi user session partitions are sorted in ascending order by time of day. Building occupancy is determined by consolidating temporally-adjacent WiFi user sessions within the same building and marking arrival time when the first WiFi user session starts and departure time when the last WiFi user session ends.

Step 3 - Define trips with rolling origin/destination designation over a sorted WiFi users session partition. Trips can now be generated from within the sorted WiFi user session partitions by defining trips with a rolling two-session window using the earlier WiFi user session as the origin and the later WiFi user session as the destination.

Step 4 - Concatenate WiFi user session partitions to produce trips dataset. The WiFi user session partitions are concatenated to create a trips dataset for all trips inferred from the original WiFi user sessions.

Step 5 - Separate weekday and weekend trips. To facilitate different weekday and weekend route heuristic optimization solutions, weekday trips and weekend trips are separated.

From the 17.6 million WiFi user sessions that were logged, 1.7 million intra-campus trips were inferred.

2.3.3 Phase III - Objective Metric Calculation

Measuring social benefit is a challenging endeavor. One approach is taking a direct measure of social satisfaction [47]. However, it may end up being highly subjective, dependent on active participation, and unlikely to see a level of adoption that allows scaling beyond a short study.

Inferring a measure of social benefit from spatiotemporal data has been approached in numerous ways. For example, Liang, Li, and Zhang [36] assigned encounters to one of four types (i.e., familiar stranger, in-role, friend, and stranger). In another study, Asatani, Toriumi, Mori, *et al.* [39] inferred that repeated spatiotemporal co-occurrences were more due to interpersonal relationships rather than familiar strangers, and Sun, Axhausen, Lee, *et al.* [38] highlighted the value of identifying "super-spreaders" for communicable diseases as their movements heavily influence the transmission thru social interaction.

It is important to define a concise social benefit metric to be used as the objective for route improvement. To the best of our knowledge, no existing metric clearly yields a measure of social benefit in public transit. To address this need, two suitable measures of social benefit applicable to public transit were developed in this research: the *Familiar Stranger* (FS) metric and the *Encounter Network Clustering* (ENC) metric.

The FS and ENC metrics result from translating identified social benefit metrics to apply to transit. Both metrics must evaluate every trip against every potentially overlapping trip with respect to time and place. Time overlaps are deemed possible to occur either while waiting at the origin bus stop for a trip or while riding in the same transit vehicle during a trip, as illustrated by the examples in Figure 2.3. In the first example, the time overlap occurs while waiting for a bus at Stop C and accounts for 10% of the total time of Trip 1 and 20% of the total time for Trip 2. The second example depicts a scenario where the overlap is due to riding on the same bus on Route A and accounts for 40% of the total time of Trip 3 and 20% of the total time for Trip 4.

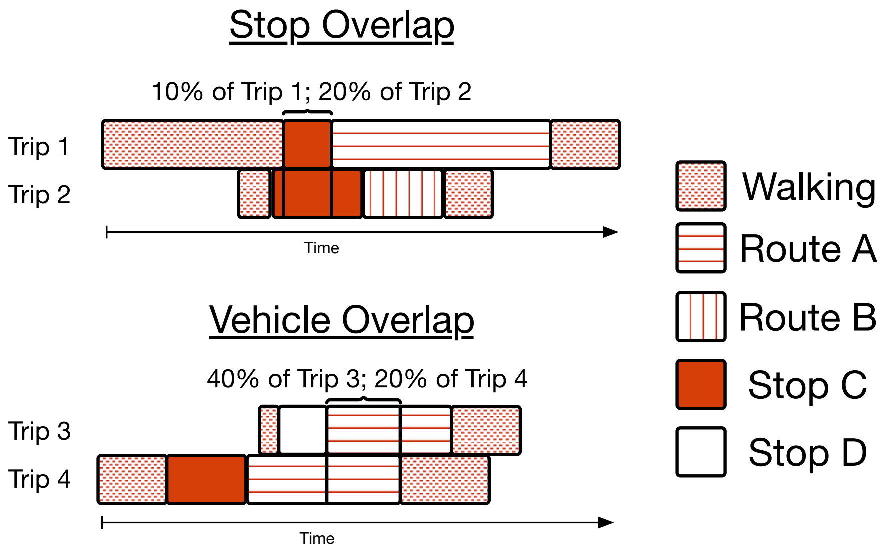


Figure 2.3: Trip Overlap Timelines

To allow this analysis to scale, the Python library Dask was used to enable distributed

Notation	**Meaning**
I	Set of individual trips
s	Time duration of trip
b	Time boarding at (or departing from) a transit stop
a	Time alighting at (or arriving to) a transit stop

Table 2.2: Metric Calculation Notation

parallel processing [48]. The Dask distributed architecture uses a scheduler/worker paradigm with tasks submitted to the scheduler to distribute and coordinate computation by workers. Partitioning metric evaluations into trip partitions that share a date plus a vehicle or a stop facilitate the distribution of metric computations amongst worker nodes. By distributing computations amongst a cluster of worker nodes whose computational capabilities exceed that of a single node, large quantities of computational resources are harnessed to increase the speed at which the FS and the ENC metrics are evaluated.

The notation used in establishing the new metrics is presented in Table 2.2.

2.3.3.1 Familiar Stranger Metric

The FS metric holds similarities to the generation of encounter networks for communicable diseases in which the amount of time exposed to an infected individual is important to understanding the likelihood of transmission. In the FS metric, the amount of *overlap* that occurs during a trip is of social interest. FSs in transit are a function of spending time in the proximity of those strangers, and their presence having an impact on the experience.

As mentioned in Section 2.3.3, trips can overlap either at the boarding bus stop or while traveling in a vehicle. The overlap time, OT, is calculated using Equation 2.1 with boarding/alighting being equivalent to arriving/departing a bus stop.

$$OT = \sum_{i \neq j \in I} \min(a_i, a_j) - \max(b_i, b_j) \tag{2.1}$$

The FS metric is defined as the sum over all trips of the percentage of total time of one trip that overlaps with another, as shown in Equation 2.2.

$$FS = \sum_{i \in I} \frac{OT}{s_i} \tag{2.2}$$

$$\max FS \tag{2.3}$$

In the example timeline in Figure 2.3, Trip 1 would contribute 0.1 to the sum for its stop overlap with Trip 2, and Trip 2 would contribute 0.2 to the sum for its overlap with Trip 1. Equation 2.3 is a possible objective of the heuristic optimization.

2.3.3.2 Encounter Network Clustering Metric

While the FS metric assesses an overarching sum, the ENC metric, C, assesses the coincidence of PWDs over the period of analysis. Influenced by the benefits of small world networks identified by Watts and Strogatz [43], the ENC metric uses a local clustering coefficient, c_v, shown in Equation 2.4.

$$c_v = \begin{cases} 0 & \text{if } deg^{tot}(v) < 2 \\ \frac{1}{2(deg^{tot}(v)(deg^{tot}(v)-1)-2deg^{\leftrightarrow}(v))} \sum_{uw} (\hat{w}_{vu}\hat{w}_{vw}\hat{w}_{uw})^{\frac{1}{3}} & \text{otherwise} \end{cases} \tag{2.4}$$

where,

$$\hat{w}_{vu} = \frac{w_{vu}}{\max(w)} \tag{2.5}$$

and

$$w_{vu} = \sum_{i \neq j, j \in I_v, I_u} \frac{\min(a_i, a_j) - \max(b_i, b_j)}{s_i} \tag{2.6}$$

Then, the objective metric of a solution is evaluated by taking the average of all PWDs, as shown by Equation 2.7.

$$C(G) = \frac{1}{n} \sum_{v \in G} c_v \tag{2.7}$$

Equation 2.8 is another possible objective of the heuristic optimization.

Notation	**Meaning**
G	PWD coincidence network
u, v	n nodes in G
vu	Edge between nodes v and u
w_{vu}	Weight of edge vu
w_{uw}	Weight of subgraph of node u as defined by Onnela, Saramäki, Kertész, *et al.* [49]
$deg^{tot}(v)$	Sum of the in and out degrees of node v
$deg^{\leftrightarrow}(v)$	Reciprocal degree of node v
I_v	Set of trips taken by node v

Table 2.3: PWD Network Notation

$$\max C(G) \tag{2.8}$$

Additional notation for the PWD encounter network is presented in Table 2.3.

Similar to the FS metric, the percentage of a trip that overlaps with another holds benefit. More specifically, the sum of trip overlap percentages is used as the edge weights of the network, as shown by Equation 2.6. As the amount of overlap is not likely symmetrical between two PWDs (i.e., nodes in the encounter network), the network must be directed. The resulting encounter network is thus weighted and directed. The average local clustering coefficient, as defined by Fagiolo [50] and Onnela, Saramäki, Kertész, *et al.* [49], was selected as the objective metric for a couple of reasons.

First, a global clustering coefficient (a measure of transitivity) is not well suited to evaluate a network with a large proportion of zero-degree nodes (e.g., walked trips). Since a global clustering coefficient is a ratio of closed triplets (i.e., three nodes in an undirected graph connected by three edges) to the total number of triplets, a network with a high global clustering coefficient could be composed of just a few high-degree nodes and many zero-degree nodes. With a local clustering coefficient, zero-degree nodes can be assigned a value of zero, and affect a large penalty when the average is taken across all nodes.

Second, the local clustering coefficient as defined by Fagiolo [50] and Onnela, Saramäki, Kertész, *et al.* [49] has been implemented in the NetworkX package [51]. NetworkX is a commonly used network analysis library in Python, but is known to be slow due to its native Python implementation. The package graph-tool [52] was also considered as it is largely implemented in C++ and is capable of quickly processing an average local clustering coefficient (albeit, the slightly different average local clustering coefficient as defined by Watts and Strogatz [43]). For the network size and system on which the average local clustering coefficient was calculated, graph-tool proved about 50x faster than NetworkX. However, graph-tool proved inconsistent when calculating graphs with many zero-degree nodes (for which there are many in this analysis). Ultimately, the stability of NetworkX's average local clustering coefficient is what led to its use for analysis.

2.3.4 Phase IV - Evolutionary Heuristic Optimization

Finding solutions for UTRP-like problems using evolutionary optimization heuristics is an area well explored in recent research [25, 27, 30, 53–56]. While many of the methods reported in prior work are not directly applicable to the UTRP developed in this research, other evolutionary optimization heuristics are a good fit for the unique nature of the social objective. Central to this research effort is the parallelization of the employed heuristics.

2.3.4.1 Parallel Optimization

The island model of parallel optimization facilitates the distribution of an evolutionary optimization amongst a cluster of computational nodes. The foundations for the island model for parallel optimization emerged in the late 1980s [57, 58], and it has been implemented since to allow distributed optimization. Conceptually, distribution is accomplished by establishing populations of solutions as islands and defining rules for evolution on an island and migration between the islands. Beyond the parameter tuning that is part of most optimization heuristics, additional inter-island parameters affect optimization performance.

Ruciński, Izzo, and Biscani [59], which were part of the team at the European Space Agency (ESA) that developed PaGMO, studied the impacts of migration topology when using the island model for parallel optimization. Fourteen different topologies were evaluated in the application of both Differential Evolution and Simulated Annealing. The authors found that migration topology impacts both the quality of solutions and the time to convergence. They also found that the algorithm used for optimization had the

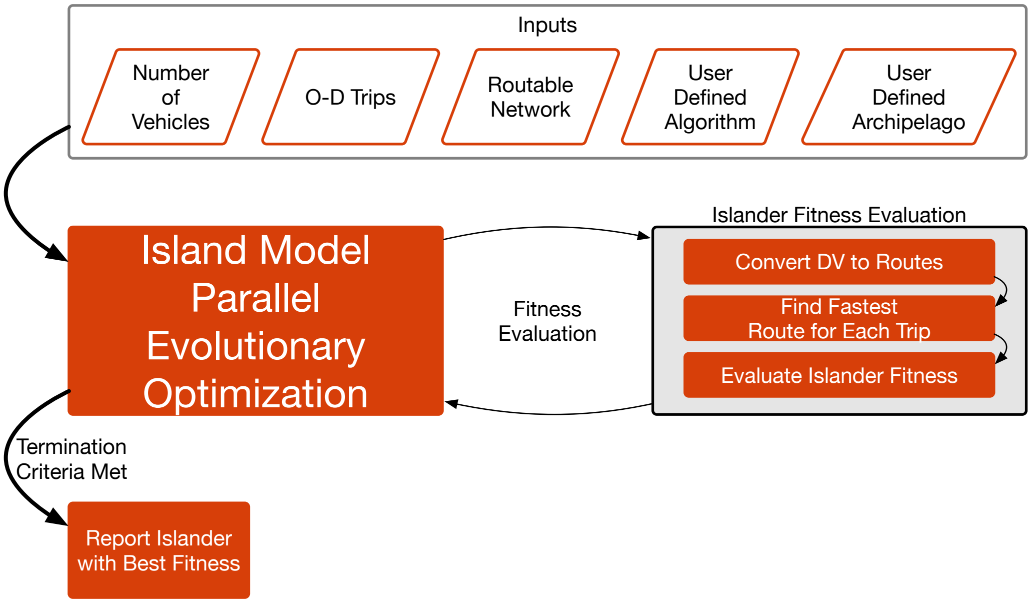

Inputs
Number of Vehicles
O-D Trips
Routable Network
User Defined Algorithm
User Defined Archipelago
Island Model Parallel Evolutionary Optimization
Fitness Evaluation
Islander Fitness Evaluation
Convert DV to Routes
Find Fastest Route for Each Trip
Evaluate Islander Fitness
Termination Criteria Met
Report Islander with Best Fitness

Figure 2.4: Evolutionary Heuristic Optimization Flowchart

biggest impact on the relative performance of the various topologies compared to other modeling aspects such as the number of islands and the problem to be optimized. Given the lack of specific guidance resulting from Ruciński, Izzo, and Biscani's work, several parameters and configurations were evaluated in this research.

2.3.4.2 Heuristic Optimization Implementation

pygmo was the parallel heuristic optimization library used to implement the island model evolutionary heuristic optimization of this framework. pygmo is based on the C++ library PaGMO for parallel optimization, which uses an asynchronous island model [60]. The pygmo library was developed at the ESA to facilitate high-dimension global optimization problems for spacecraft trajectories and part design, yet the universality of the approach and the available algorithms makes it easily employed in other global optimization problems. An overview of the island model as implemented in pygmo and adapted for this framework is shown in Figure 2.4. It is important to note that the islander fitness evaluation module depicted in Figure 2.4 is unique to this framework.

The implementation of the island model evolutionary heuristic optimization begins

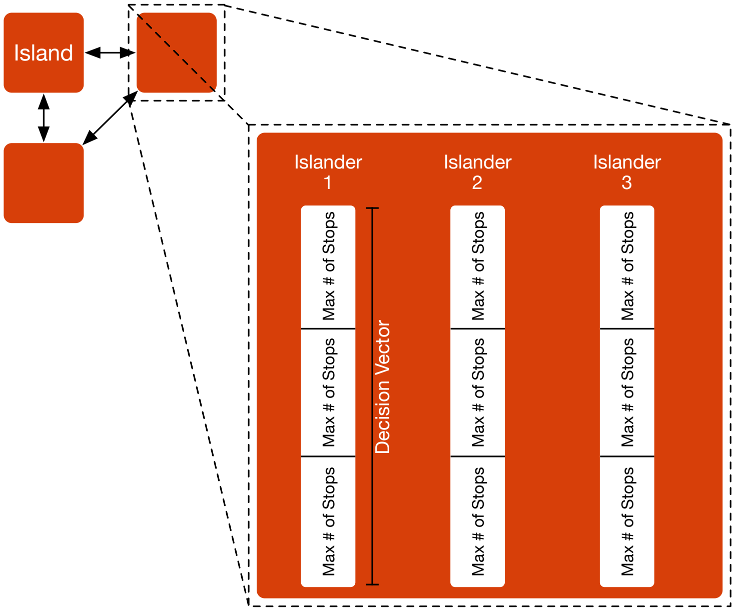


Figure 2.5: Solutions Archipelago

with creating island populations (i.e., islanders) with initial decision vectors. In the context of this research, an initial decision vector consists of a decision vector with a length equal to the product of the maximum permissible number of stops per route and the number of routes. Figure 2.5 illustrates an example of three initial decision vectors. Every value in the decision vector references a node index in the transportation network. Every node in the transportation network is a stop node, and stop nodes are assigned index values which increase southwest to northeast in the transportation network, as shown in Figure 2.6. The decision vector's initial values are randomly assigned as either an integer representing a stop index value or a null value representing no stop.

Once the initial decision vectors are generated, the evolutionary process can commence. Every generation, each decision vector is evaluated by calculating the objective metric with the two network (i.e., transit and encounter) interactions outlined in Figure 2.7 beginning in the top-left where a mapped representation of routes defined by a decision vector is shown. The first step in calculating the objective metric is to take the routes defined in the islander's decision vector and assigning all the trips in the dataset to either a route or to walk. Trips are assigned to either the route that takes the least amount of

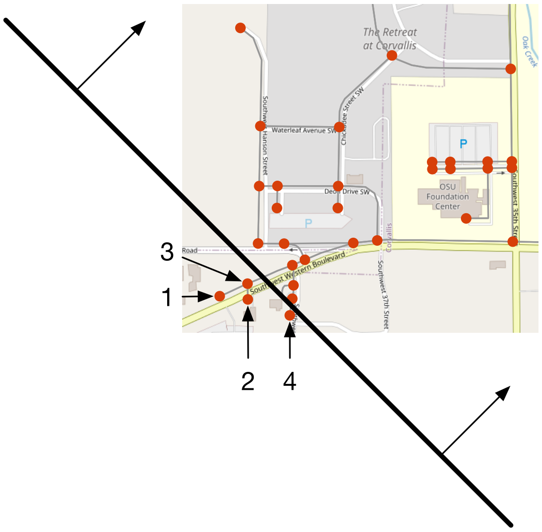


Figure 2.6: Assigning Indices to Transportation Network Nodes

time or exclusively walking. Trip assignments that involve a transfer between routes are not considered. An example is shown in Figure 2.8 where the trip would be assigned to use Route A. If Route A were not an option, the trip would walk rather than taking the longer Route B. The process of assigning trips to the fastest option occurs for every trip in the dataset.

Once all trips are assigned to either a route or exclusively walking, the FS or the ENC objective metric is calculated as described in Section 2.3.3.

Table 2.4 shows the notation used to calculate the maximum occupancy of every route. This constraint supports there being a space on a vehicle for riders for whom the vehicle's route is the fastest. The maximum occupancy of every route is determined via a cumulative sum of boardings (each boarding represented by a +1) and alightings (each alighting represented by a -1) for every stop event, as shown in Equation 2.9. The maximum cumulative sum each route experiences is determined, and the resulting objective value is either the metric calculated or zero (if it is found that a maximum occupancy is exceeded), as shown in Equation 2.10.

$$O_r^{(0)} = 0, \quad O_r^{(k)} = O_r^{(k-1)} + b_k - a_k \tag{2.9}$$

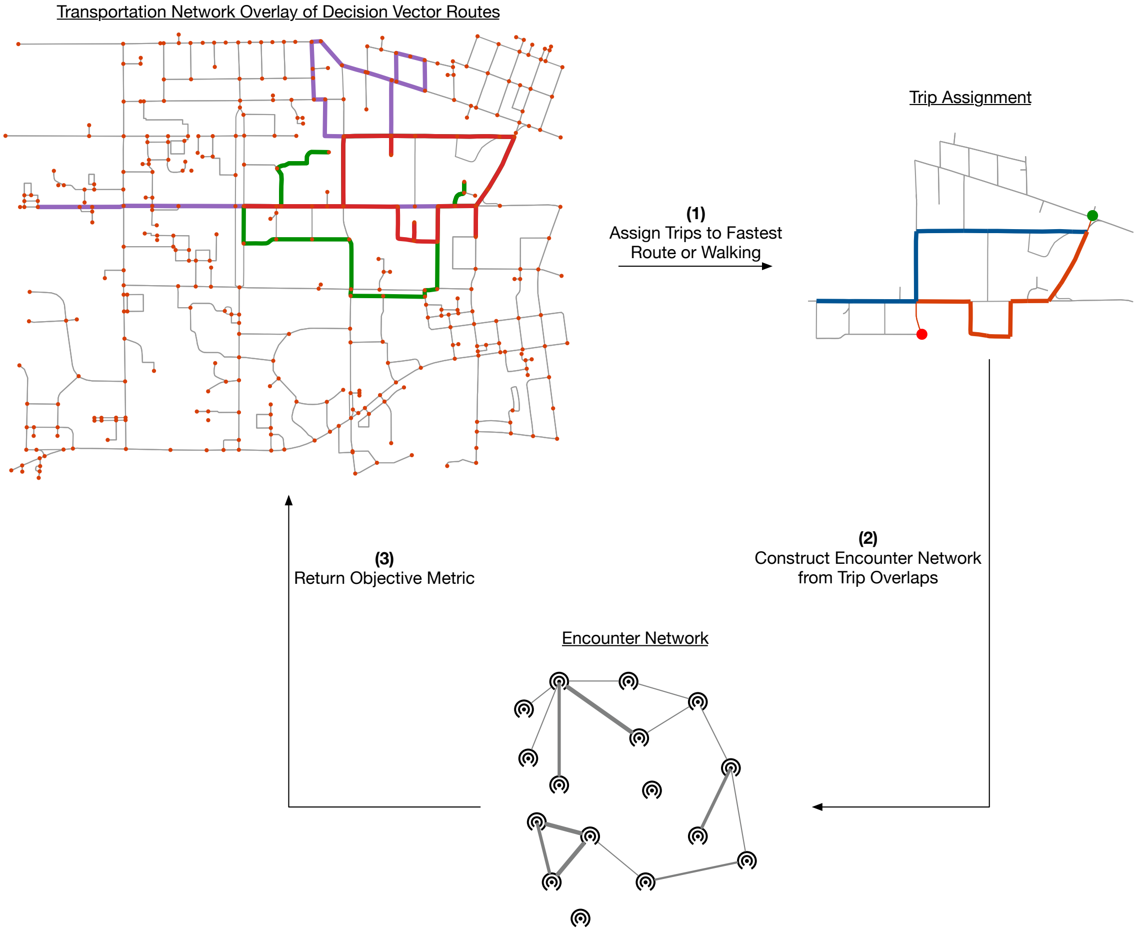

Transportation Network Overlay of Decision Vector Routes
Trip Assignment
(1)
Assign Trips to Fastest Route or Walking
(2)
Construct Encounter Network from Trip Overlaps
Encounter Network
(3)
Return Objective Metric

Figure 2.7: Two Network Interaction

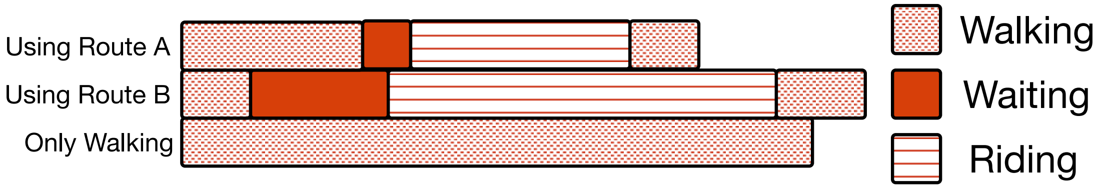


Figure 2.8: Assessing Trip Timelines

Notation	Meaning
R	Set of all routes
r	Single route in R
k	Stop occurrence
r_k	Set of all stop occurrences on route r
$O_r^{(k)}$	Occupancy at stop occurrence k on route r
O_{max}	The maximum occupancy of a vehicle
b_k	Boardings at stop occurrence k
a_k	Alightings at stop occurrence k

Table 2.4: Maximum Occupancy Notation

$$Objective = \begin{cases} 0 & \text{if } \max O_r^{(k)},\ \forall\ r \in R \wedge\ \forall\ k \in r_k > O_{max} \\ FS \vee C(G) & \text{otherwise} \end{cases} \tag{2.10}$$

Island populations evolve for the designated number of generations before a migration (as defined by selection, replacement, and topology) of islanders between the islands occurs. If the best islander objective values in the archipelago do not improve for a designated number of migrations, then the heuristic optimization is assumed to have converged.

2.3.4.3 Evolutionary Algorithm Selection

Within each island of the island model, an algorithm is employed to implement the evolutions. Three candidate algorithms were identified: Artificial Bee Colony (ABC) as defined by Mernik, Liu, Karaboga, *et al.* [61], Improved Harmony Search (IHS) as defined by Mahdavi, Fesanghary, and Damangir [62] with some pygmo-specific adaptions, and Simple Genetic Algorithm (SGA) devised by the pygmo team. These three candidate algorithms were selected because they are all appropriate for a single-objective, unconstrained optimization and are suitable for use with an integer-encoded decision vector (as stops are indexed using integer values, non-integer values would require additional processing). Trials using 10,000 chronologically contiguous trips, the FS metric, and varying some algorithm parameters showed that initial results were best using SGA. The

Operator	**Type**	**Description**
Selection	Tournament	Supports maintaining genetic diversity amongst an island's population while promoting the likelihood of well-performing route segments appearing in the next generation.
Crossover	Single Point	Facilitates passing on well-performing route segments of parents onto children.
Mutation	Gaussian	Stop index values close to one another also represent geographic proximity. Gaussian mutation results in gradual geographic route changes. The gradual changes support convergence in comparison to a distribution that would cause more disruptive mutations, like a uniform distribution mutation strategy.
Reinsertion	Pure Elitism	Pure elitism is the only reinsertion approach available for SGA within the pygmo library.

Table 2.5: Genetic operators of the SGA used in pygmo

inputs and results of these trials can be seen in Tables 2.10, 2.11, and 2.12.

2.3.4.4 Simple Genetic Algorithm

The implementation of the SGA available in pygmo was used to drive the evolution of each generation of the island model. For this research, the decision vector used in the island model heuristic optimization process becomes analogous to a chromosome in the SGA. The SGA in pygmo can execute different genetic schemes, including selection, crossover, mutation, evaluation, and reinsertion [63]. At every stage (except evaluation), the choice of a specific genetic operator within a scheme can affect the performance of the SGA [59], thus making this choice not trivial. Understanding of the problem and its chromosome representation can guide the choice of specific genetic operators. Table 2.5 shows the specific genetic operators that were preemptively chosen in this research to limit unnecessary breadth of search.

The use of SGA as an evolutionary algorithm and the choice of socially-aware objectives support the timely convergence of feasible solutions to a well-performing route plan solution.

Parameter	**Level**	**Comments**
Number of Archipelago Islands	10	Practically constrained by available processor cores
Number of Islanders per Island	50	Each islander represents its solution with a decision vector
Number of Generations per Evolution	4	Obtained by rounding up the natural logarithm of the number of islanders
Number of Vehicles	5	Parallels the number of existing routes on OSU's Beaver Bus service
Maximum Number of Stops per Vehicle	5	
Stopping Criteria	2	Number of evolutions without improvement of best solution metric value. Only two (2) evolutions were tested.

Table 2.6: Set of Fixed Experimental Parameters

2.4 Results and Discussion

The performance of the social experience heuristic optimization framework was assessed through a series of computational experiments. Since preliminary computational experiments exhibited long run times to reach the defined convergence criterion, additional computational experiments were conducted using a two-stage approach to explore the performance (i.e., fine-tuning) of the parameters of the island model SGA and the UTRP while still maintaining implementable processing times.

In the first stage of the two-stage approach, two sets of experimental parameters were used. The first set of six experimental parameters, shown in Table 2.6, used a single fixed level to better balance the long computational runs observed in the preliminary experiments. The second set of three experimental parameters, shown in Table 2.7, used two levels for each parameter. The levels for each variable experimental parameter were chosen based on the default SGA parameter implemented in pygmo [60] and an additional relaxed value that would likely increase variability between generations. In the first stage,

Parameter	**Levels**	**Comments**
Archipelago Topology	• Fully Connected • Ring	Frequently implemented connected topologies for the Island Model; the only connected topologies predefined by Biscani and Izzo [60]
SGA Crossover Probability	• 0.90 • 0.75	SGA default value [60] and relaxed value
SGA Mutation Probability	• 0.02 • 0.10	SGA default value [60] and relaxed value

Table 2.7: Set of Variable Experimental Parameters

the first 10,000 trips chronologically were used as input data to examine the impact of the variable parameters. The first inferred trip in the 10,000-trip dataset started at 12:03 am on Monday, 18 September 2017, and the final inferred trip concluded at 1:32 pm on Wednesday, 20 September 2017.

The computational results from stage one were used to inform stage two, in which the inferred trips of the busiest five consecutive weekdays present in the data were used as input. Using the five consecutive weekdays with the largest number of inferred trips allowed route analysis when it was expected that the transit service would be experiencing its highest demand. The busiest five consecutive weekdays in the dataset were Monday, 9 October 2017, to Friday, 13 October 2017. On Saturday, 14 October 2017, there was a home football game, which may explain an increased level of campus activity on the days leading up to the game. The week is also roughly the middle of the Fall 2017 term, with mid-term preparations potentially contributing to campus activity. During the busiest five consecutive weekdays, 79,879 trips were inferred.

2.4.1 Familiar Stranger Metric Results

As described in Section 2.3.3.1, the FS metric assesses the proportion of a trip spent in proximity of another transit user. This metric is cumulative and is indifferent to any prior interactions.

2.4.1.1 Stage One - Parameter Tuning

A 2^3 full factorial designed experiment was conducted to better understand the effect of the parameters archipelago topology, crossover probability, and mutation probability on the FS metric. The first 10,000 chronological trips were used as input data in these experiments.

Each experimental parameter had a low and a high level, as shown in Table 2.7, which resulted in eight treatment combinations. Two replications were performed per treatment combination using a different pseudo-randomly generated initial solution for a total of 16 computational runs. Each replication was run until the stopping criterion of two evolutions without improvement of the best solution metric value was met. The average time for the stopping criterion to be met was approximately 5.5 hours per treatment combination replication. The resulting value of the FS metric for each replication is shown in Table 2.13.

Figure 2.9 shows the results from every replication separated by parameter level. Each plot in Figure 2.9 shows the resulting metric values for the four treatment combinations, each with two replications, that used the level indicated. The replication with the single highest value for the FS metric (i.e., 3076.93) used a fully connected archipelago topology, a 0.9 crossover probability, and a 0.02 mutation probability.

2.4.1.2 Stage Two - Busiest Week Analysis

In this stage, the parameter values for a fully connected archipelago topology, a 0.9 crossover probability, and a 0.02 mutation probability along with input data in the form of 79,879 trips from five contiguous workdays were used for the route improvement process. Due to time constraints, only a single repetition was performed, which met the stopping criterion after approximately 19 hours. As shown in Figure 2.10, the improved routes deliver a 242% improvement in the FS metric over the existing baseline routes. This improvement is accompanied by a 45% increase in the number of trips served, a measure that was not an objective. As shown spatially in Figure 2.11, the improved routes heavily concentrate along the perimeter of two centerpieces of the campus life (i.e., the library and the student activity center) instead of near the physical center of campus.

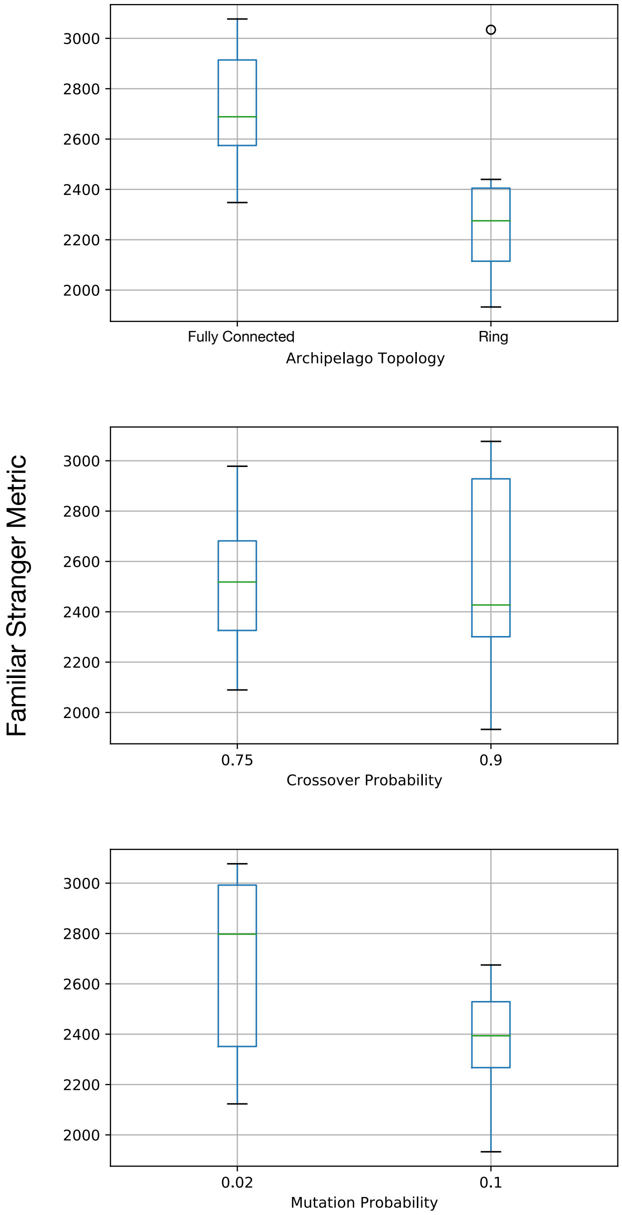


Figure 2.9: Familiar Stranger Trials

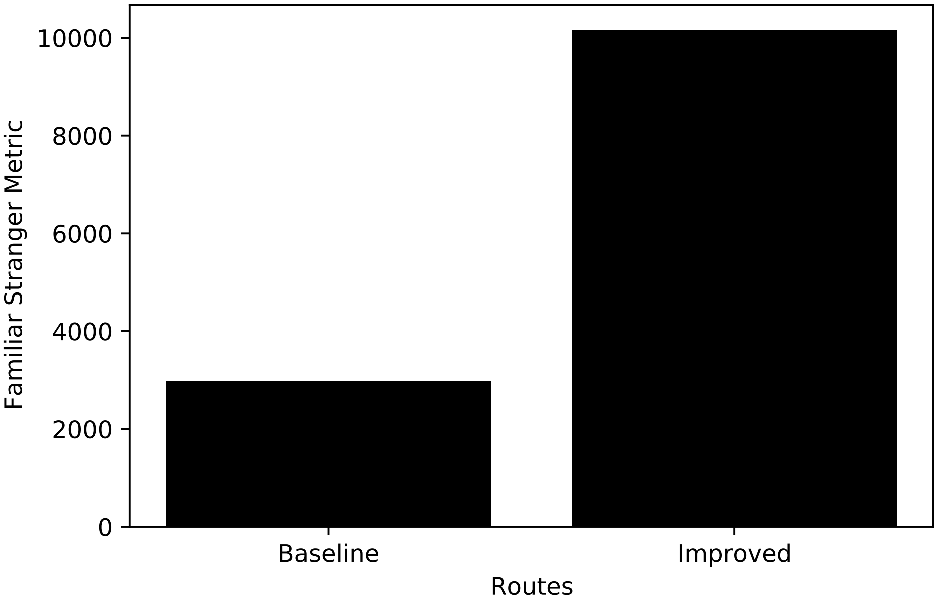


Figure 2.10: Familiar Stranger Metric Comparison

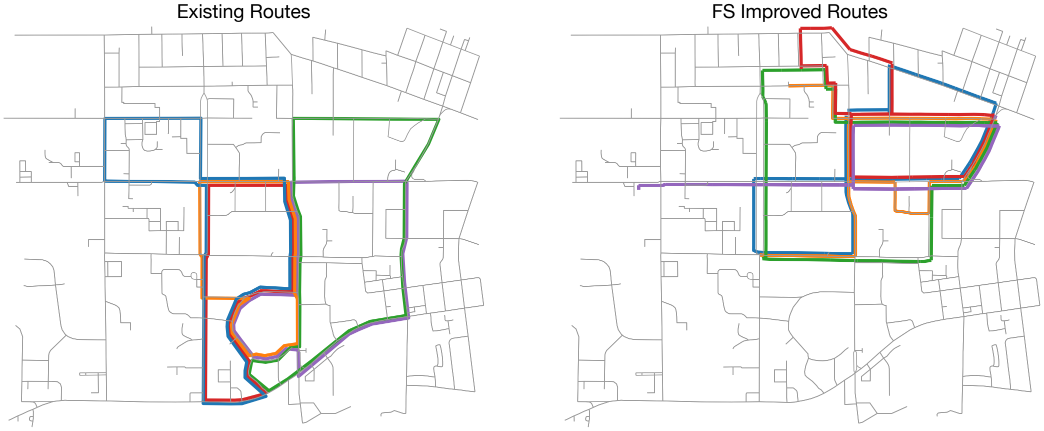


Figure 2.11: Improved Familiar Stranger Routes

2.4.2 Encounter Network Clustering Metric

As described in Section 2.3.3.2, the ENC metric assesses how the encounter network potentially serves the development of socially beneficial networks. The ENC metric accounts for prior interactions, and the performance of trials demonstrates the impacts of increasing the analysis period from the 10,000 trip sample to the 79,879 trip sample.

2.4.2.1 Stage One - Parameter Tuning

A 2^3 full factorial designed experiment was also conducted to better understand the effect of the parameters archipelago topology, crossover probability, and mutation probability on the ENC metric. Similarly, the first 10,000 chronological trips were used as input data in these experiments.

Each experimental parameter had a low and a high level, as shown in Table 2.7, which resulted in eight treatment combinations. Two replications were performed per treatment combination using a different pseudo-randomly generated initial solution for a total of 16 computational runs. Each replication was run until the stopping criterion of two evolutions without improvement of the best solution metric value was met. The average time for the stopping criterion to be met was approximately 10.2 hours per treatment combination replication. The resulting value of the ENC metric for each replication is shown in Table 2.14.

Figure 2.12 shows the results from every replication separated by parameter level and is presented in the same manner as Figure 2.9. The replication with the single highest value for the ENC metric (i.e., 0.05931) used a ring archipelago topology, a 0.75 crossover probability, and a 0.02 mutation probability.

2.4.2.2 Stage Two - Busiest Week Analysis

In this stage, the parameter values for a ring archipelago topology, a 0.75 crossover probability, and a 0.02 mutation probability along with input data in the form of 79,879 trips from five contiguous workdays were used for the route improvement process. Due to time constraints, only a single repetition was performed, which met the stopping criterion after approximately 19 hours. As shown in Figure 2.13, the improved routes translate into a 119% improvement in the ENC metric over the existing baseline routes. This

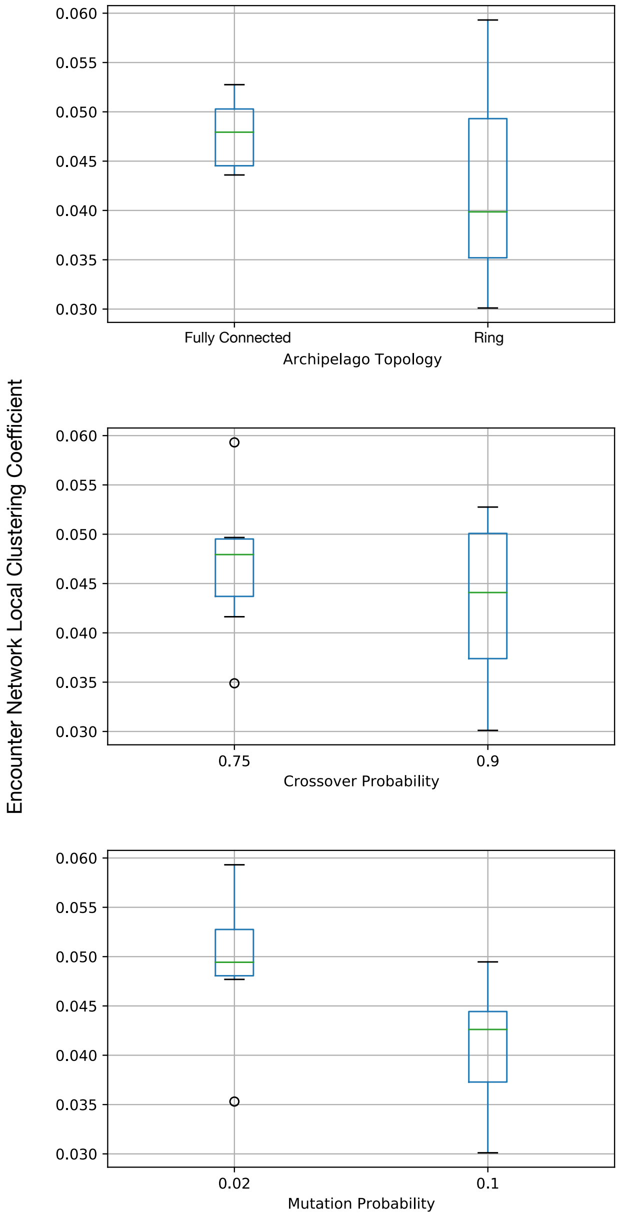


Figure 2.12: Encounter Network Average Local Clustering Coefficient Trials

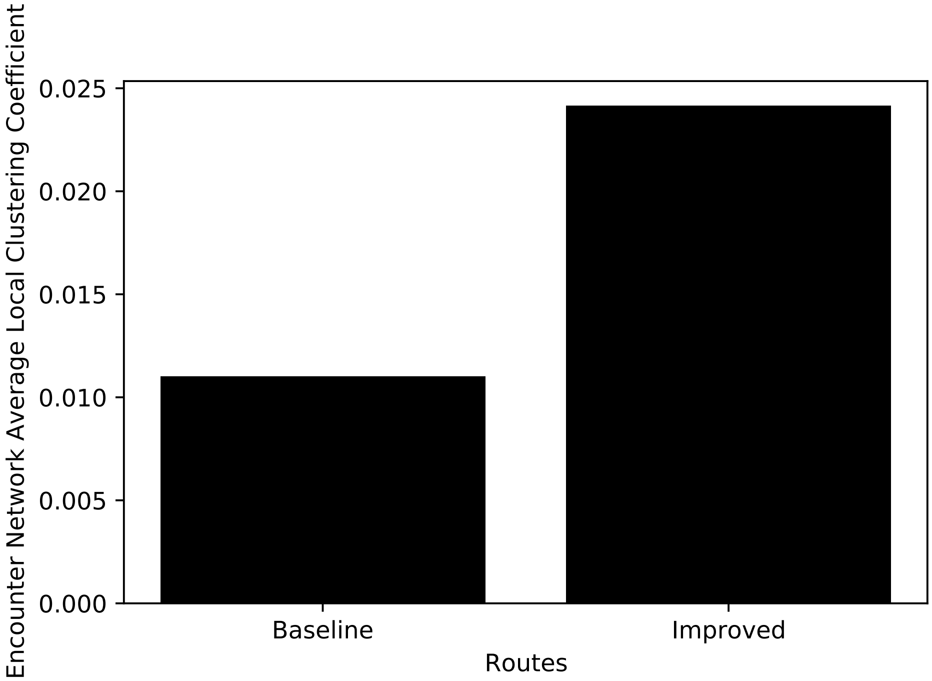


Figure 2.13: Encounter Network Average Local Clustering Coefficient Metric Comparison

improvement is accompanied by a 22% increase in the number of trips served, a measure that was not an objective. Similar to what is seen in Figure 2.11, though to a lesser degree, Figure 2.14 is centered along the perimeter of two centerpieces of the campus life (i.e., the library and the student activity center).

2.4.3 Ridership Comparison

A simple approach to increase the FS metric would be to increase the number of riders. Increasing the number of riders likely increases the number of interactions among these riders and, as a consequence, would drive an increase in the FS metric. To better understand if the improvement seen in the FS metric is merely a function of increased ridership or a reconfiguration of ridership patterns, it is reasonable to examine the relationship between the FS metric improvement and an increase in ridership. The FS metric's improvement is 5.4 times that of ridership, suggesting that more than just an

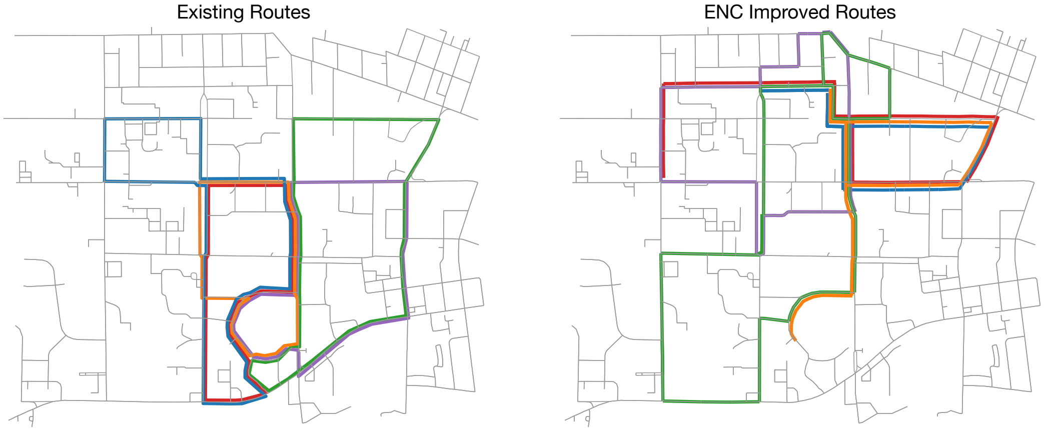


Figure 2.14: Improved Encounter Network Average Local Clustering Coefficient Routes

	FS Improved	ENC Improved
Metric	242%	119%
Trips Served	45%	22%
Improvement Ratio	5.4	5.4

Table 2.8: Relative Measure Improvement

increased ridership is the cause of the FS metric's gain.

A likely better indicator is the relative improvement in the ENC metric. Assuming a random network generation of the baseline routes, adding additional riders at random to the network would not affect the ENC metric. Seeing a 5.4x improvement in the ENC metric compared to ridership is a strong indication that the encounter network generated from the improved routes is not merely a larger network with more riders, but a network that better delivers the ENC objective.

2.4.4 Encounter Network Evaluation

Social contact networks are often scale-free networks [64, 65]. The node degree distribution of scale-free networks follows a power law, where the probability P of node degree k is approximated by a distribution of the form $P(k) \sim k^{-\gamma}$. Visually, the degree distributions of the encounter networks for baseline routes, FS improved routes, and ENC improved routes all fit a power law distribution, as seen in Figure 2.15. However, Broido

	Power law	**Exponential**	**Log-normal**	**Weibull**
Baseline	*1.00*	1.02	1.04	1.05
FS Improved	1.00	0.88	0.90	*0.84*
ENC Improved	1.00	1.04	1.07	*0.99*

Table 2.9: Scaled RSS Distribution Comparison

and Clauset [66] found that "[s]cale-free networks are rare" in empirically observed networks. The authors found that few node degree distributions of the observed networks fit the power law, as would be found in a scale-free network, better than the three evaluated alternative distributions: exponential, log-normal, and Weibull. Evaluating using the residual sum of squares (RSS), the node distributions of the baseline, FS improved, and ENC improved encounter networks were fit to the four distributions. The RSS values shown in Table 2.9 have been linearly scaled such that the power law results, considered the baseline for comparison, are 1.00. The best performing distribution for each encounter network, determined by the lowest RSS, is italicized. Similar to Broido and Clauset's results, a power law distribution was not consistently the best fit. However, with only evaluating three networks, compared to the 3,662 that Broido and Clauset evaluated, the sample size is significantly smaller. The authors' observation that "social networks are at best only weakly scale free, and even in cases where the power-law distribution is plausible, non-scale-free distributions are often a better description of the data" is in keeping with the results of fitting the four distributions to the encounter networks, and while not conclusive as to the encounter networks mimicking observed social networks, demonstrate another shared property.

2.5 Conclusions

This research introduced two new socially-aware objectives that can be used with the demonstrated socially-optimized route framework to produce routes that serve to improve the social experience of riders. Using empirical WiFi user session data to infer trips on a university campus showed how the proposed framework could be applied. The effectiveness of the framework was demonstrated with a 242% and a 119% improvement in the FS and ENC metrics, respectively, when evaluated using the busiest week of the university campus's term.

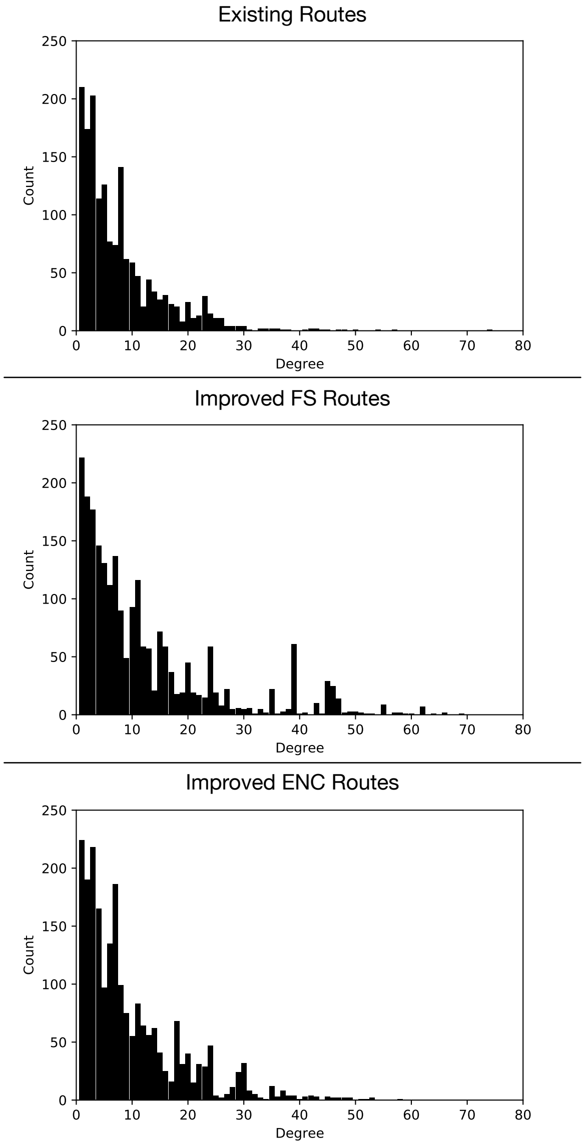


Figure 2.15: Routes' Encounter Network Degree Distributions

As earlier introduced, the improvement of these metrics is aimed to enhance the rider experience with no additional capital resources required from a transit agency. Especially in these times, these organizations understand that what distributes more resources to one service, takes away from another. Metrics like FS and ENC are but two of many metrics that a transit agency will need to consider when planning routes. Objectives like socioeconomic equality and serving previously under-served groups are two additional socially-aware metrics likely to be considered by transit planners. The demonstrated framework is a starting point to introduce additional socially-aware objectives in solutions improved using UTRP methods.

Observed in the university campus demonstration were numerous secondary effects that were not original objectives of the framework. Two potentially significant are an increase in ridership and a redistribution of the most heavily served areas. Public transit agencies see these impacts as they work to balance serving as many users as possible with useful public transit while not neglecting low volume areas that may have more vulnerable populations. This research implements a single-objective that seeks to maximize a single social metric, but future work, intended for real-world implementation, should include additional objectives so that few secondary effects are merely coincidental.

While employing only a single-objective in this framework, nothing is constraining using the two new socially-aware objectives in a multi-objective framework accounting for other transit objectives. The SGA demonstrated as effective in this research has multi-objective genetic algorithm corollaries that could serve the two new socially-aware objectives alongside other transit objectives. By integrating multiple transit objectives, entities tasked with improving transit could develop holistic system planning frameworks that plan transit systems to serve riders better.

2.6 Framework Output

Trial	Population	Max Stops	Vehicle Count	Topology	Stop Generations	ABC Limit	Best Metric Value	Bus Ride Count
0	50	5	5	fully_connected	2	1	-445.5087646	306
1	50	5	5	fully_connected	2	2	-445.5087646	306
2	50	5	5	ring	2	1	-445.5087646	306
3	50	5	5	ring	2	2	-445.5087646	306
4	50	5	5	fully_connected	2	1	-407.3311206	212
5	50	5	5	fully_connected	2	2	-256.4902686	274
6	50	5	5	ring	2	1	-375.4041644	288
7	50	5	5	ring	2	2	-950.8992567	371
8	50	5	5	fully_connected	2	5	-368.9159632	155
9	50	5	5	fully_connected	2	10	-216.4341208	210
10	50	5	5	ring	2	5	-262.7700535	171
11	50	5	5	ring	2	10	-460.5184911	219

Table 2.10: Evolutionary Algorithm Selection - ABC Output

Trial	**Population**	**Max Stops**	**Vehicle Count**	**Topology**	**Stop Generations**	**Best Metric Value**	**Bus Ride Count**
0	50	5	5	fully__connected	2	-445.5110799	168
1	50	5	5	ring	2	-473.1548545	173
2	50	5	5	fully__connected	3	-688.7310816	339
3	50	5	5	ring	3	-533.1894145	260
4	50	5	5	fully__connected	4	-1144.695965	400
5	50	5	5	ring	4	-739.3869472	276

Table 2.11: Evolutionary Algorithm Selection - IHS Output

Trial	**Population**	**Max Stops**	**Vehicle Count**	**Topology**	**Stop Generations**	**SGA Crossover**	**SGA Mutation**	**Best Metric Value**	**Bus Ride Count**
0	50	5	5	fully_connected	2	0.9	0.02	-3076.931677	977
1	50	5	5	fully_connected	2	0.9	0.1	-2506.317444	1005
2	50	5	5	fully_connected	2	0.75	0.02	-2978.279272	1065
3	50	5	5	fully_connected	2	0.75	0.1	-2674.905753	826
4	50	5	5	ring	2	0.9	0.02	-2224.281375	864
5	50	5	5	ring	2	0.9	0.1	-2326.358105	922
6	50	5	5	ring	2	0.75	0.02	-2393.421767	970
7	50	5	5	ring	2	0.75	0.1	-2439.673026	892

Table 2.12: Evolutionary Algorithm Selection - SGA Output

Trial	**Population**	**Max Stops**	**Vehicle Count**	**Topology**	**SGA Crossover**	**SGA Mutation**	**Best Metric Value**	**Bus Ride Count**	**Duration (sec)**
0	50	5	5	fully_connected	0.75	0.02	-2978.279272	1065	38456
1	50	5	5	fully_connected	0.75	0.02	-2702.07246	933	26756
2	50	5	5	fully_connected	0.75	0.1	-2674.905753	826	21598
3	50	5	5	fully_connected	0.75	0.1	-2597.065347	836	13341
4	50	5	5	fully_connected	0.9	0.02	-3076.931677	977	29881
5	50	5	5	fully_connected	0.9	0.02	-2892.810586	727	11937
6	50	5	5	fully_connected	0.9	0.1	-2506.317444	1005	10957
7	50	5	5	fully_connected	0.9	0.1	-2347.752419	784	9816
8	50	5	5	ring	0.75	0.02	-2393.421767	970	39373
9	50	5	5	ring	0.75	0.02	-2123.08255	823	31399
10	50	5	5	ring	0.75	0.1	-2439.673026	892	16577
11	50	5	5	ring	0.75	0.1	-2089.493034	561	11201
12	50	5	5	ring	0.9	0.02	-3034.318787	945	18378
13	50	5	5	ring	0.9	0.02	-2224.281375	864	17626
14	50	5	5	ring	0.9	0.1	-2326.358105	922	11330
15	50	5	5	ring	0.9	0.1	-1932.763573	844	10027

Table 2.13: FS Trials Output

Trial	Population	Max Stops	Vehicle Count	Topology	SGA Crossover	SGA Mutation	Best Metric Value	Bus Ride Count	Duration (sec)
0	50	5	5	fully_connected	0.75	0.02	-0.048182602	1000	63061
1	50	5	5	fully_connected	0.75	0.02	-0.04769152	991	35687
2	50	5	5	fully_connected	0.75	0.1	-0.04946552	970	39999
3	50	5	5	fully_connected	0.75	0.1	-0.044383807	1035	48816
4	50	5	5	fully_connected	0.9	0.02	-0.05275397	1211	56209
5	50	5	5	fully_connected	0.9	0.02	-0.052753875	1145	44429
6	50	5	5	fully_connected	0.9	0.1	-0.044580334	979	24687
7	50	5	5	fully_connected	0.9	0.1	-0.043597588	977	39751
8	50	5	5	ring	0.75	0.02	-0.059311272	1326	61828
9	50	5	5	ring	0.75	0.02	-0.049666946	1043	30077
10	50	5	5	ring	0.75	0.1	-0.041632636	911	19560
11	50	5	5	ring	0.75	0.1	-0.034885781	791	11543
12	50	5	5	ring	0.9	0.02	-0.049192128	998	39806
13	50	5	5	ring	0.9	0.02	-0.035305488	727	35073
14	50	5	5	ring	0.9	0.1	-0.038082986	773	30774
15	50	5	5	ring	0.9	0.1	-0.030109401	679	7230

Table 2.14: ENC Trials Output

Metric	**Population**	**Max Stops**	**Vehicle Count**	**Topology**	**SGA Crossover**	**SGA Mutation**	**Best Metric Value**	**Bus Ride Count**	**Duration (sec)**
fs	50	5	5	fully_connected	0.9	0.02	-10164.69115	3486	69335
enc	50	5	5	ring	0.75	0.02	-0.024140514	2925	68472

Table 2.15: Busy Week Output

STRATEGIC ROUTE PLANNING TO MANAGE TRANSIT'S SUSCEPTIBILITY TO DISEASE TRANSMISSION

Sylvan Hoover, J. David Porter, and Claudio Fuentes

Transportation Research Record: Journal of the Transportation Research Board
Washington, DC
OnlineFirst

Chapter 3: Strategic Route Planning to Manage Transit's Susceptibility to Disease Transmission[1]

3.1 Introduction

Transit agencies have experienced dramatic changes in service and ridership due to the COVID-19 pandemic. Current mitigation actions are primarily operational, but as communities transition to a new normal, strategic measures are needed to support continuing disease suppression efforts. Strategic measures that are to be implemented in the pursuit of disease mitigation need to balance potentially competing interests like efficiency, effectivity, and privacy.

Prior research examined the potential for a better understanding of disease transmission among transit riders. This research expands that analysis to provide actionable results to transit agencies in the form of improved transit routes. A multi-objective heuristic optimization framework employing the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) generates multiple route solutions to planners. The route solutions allow transit agencies to balance the utility of service to riders against the susceptibility of routes to enable the spread of disease. Where other research focuses primarily on individual riders, the route solutions generated by the proposed multi-objective heuristic optimization framework facilitate the disease mitigation goals of a transit agency while not taking actions against any rider on an individual basis. A case study of transit at Oregon State University (OSU) is presented with multiple transit network solutions evaluated and the resulting encounter networks investigated.

[1] S Hoover, JD Porter, and C Fuentes, "Strategic Route Planning to Manage Transit's Susceptibility to Disease Transmission", en, *Transportation Research Record: Journal of the Transportation Research Board*, p. 036 119 812 199 781, Mar. 2021, ISSN: 0361-1981, 2169-4052. DOI: 10.1177/0361198121997815. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0361198121997815 (visited on 05/11/2021)

3.2 Literature Review

This research draws from two distinct and well-explored areas: the Urban Transit Routing Problem (UTRP) and the network modeling of disease transmission. The intersection between these two areas occurs when the individuals served by a UTRP solution represent nodes in a network model of disease transmission. By using metrics derived from the disease transmission network model as an objective in a UTRP, the two representations, one of the transit network and the other of the disease transmission network, evolve to provide a feasible UTRP solution that delivers a network model of disease transmission favorable to mitigating the spread of disease through transit.

3.2.1 Urban Transit Routing Problem

Fundamentally, the UTRP aims at improving transit routes for a defined objective or a set of objectives. However, the UTRP manifests in various forms. In addition to route improvement, for example, some approaches focus on stops, service schedule, and vehicle outputs, among other associated decisions. A UTRP can be either single- or multi-objective, with metrics contributing to the perspective of riders, operators, or both.

In 2008, Guihaire and Hao [25] published a review paper that synthesized prior work addressing the UTRP dating back to 1925. In this work, the authors attribute the first use of evolutionary optimization (e.g., a genetic algorithm) to Xiong and Schneider [26]. Approaches have evolved since Xiong and Schneider's work, but the fundamental aim of improving a transit network for economic efficiency has remained consistent.

Some UTRPs do not even consider route improvements. Mahmoudzadeh and Wang [28] developed a cluster-based scheduling approach to improve the alignment of a university shuttle service with travel behaviors and class times. Using automated passenger counter and automatic vehicle location data from a campus shuttle service, they selected months during which travel patterns appeared consistent (e.g., middle of a term at the university), and used both supervised and unsupervised clustering methods to assess different approaches to schedule departure times to improve efficiency. Mahmoudzadeh and Wang's entire work achieved its goal without a single change to the physical route.

The UTRP can be adapted to model the characteristics of a specific real-world problem. For example, a variant of the UTRP known as the School Bus Routing Problem (SBRP) manifests similar characteristics to the problem addressed with the multi-objective heuristic optimization framework proposed in this research. More specifically, both problems have riders with known origins and destinations, time constraints for when those riders are to be transported, a fixed number of available vehicles, and flexibility in stop location assignment.

3.2.1.1 School Bus Routing Problem

An application of the SBRP that gathered significant coverage in the general media [68–70] involved the bus network of the Boston Public Schools [29]. The authors introduced another variation of the UTRP called bio-objective routing decomposition (BiRD) to model and obtain feasible solutions to the bus time selection problem. Breaking optimization problems into sub-problems to be optimized, recognizing the different issues faced at each stage, is not unique to BiRD.

Lemos, Joshi, D'souza, *et al.* [71] explored the effect of different heuristic optimization methods (i.e., ant colony, honey bee, and greedy randomized optimizations) when finding feasible solutions to the SBRP. The results showed no heuristic optimization approach to be dominant, which suggests that situational interests (e.g., number of buses, distance traveled, etc.) drive which approach is best suited. Although this outcome does not lead to any specific heuristic to be selected for this research, it suggests that multiple heuristics are capable of providing feasible solutions.

In another study, Leksakul, Smutkupt, Jintawiwat, *et al.* [31] used machine learning and evolutionary optimization to plan bus stops and routes for a factory shuttle, which is, effectively, an SBRP. The proposed approach reduced employees' walking distance to bus stops by 79%. Again, breaking a problem of stop identification and route determination into sub-problems to employ appropriate optimization methods (be they heuristic, a machine learning method, integer optimization, etc.) proves an effective approach to solving the greater problem.

The breadth of approaches to find optimal solutions to the UTRP is evident in research performed for various areas of focus within transportation. Rider choice is not always addressed with a UTRP, but Liu, Liu, Yuan, *et al.* [32] employed data from taxi and bus trips in Beijing, China, in a three-phase approach to predict bus route demand and optimize route planning to suit rider demands. Although Liu, Liu, Yuan, *et al.* used origin-destination (O-D) data in a manner that is typical of SBRP research, they recognized that riders have a mode choice and used that recognition to employ an objective that aimed at maximizing the number of riders that choose mass transit.

3.2.2 Network Analysis of Disease Transmission

The disease network needed for this research differs from how disease networks are often constructed. Disease networks that show a temporal quality, where most research is focused, adopt either a susceptible-infectious-recovered (SIR) or susceptible-infectious-susceptible (SIS) network model depending on the disease of focus [72]. SIR and SIS network models have nodes that evolve their state based on their stage in the epidemic. These network models have temporal qualities that are useful when studying how epidemics spread. For this research, rather than investigating how a disease spreads and affects the state of a node, the focus is on minimizing the metrics of a cumulative encounter network so that disease has little opportunity to spread.

Encounter networks exist to represent an interaction between nodes. Some are temporally evolving, like those used in SIR and SIS disease network models. Others are fixed at a time point of interest, like those looking back at a past period of contacts [73]. Knowledge about what constitutes an encounter for the intended use of the encounter network dictates when an edge is to be created between nodes and what attributes those edges may hold. In addressing disease transmission within an encounter network, the results are invariably driven by decisions made about the generation of the encounter network.

The cumulative encounter network does not evolve with time, for it represents all encounters as if they have already occurred. By recognizing how a disease is allowed to spread in an encounter network, transit can be designed to produce a resulting encounter network that minimizes transit's contribution to the spread of disease. The difference between those with interest in a cumulative encounter network and an evolving encounter network is described by Keeling and Eames [72] as: "the research in graph theory and social sciences generally considers an understanding of the network itself to be the ultimate goal; in contrast, epidemiological interest is focused on the spread of the disease, in which case the network forms a constraining background to the transmission dynamics." While the focus of the multi-objective heuristic optimization framework presented in the following sections is on the epidemiological spread, understanding the disease network is more aligned with other fields.

Regardless of the approach, the goal is to reduce the spread of disease. Xu, Connell McCluskey, and Cressman [74] modeled a hub-based transportation system to investigate the effects of mitigating the spread of disease and concluded that reducing the likelihood of infected travelers using transit is not sufficient to stop an outbreak. Instead, reducing an outbreak can be achieved by increasing system efficiency (i.e., reducing contact between travelers) or reducing overall travel by the population. While approaches that target individuals may serve to limit the spread of disease effectively [65, 75, 76], the implications from both a privacy and civil liberties perspective are concerning. Therefore, strategic and systemic measures that do not target individuals should be considered to limit the spread of disease within transit.

3.2.2.1 Disease Transmission in Transit

Reducing the spread of disease in pandemic environments requires multi-faceted approaches. *What Bus Transit Operators Need to Know About COVID-19* [77] is a resource for transit organizations to address their needs during pandemics for minimizing the overall effects of infectious diseases. The report intentionally does not give prescriptive measures for transit agencies to take but rather outlines the areas transit agencies should focus upon to prepare for the impacts of a pandemic. The interventions and actions suggested are largely tactical and operational and reflect the immediate requirements of transit agencies to protect riders, workers, and their community.

Identifying the operational risks is one strategy for minimizing the spread of disease. Goscé and Johansson [78] showed a neighborhood-level correlation between the amount of time spent in enclosed stations by residents and the rates of influenza-like illness in the neighborhood. Bota, Gardner, and Khani [79] proposed a three-stage approach to detect components (e.g., stops and trips) of a public transit system most contributing to the spread of disease with a follow-up proposing a more efficient approach [80]. The commonality between these two studies is the recommendation to increase surveillance and mitigation measures on the vehicle trips identified as most likely to spread disease [79, 80]. All three studies identify the link between specific places in transit (stations in [78] and vehicle trips in [79, 80]) and the spread of disease; however, other than operational surveillance, they do not address strategic changes to reduce the existence of problematic places.

Rather than identify problematic places in transit, Shoghri, Liebig, Gardner, *et al.* [81] identified problematic patterns of people. Simulating various kinds of passengers, primarily belonging to the groups "returners" and "explorers," the impacts of different passenger qualities on the spread of disease were assessed. The most significant contribution to the spread of disease came from "large distance returners," and the smallest contribution was from "small distance returners." These results led the researchers to conclude that the latter group could be excluded from containment measures. This study is an example where the changes proposed to the transit system are not merely operational but would force the targeting of individuals and assign risk to their actions. Perhaps a redesign of transit to not enable problematic passenger patterns would serve to address their existence without specifically targeting individuals strategically, but such changes are not mentioned.

Perhaps the most comprehensive work on the spread of disease, and timely as it focuses on the specific epidemiological characteristics of COVID-19, was authored by Mo, Feng, Shen, *et al.* [82] in which a susceptible-exposed-infectious-recovered-susceptible (SEIRS) disease model was used. The SEIRS disease model introduced both a "novel theoretical solving framework for the epidemic dynamics with time-varying and heterogeneous network structure" allowing population-size analysis with what is characterized as "low computational costs," and an evaluation of both public health and transportation policies that affect the spread of disease. Control policies evaluated include adjusting parameters that proxy individual and medical provider behavior, varying trip occurrence rate, changing the distribution of individual passenger departure times, closing bus routes (employing a variety of route selection strategies), limiting maximum passenger load in vehicles, and isolating specific individuals identified as likely to have a high impact on the spread of disease. Researchers concluded that "[t]he most effective approach is isolating influential passengers at the early stage." Strategic changes investigated did not show the same effectiveness as targeting individuals. Perhaps individual targeting will prove to be most effective, but until all non-individualized options are assessed, it is premature to sacrifice the societal cost of individualized measures.

3.2.2.2 Network Model Growth

Understanding the success of efforts to reduce the susceptibility of the spread of disease, both existing baseline and model baseline provide points of comparison. While the data will provide an existing baseline, a network model similar to the framework's environment needs to be identified for comparison. Amati, Lomi, and Mira [83], noting the growth of social network analysis, undertook a review of social network models. Finding both prior general reviews to be insufficient, and other reviews focusing on special classes of social network models, Amati, Lomi, and Mira focused on "models for the analysis of cross-sectional network data, that is, data generated by one single observation of a (complete) network." Amati, Lomi, and Mira's review is heavily focused on exponential random graph models (ERGMs), citing others' beliefs in the models' suitability at expressing a moment in time for a social network.

Toivonen, Kovanen, Kivelä, *et al.* [84] used ERGMs, network evolution models (NEMs), and nodal attribute models (NAMs), to social network models to empirical datasets of social networks to assess the resemblance. While multiple network metrics were compared by Toivonen, Kovanen, Kivelä, *et al.*, of interest for research looking to reduce the susceptibility of the spread of disease through network edge minimization is the node degree distribution. When fit to the empirical datasets, the social network models produced different node degree distributions. For one of the NAMs, BPDA, the empirical datasets fit well to the Poisson degree distribution. The Váz model, a NEM, produced a degree distribution that decays as a power law, $P(k) \sim k^{-\gamma}$. The other NEMs and a ERGM model produced degree distributions that "all appear to decay slower than the Poisson distribution, but faster than power law."

The decision as to which social network model is used for comparison is not blatantly clear. Studied empirical social networks possess varying resemblance to social network models. Random networks, like ERGMs, with resulting Poisson node degree distributions do not account for high degree nodes present in many empirical social networks [64]. Alternatively, scale-free networks with resulting power law node degree distributions (with $2 < \gamma < 3$) support the high degree hubs observed in citation, email, and internet networks [64]. Fitting empirically observed degree distribution presents a useful indicator as to which social network model may best model an empirical social network.

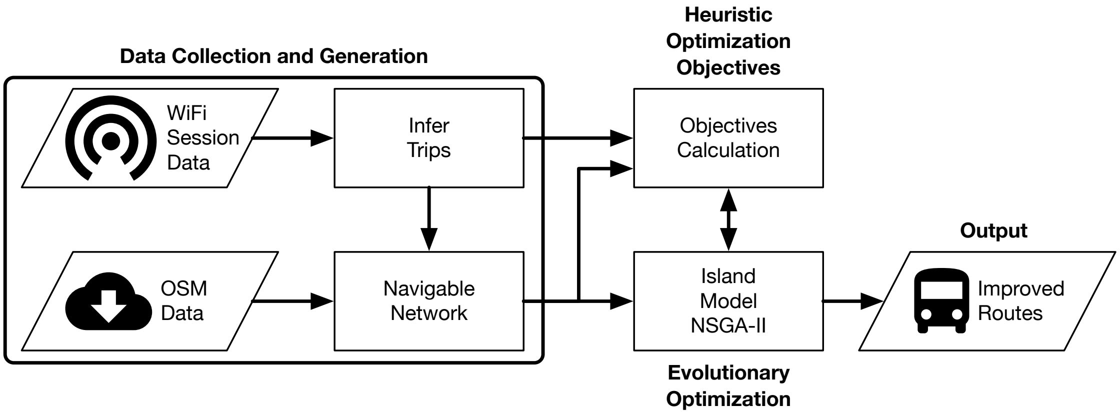


Figure 3.1: Main Stages of the Multi-Objective Heuristic Optimization Framework

3.3 Methodology

The framework introduced in this research serves to ingest O-D mobility data and use it to present feasible route solutions to transit planners that balance the two objectives of (a) maximizing the number of observed trips served by transit, while (b) minimizing the susceptibility of transit to spreading disease. The framework uses a multi-objective UTRP heuristic optimization with an encounter network metric as one of the objectives. Every candidate solution represents the existence of two interconnected networks. The first is the transit network that defines the candidate solution, and the second is an encounter network generated from the simulation of trip mode choice using the candidate transit network.

Figure 3.1 depicts the main stages of the framework, which begins with the processing of WiFi session data. The inferred trips define the geographic area for which a navigable network must be generated. The navigable network defines potential stops that constitute the decision vector's values that the evolutionary optimization evolves. The objective calculation sits outside the evolutionary optimization process, and when called to evaluate the fitness of a decision vector, employs the inferred trips and navigable network to assess objective values. The output is the improved routes once the stopping criterion, a defined number of generations without change to the cumulative Pareto front, is met.

3.3.1 Data Collection and Generation

Two data sources are used as the main inputs. The first data source consists of WiFi user sessions covering the area of interest. The WiFi user session data is used to infer trips that occur between locations within the area of interest. The second data source is a network representation of navigable paths. The scope of the network is defined once the scope of mobility in the WiFi user session data is known. Therefore, the WiFi user session data must be prepared first before any subsequent steps. The two datasets interface later in the framework to assess if an inferred trip occurs faster over the network by foot or by using a candidate solution transit route.

3.3.1.1 Trip Inference

Trip inference occurs by identifying when WiFi user sessions established by personal wireless devices (PWDs) are logged as ending in one building and WiFi user sessions are logged as starting in another. The data logged for each WiFi user session are shown in Table 3.1. All recorded WiFi user sessions during the period of analysis are used. In establishing which WiFi user session pairs constitute trips between locations, considered factors include session duration (including prior, current, and subsequent sessions), distance between WiFi user session access points (APs), and temporal order of WiFi user sessions. Every PWD is assumed to represent an individual, so a single individual with multiple PWDs establishing and terminating WiFi user sessions would be analyzed as multiple people. Future work may better address both individuals with multiple PWDs and those without any PWD.

3.3.1.2 Road Network Generation

Road network data was generated from OpenSteetMap data using the OSMnx package [46]. A spatial envelope slightly larger than the area that contained the locations of WiFi user sessions was used to limit the size of the road network. To best represent the roads available to transit, the OSMnx network designation of 'drive_service' was used, providing all drivable streets as edges with all intersections as nodes.

Every node in the road network is considered a potential stop node and potential stop nodes are assigned index values. Figure 3.2 illustrates the procedure to assign index

Table 3.1: Relevant Fields of the WiFi User Session Data

Field Name	**Data Type**	**Description**
MacPIN	String	Anonymized representation of the MAC address of a personal wireless device (PWD).
SessionStart_Epoch	Unix timestamp	Date and time a PWD initiated a connection with an access point (AP).
SessionEnd_Epoch	Unix timestamp	Date and time a PWD terminated a connection with an AP.
AP_Mac	String	MAC address of AP network back haul.
BuildingCode	String	Building abbreviation where AP is located on campus.

values to potential stop nodes. The solid black line, which moves from the southwest to the northeast corner of the road network, facilitates searching nearby potential stop nodes represented by the orange circles. Potential stop nodes are assigned an incremental index value as they are intersected by the moving black line with Figure 3.2 showing example values.

3.3.2 Evolutionary Optimization

Finding feasible solutions for UTRP-like problems using evolutionary optimization heuristics is an area well explored in recent research [25, 27, 53–56, 71].

The parallel optimization library used for the island model evolutionary optimization of this framework was pygmo. pygmo is based on the C++ library PaGMO for parallel optimization, which uses an asynchronous island model [60]. The pygmo library was developed by a team at the European Space Agency to facilitate high-dimension global optimization problems for spacecraft trajectories and part design. Yet, the universality of the approach and the available algorithms makes it easily employed in other global optimization problems.

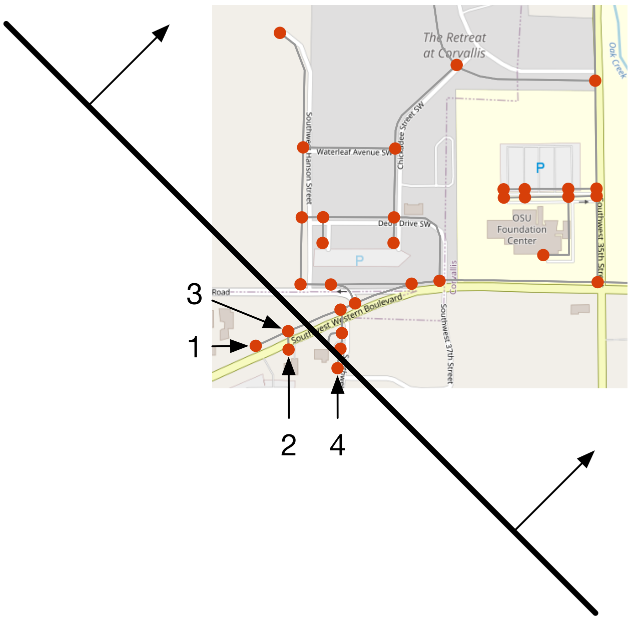


Figure 3.2: Process to Assign Indices to Road Network Nodes

3.3.2.1 NSGA-II and Island Model Parallel Optimization

The parallelization of the multi-objective evolutionary algorithm (MOEA) known as NSGA-II [85] is implemented in pygmo through the use of an approach known as the island model [57]. In the island model, multiple parallel optimizations are run (i.e., the islands), and candidate solutions are shared between the islands through a defined graph topology (i.e., the archipelago). Each candidate solution within an island is referred to as an islander with the solution's values being the islander's decision vector. Using the parallelization that pygmo facilitates, the optimization can scale an archipelago to utilize all available cluster resources while still operating within the constraints of any single island process. The latter is done by limiting the size of an island's optimization to be within the available node resources. While many multi-objective heuristic optimization algorithms exist, NSGA-II was selected for its prevalence as a benchmark against which other multi-objective heuristic optimization algorithms are measured. Future work could investigate other algorithms and parallelization approaches, but such is not the focus of this work.

Table 3.2: Heuristic Optimization Objectives Notation

Notation	**Meaning**
G	Encounter network
$E(G)$	Edge count of G
s	Inferred trip
S	Set of inferred trips
m_s	Mode of trip s

3.3.3 Heuristic Optimization Objectives

Balancing the needs of providing transit to a population against the aim of minimizing the spread of disease is the purpose of this work. To accomplish both, quantifiable objectives are defined. Maximizing the provision of transit is measured by simulating potential rider mode choice using inferred trips and the candidate transit network and counting the number of trips served by the candidate transit network, as shown in Equation 3.1. Minimizing the spread of disease is accomplished by minimizing the number of encounters in shared spaces resulting from transit use (i.e., minimizing the number of edges in the encounter network), as shown in Equation 3.2. Table 3.2 outlines the notation employed in the objectives.

$$\text{Maximize: } \sum_{s \in S} [m_s \neq \text{"walking"}] \tag{3.1}$$

$$\text{Minimize: } E(G) \tag{3.2}$$

3.3.3.1 Rider Mode Choice Simulation

The first step in calculating the objective metric is to take the routes defined in the islander's decision vector and assigning all the trips in the dataset to either a transit route or to walk. Trips are assigned to either the transit route that takes the least amount of time or exclusively walking. Trip assignments that involve a transfer between routes are not considered. An example is shown in Figure 3.3, where the trip would be assigned to use Route A. If Route A were not an option, the trip would walk rather than taking the longer Route B. The process of assigning trips to the fastest option occurs for every

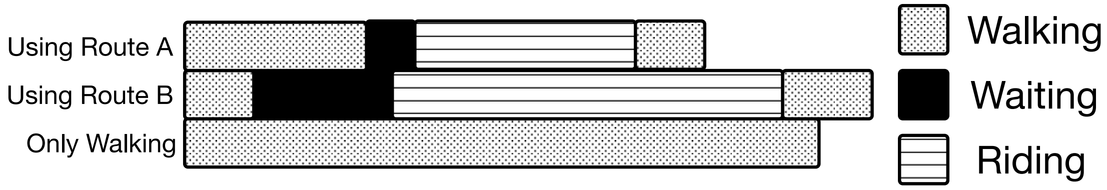


Figure 3.3: Assessing Trip Timelines

trip in the dataset.

Maximizing the count of trips assigned to a transit route as the fastest option is the first objective.

3.3.3.2 Transit Encounter Network

Once trip assignments are made, an encounter network of transit users can be generated. Using the original departure time (i.e., the end time of a WiFi user session), timestamps are generated from the simulation for (a) the duration of the walk to the origin transit stop, (b) the duration of the wait at the origin transit stop, and (c) the duration of the ride in the transit vehicle. From the timestamps and the associated stops and vehicle for all trips using transit, an encounter network is generated from all trips whose timestamps establish a shared presence at either a transit stop or a vehicle.

Remaining generalized in the assumptions about disease transmission, any shared presence at a transit stop or in a transit vehicle is considered an encounter. By defining the objective as the minimization of edges in the encounter network, fewer opportunities exist in those environments exist for direct transmission of disease. Future work could be more tailored to the transmission properties of a specific disease or expand the encounter analysis beyond waiting at the origin transit stop or sharing a transit vehicle.

3.4 Limitations of Selected Approach

Worth recognizing, in addition to the above decision justification, are the limitations inherent in this work. Most prominent is the source of O-D trip data. Using WiFi user session data as the source of O-D trip data presents a large sample size but also limited to those associated with the source network. In environments like universities or large corporate campuses where many within the area of interest are associated with the organization and carry with them devices that associate to the organization's network, this presents a low-investment approach to highly detailed mobility data. In other environments where many within the area would not establish an association with the prevailing WiFi network, using WiFi user session data would not provide a useful capture of mobility. As the heuristic optimization only relies on trips existing within the area of interest, other means of trip generation can be used when WiFi user session data is not well suited. For example, Bota, Gardner, and Khani [79] took one day's AM peak trips and repeated it four times to generate a representative AM peak trips for a workweek.

Next to consider is the use of a disease network model. While standard disease network models like SIS, SIR, and SEIRS can be used in agent-based modeling (or a simplification thereof as done in Mo, Feng, Shen, *et al.* [82]), doing so requires both significant computational resources and disease-specific knowledge. By simply minimizing the number of direct encounters transit users experience (while some assumptions about latent infectiousness of a place are made), other assumptions about the spread of disease are avoided, and a high-caution approach is taken (e.g., no minimum for the duration of contact and the assumption that an encounter occurs every time a space is shared). The edge count of an encounter network is less computationally intensive than an agent-based simulation using the same encounter network.

Lastly, only encounters while using transit in the area of focus are considered. This is less of an issue for this research than for work where agent-based modeling is used [79, 80, 82] as no assumption of disease status is made about parties to an encounter. Mo, Feng, Shen, *et al.* [82] tries to address the shortcoming of outside-of-transit interactions for an agent-based model by considering shared origins and destinations but requires significant assumptions to be made. The downside of using the edge count of an encounter network is that encounters that may have been perfectly safe (e.g., between two susceptible but not infectious individuals) are treated the same as an encounter where disease could spread (e.g., between a susceptible individual and an infectious individual). This downside may needlessly limit encounters that would not have contributed to the spread of disease.

3.5 Results and Discussion

A case study was conducted with WiFi user session data collected during the fall 2017 quarter at OSU. A workweek was chosen with the most inferred trips to demonstrate the application of the proposed multi-objective heuristic optimization framework under the campus's busiest state. The busiest five consecutive weekdays in the dataset were Monday, 9 October, 2017 to Friday, 13 October, 2017. On Saturday, 14 October, 2017, there was a home football game, which may explain an increased level of campus activity on the days leading up to the game. The week is also roughly the middle of the fall 2017 quarter, with mid-term preparations contributing to campus activity. During the busiest five consecutive weekdays, 79,879 trips were inferred.

3.5.1 Inferred Trip Characteristics

The trips inferred for the week of interest have patterns that could be expected of a university campus. Figure 3.4a shows the distribution of the day-of-week in which the inferred trips occurred. The Thursday peak is not regularly witnessed in other weeks, and it could possibly be attributed to football game-related activities; the Friday level is also atypical as other weeks will have reduced activity on Fridays. Both discrepancies suggest this week is not representative of a typical workweek, but that does not impede the utility of testing the multi-objective heuristic optimization framework against the campus's busiest week. Figure 3.4b shows the distribution of the trip start times. The peak observed in the morning with the greatest number of trips around 10 AM is typical of workweeks at the OSU campus, likely associated with class schedules. The late-night peak increasing from 6 PM to 11 PM may be associated with coordinated evening activities on campus. Figure 3.4c shows the observed duration of the inferred trips (i.e., WiFi user session end time in origin building to WiFi user session start time in destination building). The long tail demonstrates a limitation of using WiFi user sessions to infer intra-campus trips. For trips with origins or destinations outside of campus, if the device identifier establishes a new WiFi user session during the same day, it is inferred that an intra-campus trip occurred. To limit the effect of outside-of-campus trips, only trips with an observed duration of 60 minutes or less are used to calculate objectives.

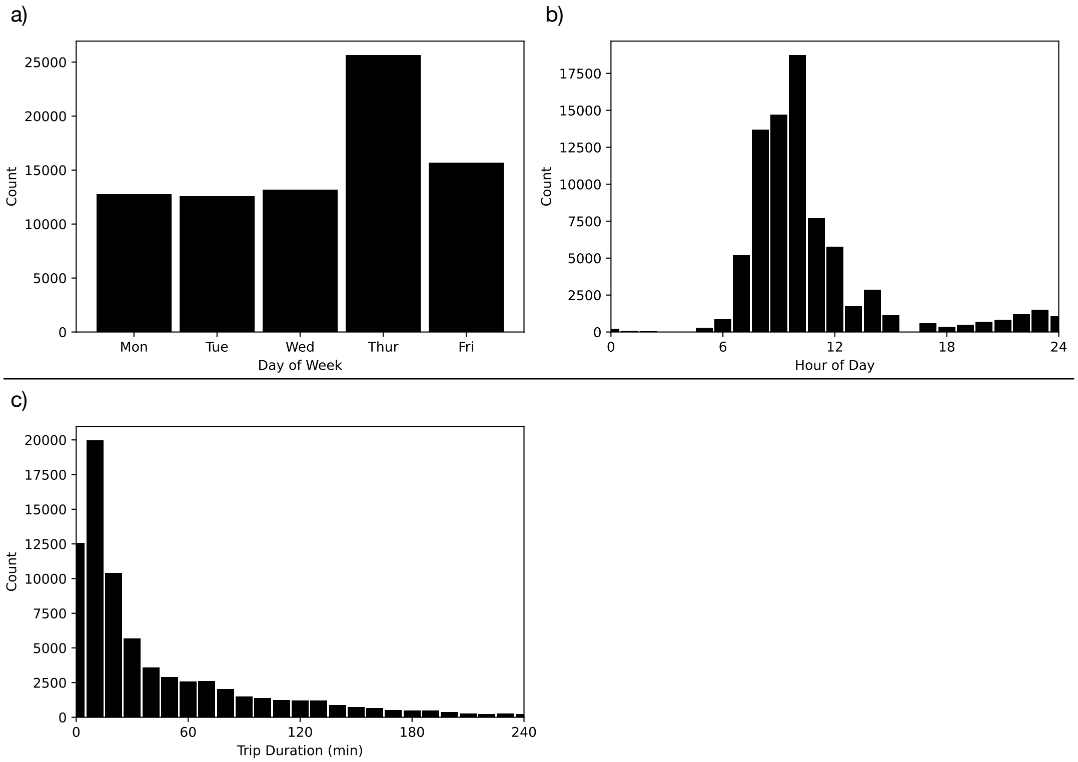


Figure 3.4: Temporal Distributions of Inferred Trips

3.5.2 Comparison to Existing Routes

As the proposed multi-objective heuristic optimization aims to improve the routes used by transit, a baseline comparison is made against the routes contemporary to the WiFi user session data. Figure 3.5 plots the set of non-dominated solutions that constitute the Pareto front alongside the baseline routes subjected to the objective calculations. The baseline routes are dominated by multiple solutions provided by heuristic optimization. The feasible solution with the closest number of riders (1.2% more than baseline) provides a 10.7% reduction of encounter network edges; the feasible solution with the closest number of encounter network edges (0.7% fewer than baseline) provides a 5.8% gain in riders. Figure 3.6 plots the routes of the feasible solution with the closest number of riders (when compared to the baseline) and the baseline routes onto the road network of the OSU campus for comparison. The grey lines in Figure 3.6 represent the road network edges with every intersection being a node. Each non-grey line plotted atop the road network is a bus route. The routes produced by the feasible solution appear less centralized than the baseline routes that at some point each operate adjacent to the campus's football stadium.

3.5.2.1 Encounter Network Comparison

Assessing the performance of the framework can be done by comparing the encounter network metrics of the baseline existing routes to a non-dominated solution serving a similar number of riders. The two distributions are compared in Figure 3.7.

Figure 3.7a depicts the distribution of node degrees in the encounter networks. Aside from reducing the number of edges in the encounter network of the feasible solution, the distribution of node degrees also changes. While there are fewer nodes with degree greater than 10, there is an increase in the number of nodes with degree 10 or less. Although there is not necessarily a direct connection between the number of nodes in the encounter network and the number of inferred trips served by transit, the difference between the two distributions suggests minimizing edges in encounter networks while still serving trips is done through the reduction of riders having encounters with many others. This concept is similar to the idea of reducing super-spreaders suggested in other research [65, 75, 76, 82].

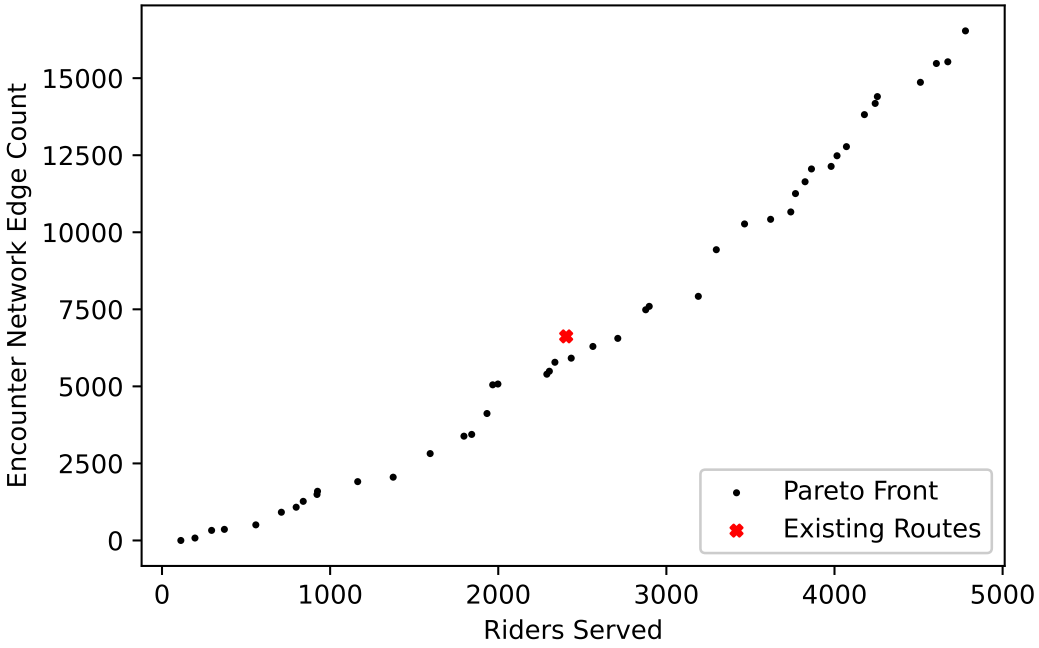


Figure 3.5: Pareto Front Comparison to Baseline

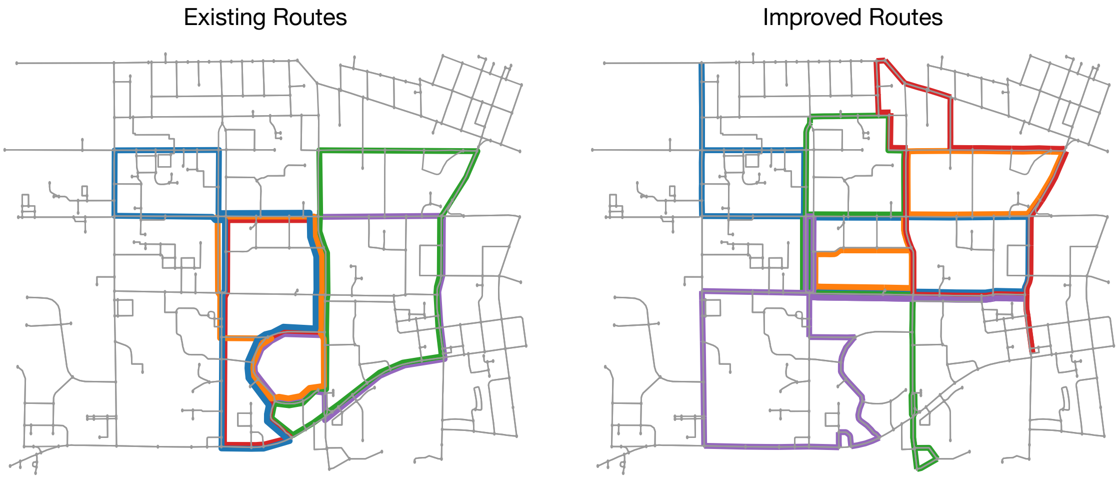


Figure 3.6: Closest Rider Count Solution Comparison to Baseline

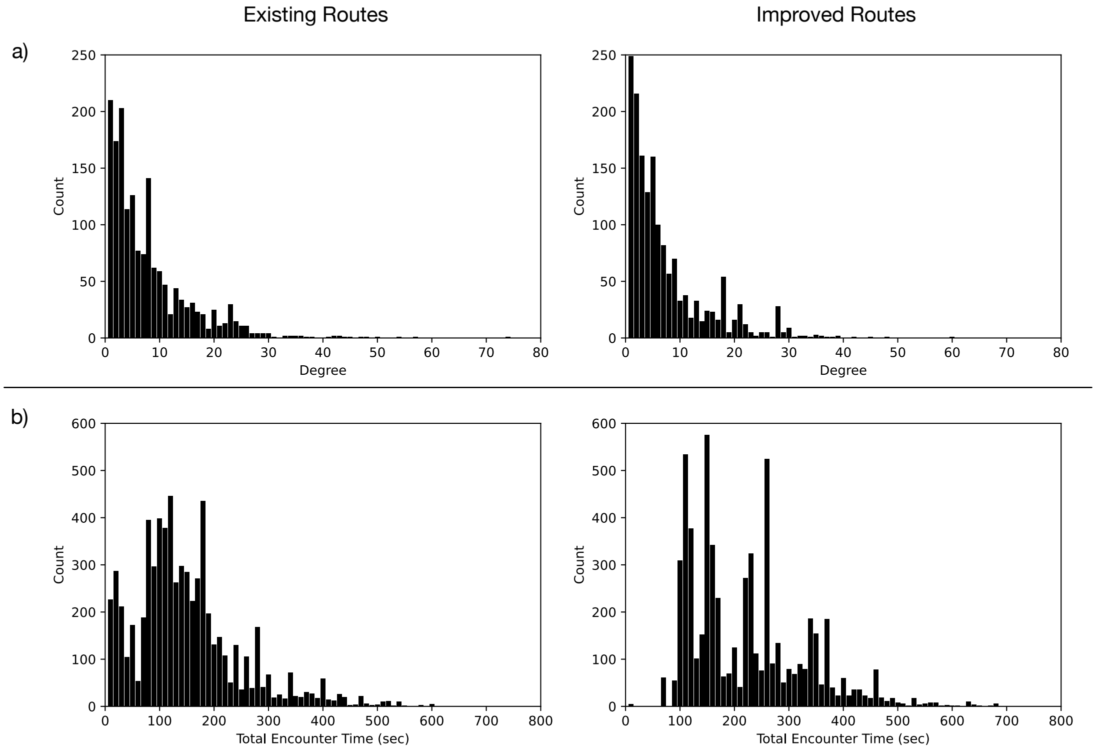


Figure 3.7: Encounter Network Distributions

Figure 3.7b is the distribution of total contact time between nodes in the encounter networks. Compared to the change in the node degree distribution, the difference between the baseline and the feasible solution's encounter network is stark. Over the analyzed week, the feasible solution's encounter network almost eliminates cumulative encounters between nodes that last less than 80 seconds, whereas the baseline encounter network experiences a peak in that area. The change in distribution indicates a shift in cumulative encounters towards riders that are to find each other having more prolonged encounters and edges in the encounter network that are fleeting being all but eliminated.

3.5.3 Comparison to Network Growth Models

The Pareto front in Figure 3.5 shows growth in encounter network edges as the number of nodes (i.e., riders) increases. Comparing the growth of the Pareto front to known social network models provides an indication of the performance of the proposed multi-objective heuristic optimization framework at minimizing encounter network edges. The selection of the comparative model is dependent on the generalized representation of the environment in which the proposed multi-objective heuristic optimization framework's encounter network grows. Social contact networks are often scale-free networks [64, 65]. Since the encounter network of the proposed multi-objective heuristic optimization framework has a fixed number of nodes (i.e., riders), it is possible to further specify a suitable model as a bounded scale-free network [65, 86]. Figure 3.8 shows a comparison of the Pareto front, earlier used in Figure 3.5, to the typical bounds of a social contact scale-free network, where the probability P of node degree k is approximated by a power-law distribution of the form $P(k) \sim k^{-\gamma}$. The shaded region between the lines where $\gamma = 2$ and $\gamma = 3$ shows where social contact networks that exhibit bounded scale-free network characteristics typically reside. The use of a scale-free network as a comparative model is further supported by the empirical distributions shown in Figure 3.7a that align with a power-law distribution. Figure 3.8 shows that below 1,500 riders served, the proposed multi-objective heuristic optimization framework returns expected or better than expected levels of performance when measured by the number of encounter network edges. Above 1,500 riders, the proposed multi-objective heuristic optimization framework returns more edges than expected of a typical bounded scale-free network.

The shift from out-performing to under-performing a bounded scale-free network may be the result of another property of the environment not considered in the bounded scale-free network model, i.e., finite encounter spaces. Regardless of size of the encounter network, the number of spaces where an encounter can occur (i.e., an origin bus stop or vehicle) is fixed. When only a relatively low number of riders use transit, the fixed number of encounter spaces is not an impacting constraint. As the number of riders increases, the number of spaces does not, likely forcing an increase in the number of encounters each rider experiences (i.e., the mean degree of nodes increases). A network model addressing the fixed number of encounter spaces could not be identified.

3.5.4 Heuristic Optimization Stopping Criteria

Selecting stopping criteria is central to implementing an MOEA well, and a parallelized MOEA is no exception. Gutierrez, Adamatti, and Bravo examined stopping criteria

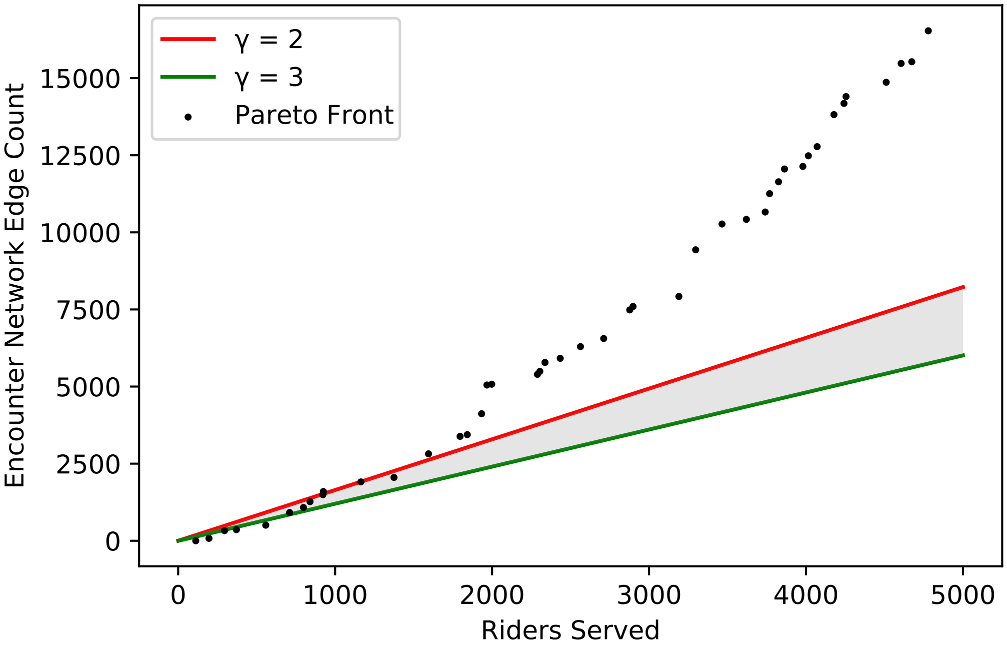


Figure 3.8: Bounded Scale-Free Network Growth Comparison

for multiple MOEA, including NSGA-II, and developed new stopping criteria that they demonstrated reduced the amount of time needed to achieve comparable results to prior common stopping criteria. Of the three common stopping criteria cited by the authors, only one does not require significant prior knowledge about anticipated results, i.e., "stopping when no significant improvement in the quality of the solution set can be obtained by performing further generations." For this research, the stopping criterion was assessed using a cumulative set of non-dominated solutions. If the cumulative set of non-dominated solutions did not improve for a defined number of generations, then the stopping criterion was met. A chart of the resulting Pareto front from four different stopping criterion is shown is Figure 3.9. Figure 3.10 shows the time needed by the heuristic optimization to meet the stopping criterion.

Changing the stopping criteria by increasing the number of generations needed without a new non-dominated solution results in solutions with additional simulated transit riders but at the cost of a substantial increase in the amount of computational time. Importantly, the solutions of faster stopping criteria are comparable in edge count for a given number of riders. Recognizing that the methods explored in this research would be implemented in environments where there is an interest in minimizing the spread of disease, the solutions with the lower encounter network edge counts are likely of greater interest to planners. While it should be recognized that more computational resources would be necessary for solutions providing higher ridership, the options with a stopping criteria requiring a lower computational time will provide comparable solutions in the areas planners are likely to seek.

3.6 Conclusion

This research has introduced a multi-objective heuristic optimization framework that provides actionable results to transit agencies to balance the utility of service to riders against the susceptibility of routes to enable the spread of disease in a community. A case study using empirical WiFi user session data to infer trips on a university campus showed how the proposed framework could be applied. The choice of which route set to implement is ultimately up to the transit planner with the framework allowing for a mix of defined and framework-generated routes to provide additional flexibility.

Multi-objective heuristic optimizations of the UTRP allow for the addition or re-

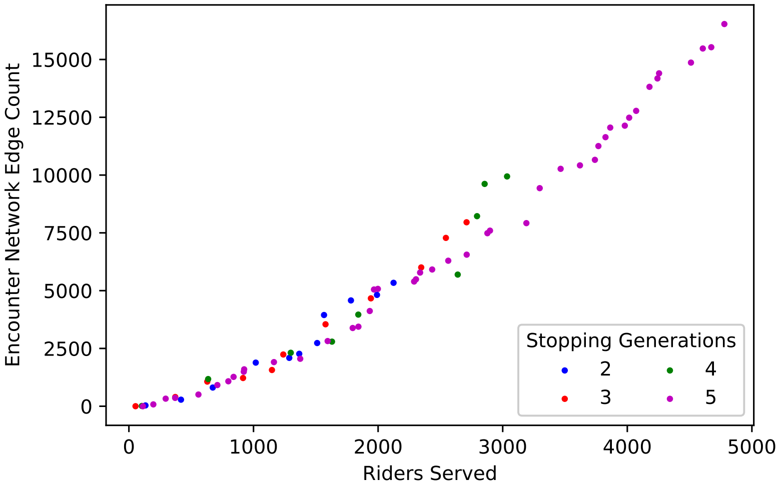


Figure 3.9: Pareto Front Variations with Stopping Criteria

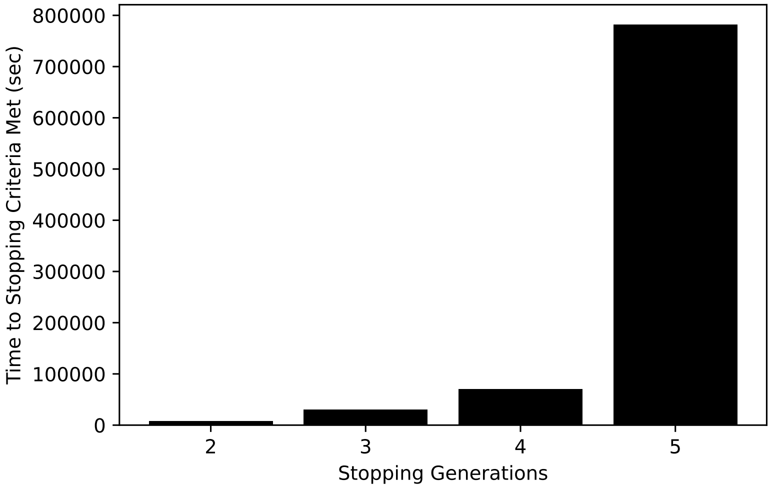


Figure 3.10: Heuristic Optimization Duration for Stopping Criteria

moval of objectives as desired. This work includes encounter network analysis in the heuristic optimization, which the researchers believe is novel for UTRP. Future work into integrating encounter network objectives into other UTRP landscapes would enable transit agencies to add a disease transmission component to the other objectives they are already pursuing. Other UTRP variables like improving equipment assignment efficiency or dynamic route schedules are feasible with the existing objective, but require an expansion of the heuristic optimization search space demanding additional computational resources. Additionally, using the encounter network generated in this framework, other encounter network characteristics could be used as objectives such as encounter duration or repeated encounters that may better represent disease susceptibility.

Chapter 4: Unsupervised Multi-Source Anomaly Detection Framework for Vehicular Traffic Counts[1]

4.1 Introduction

Transportation agencies collect traffic data to understand how a road network is utilized. Traffic data may be collected through methods like inductive loop detectors, radar, or third-party data aggregators. Traffic data sources are relied upon to be accurate as the data collected is used to support planning, management, and improvement decisions. Past anomaly detection mechanisms used only a single traffic data source for a given site population, relying on approaches like modeling past behavior or developing networks from neighboring sources. With emerging traffic data collection technologies, it is now reasonable to have multiple sources for a single population.

In data collection environments where multiple data sources exist, each data source has biases with its collection approach, though many claim a 100% sample of the population. Recognizing that different systems exhibit different biases (e.g., some may be affected by weather while others under-count motorcycles), understanding the relationship in sampling behavior is necessary to recognize when the behavior changes, which possibly indicates an equipment malfunction.

The gains to be experienced by implementing anomaly detection mechanisms that employ methods presented in this work could be profound. Currently, anomaly detection mechanisms employed by agencies utilize high thresholds before generating an alert because their coarse processes would unduly burden staff with false alerts at lower thresholds. The combination of high thresholds and coarse processes reduces data quality under certain anomalous conditions without the agency being made aware. By assessing equipment performance against comparable equipment measuring a like population, the need for regularly scheduled calibrations could be significantly reduced without diminishing ongoing confidence of data accuracy. Significant changes in traffic data would be

[1] Research methodology presented as S Hoover, *Methodology to Validate Accuracy of Automated Systems Using Cell Phones*, Conference on Performance and Data in Transportation Decision Making, Atlanta, Georgia, Sep. 2019

accompanied by a high degree of certainty that the change is not merely an equipment malfunction; how the change would appear across data sources under nominal conditions would already have been predicted. Any differing behavior would be indicative of an anomaly. The benefits to transportation agencies would come both in reducing staff time and resources in maintaining traffic data collection equipment and in an increase in confidence that the data being used to make decisions accurately reflect real-world conditions.

4.2 Literature Review

Emerging traffic data collection technologies provide an opportunity for new anomaly detection methods to be employed. To understand the potential for new anomaly detection methods, past methods that are largely dependant on a single data source for a given location are identified and reviewed for their contributions and limitations. Next, a developing area of recurrent neural networks (RNN) known as Long Short-Term Memory (LSTM) networks are identified as a potential method to support anomaly detection. Introducing a new anomaly detection mechanism to support identifying potential data collection malfunctions could provide data quality assurance to those seeking to use the data while reducing the workload required for collecting the data to validate its accuracy periodically.

4.2.1 Vehicle Count Anomaly Detection

New sources of traffic data that can provide independent counts are only recently emerging. Lack of perceived benefit of seemingly redundant counters may have limited the availability of multiple independent sources. As such, anomaly detection mechanisms for vehicle counts evolved around single-source anomaly detection methods. Anomaly detection methods developed for a single-source exhibit similar challenges to what would be faced by a multi-source method.

Alam, Gerostathopoulos, Amini, *et al.* [89] developed a method for anomaly detection of univariate time-series traffic data in which every data point is assessed two scores: a point-distance score and a difference-distance score. The point-distance score evaluates a data point against others of the same day-of-week and time-of-day over a year, whereas the difference-distance score evaluates a data point against the one immediately proceeding it in the time-series. Thresholds for each score are then established to identify if a data point is anomalous. Identifying a lack of traffic data labeled as nominal or anomalous, when available, subject matter experts (SMEs) were engaged to annotate anomalous traffic data. In the absence of SMEs annotation, an automatic threshold selection process using a modified z-score was developed. Locations of traffic data were clustered according to their Dynamic Time Warping distance using k-means. The clustering allowed the researchers to employ thresholds identified for a location at other locations within its cluster. The authors note that identifying an anomaly is still a largely subjective undertaking, with their efforts being to provide some efficiencies through clustering and applying subjective inputs from SMEs and developing an automated process that may be used. The use of SMEs to designate what constitutes anomalous counts and then extending the trained thresholds is one approach to defining anomalous counts.

Megler, Tufte, and Maier [90] focused on novel data cleaning methods and employing the processed data for travel time prediction. While indicating initially that supervised learning had been the intent, consulting with SMEs revealed the difficulties in labeling data. An unsupervised method, k-means, was used to classify data. Training data from the month of May was used, citing May as a "representative month" due to the lack of holidays and inclement weather. Two different kinds of anomalies were identified: "cases where there are enough instances of an anomaly for them to form their own cluster; and outliers, where some observations are unlikely to be valid but are each dissimilar from the others." The validation included both statistical methods and input from SMEs, which was noted as being heavily relied upon. Removing identified anomalous data from the inputs to predict travel time resulted in $\tilde{2}0\%$ of predicted times differing from initial predictions. The gains over the initial method are marginal, but the authors "believe that with refinement these results can be further improved, and that they warrant additional experimentation."

Relying on SMEs to determine what is and is not anomalous places a heavy burden on subjective human judgment. To try to move toward more automated methods Riveiro, Lebram, and Elmer [91] developed a visual approach to explore traffic data, analyze models built from the data, detect anomalies in the data, and provide an explanation for the anomalies. Noting the complexities of automated assessments of anomalies, the authors believe that their visual approach "may close the gap between the domain experts and the data mining engines." To identify anomalies, k-means clustering and Gaussian Mixture models were employed in tandem to identify anomalies. While not solving what may constitute an anomaly, the research demonstrates that increased automated anomaly detection is possible with further development.

Not all current methods are single-source based. Liu, Zhao, Sharma, *et al.* [92] developed a spatiotemporal pattern network (STPN) to identify traffic patterns. By evaluating traffic pattern characteristics using the STPN, the researchers could use multiple data sources despite each source counting a different vehicle population. By comparing the similarity of STPN metrics daily using a structural similarity index, days with anomalous characteristics are identified. The researchers associated weather and traffic incident information with days identified as anomalous by their method to verify the effectivity of their approach with mixed results.

Evolving single-source anomaly detection to gain from new data sources requires introducing new approaches that can use multiple-source sampling from the same population as input. Some challenges may not be overcome with these new approaches, like the subjective assessment of what constitutes an anomaly, but the expansion of potential data sources could facilitate even further model training to move toward automated anomaly determination.

4.2.2 Long Short-Term Memory Networks

The potential coexistence of two sets of regularly imperfect sample data from the same population (e.g., loop-detector and radar data) presents an opportunity. The data generated by these two sources can be construed as a time-series of vehicle counts, and when processed to find where they overlap in both space and time, the result is a multivariate time-series available for analysis. This research aims to minimize the burdens of equipment validations; using the collected data for unsupervised training (i.e., not knowing what is nominal and anomalous) serves to fulfill that aim. Examining the current landscape of unsupervised anomaly detection frameworks that employ multivariate time-series data leads to Long Short-Term Memory (LSTM) networks as a clear choice.

LSTM networks, an evolution of RNNs, are suitable for developing an understanding of what is typical and facilitating identification when data is atypical or anomalous. Anomaly detection frameworks based on LSTM networks have been developed and applied in a wide variety of application domains, including transportation, medicine, and computer networks. Feng, Yuan, and Lu [93] developed a framework for detecting abnormal events in a surveillance video. In one experiment, the researchers processed the surveillance video of a pedestrian walkway for which expected anomalous events included the presence of a car or a bicycle. Using the proposed LSTM framework, an 11.1% error rate was achieved, outperforming ten other comparative frameworks. Taylor, Leblanc, and Japkowicz [94] employed an LSTM network to learn typical bus command traffic on a controller area network (CAN) bus and alert to anomalies if unexpected data appeared. Due to the nature of the interactions between the operator and the vehicle, the focus was on achieving a false positive rate (FPR) level that would prevent operator fatigue (i.e., causing vehicle operators to ignore alerts, thus nullifying the purpose of the system). Performance varied across the devices on the CAN bus highlighting the variable effectivity based on device behavior and demonstrating the challenge of limiting FPR while maintaining utility as measured by the true positive rate (TPR). The researchers also explored what would be required to achieve a 100% TPR, which resulted in some devices having as high as a 63.4% FPR (a rate likely to lead to operator fatigue). It is evident from the research study that the anomaly detection performance of an LSTM network cannot be a simple determination but relies heavily upon both the nature of the typical data and the atypical anomaly.

Vinayakumar, Soman, and Poornachandran [95] used LSTM networks to detect anomalous behavior in computer network log entries. Their approach was distinct in its use of stacked-LSTM (S-LSTM), whereby a recurrent LSTM network layer was added to the existing LSTM network within a hidden layer. Using sensor data from the infrastructure of a cloud service provider labeled as normal or anomalous, this approach had an accuracy of 99.6% in detecting an anomaly and a FPR of 2%. The researchers believe the success of the S-LSTM approach compared to others was due to “the fact that the S-LSTM has capability to learn long-range temporal dependencies quickly with sparse representations in the absence of preliminary knowledge on time order information,” a situation similar to that faced with mobility data.

4.2.3 Literature Review Summary

The review of previous literature shows an opportunity to develop new methods for vehicle count anomaly detection. Data sources, both explicit and implicit, are increasingly available as secondary checks against primary sources of vehicle count data. However, neither collection source can be guaranteed to function without fault, nor can the relationship between the sampling of each source be assumed to be as simple as a linear relationship. Thus, employing modern machine learning techniques like LSTM networks that can be trained to model complex multi-source relationships presents an opportunity to employ vehicle count data sources better to ensure consistent collection behavior.

4.3 Methodology

The main objective of the framework proposed in this research is to reveal whether or not two spatially and temporally overlapping datasets show evidence of a shift in counts present in one dataset that does not appear in the other. Such an anomaly could indicate an undesired malfunction in the operating state of the equipment commonly used to automatically count vehicles (e.g., loop detectors). To meet this objective, the anomaly detection framework requires two distinct phases: dataset preprocessing and anomaly detection.

4.3.1 Datasets Preprocessing for LSTM

The anomaly detection framework requires two datasets of counts from the same vehicle population as input. The datasets need to be independently collected. The independent collection is necessary because any malfunction of equipment shared by the collection methods may affect both methods and obscure the malfunction from the framework. By requiring fully independent collection, any malfunction that the framework is seeking to detect would only affect one dataset. The sampling methodology of the two dataset collection methods may be the same or different with the role of the LSTM layer to predict how the methods relate under varying source inputs.

Independent datasets do not necessarily possess the same spatial points of collection, the same temporal periods, or the same temporal sampling rates. Therefore, it is sometimes necessary to rectify the spatial or temporal states of the datasets to facilitate comparison.

The focus of the rectification process is ensuring that the sample counts represent the same population. Depending on the data sources, this can include:

- Identifying common spatial groups.
- Understanding that while populations may be sampled at different spatial and temporal points, that the populations are the same.
- Resampling any differences in spatial and temporal scope or frequency to align the two sources.

Implemented in this framework is a temporal rectification. The datasets are collected at spatially different locations, but from the same population, and with a temporal offset that can be estimated with empirical data. Applying the empirically estimated temporal offset to the data source with the higher frequency of collection, a rectified temporal alignment is achieved by offsetting the time series data by the nearest multiple of the frequency of collection to the empirically estimated temporal offset. An example is shown in Figure 4.1 between two data sources, A and B, with each collection period represented by a lowercase letter and a sequential number (e.g., a_0 and a_1 are the first and second collection periods of source A, respectively). A data source with a 20-second collection frequency, depicted as A, and a 38-second offset estimation, that which is between A and B in the example, will have the time series shifted by two points in the time-series, as shown in Step 1 of Figure 4.1.

Once the necessary temporal alignments are implemented, each data source must be aggregated to the coarsest spatial and temporal resolution present in the source data by summing the finer resolution data. An example of temporal aggregation is shown in Step 2 of Figure 4.1. The aggregation supports the comparison of like populations.

4.3.2 LSTM Anomaly Detector

The first step of this phase is to build a prediction model of the behavior of the relationship between the two data sources. The collected data, preprocessed and presented as the difference between the two data sources, is then compared against the prediction model generated by the LSTM network of typical behavior for the difference of the two data sources to facilitate identification of any periods of potentially anomalous data. Periods of potentially anomalous data are indicated and may warrant additional inspection. Based

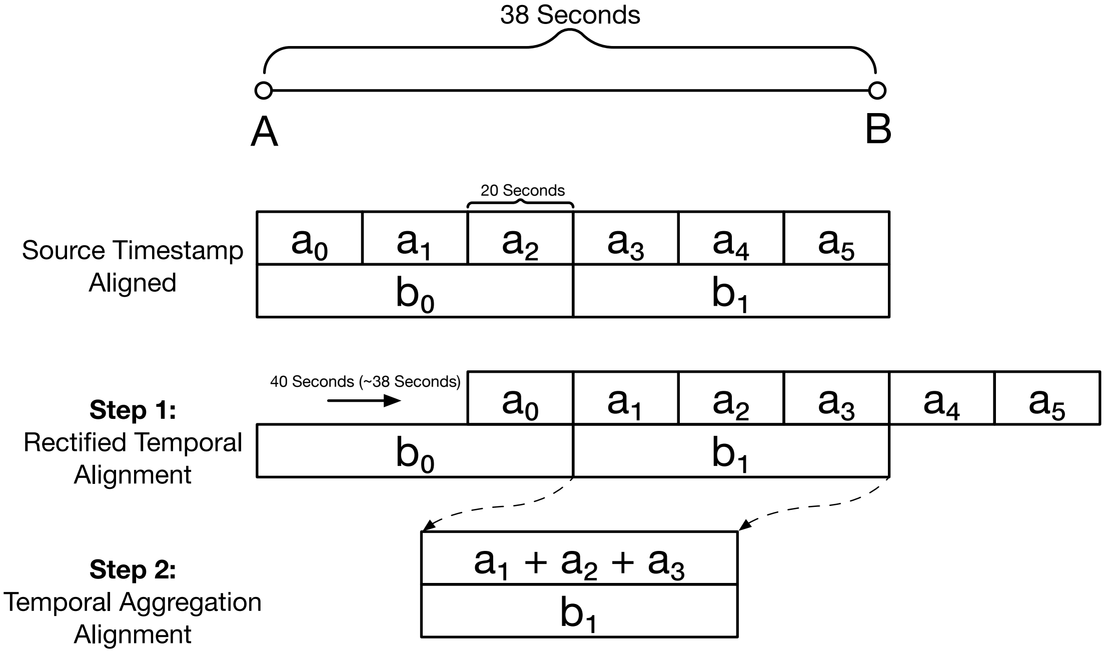


Figure 4.1: Temporal Rectification Between Sources A and B

Notation	**Meaning**
I_n	Vehicle Count Data from Source n
O	Network Output
t	Point in Time of Time-Series
m	Metric Value
q	Standard Deviation Multiplier

Table 4.1: Anomaly Detection Calculation Notation

on prior work [93, 94], an RNN with LSTM layers was selected as a suitable approach for unsupervised training of a model using multivariate (e.g., in this instance, multiple data sources) time-series data.

RNNs with LSTM layers allow the processing of value sequences. In the case of this research, a time-series is employed to represent vehicle counts. The evolution of the LSTM cell in an RNN allows the network to incorporate long-term trends that exist in the data sequences that would otherwise be lost due to vanishing gradients [96] (e.g., a once-a-week occurrence may be lost in a 'vanilla' RNN if using a sampling period of one-day). For vehicle counts, it is beneficial to include long-term trends in predictive approaches as changes on both a weekly and seasonal basis are present [97].

The Keras library was used as the interface to employ the LSTM network because of the availability of LSTM layers, its backend flexibility, and perceived ease-of-use. Backend flexibility was important when developing this framework as different backends were needed to test on older hardware not supporting modern instruction sets (since version 2.4, Keras only supports TensorFlow).

Employing Keras as a high-level API for both Theano and TensorFlow backends (backend selection is dependent on the platform being employed), a sequential model is created with one LSTM layer and one densely-connected neural layer, as shown in the box labeled "Neural Network" in Figure 4.2. The notation used between steps is defined in Table 4.1. The prediction model states are reset after each training epoch as each new epoch begins again at the start of the timeline.

Input features are the vehicle counts from each dataset and engineered features derived from the vehicle counts.

The prediction model is employed as an autoencoder, using the datasets' count difference as the output, as shown in Table 4.2. Once the prediction model is trained,

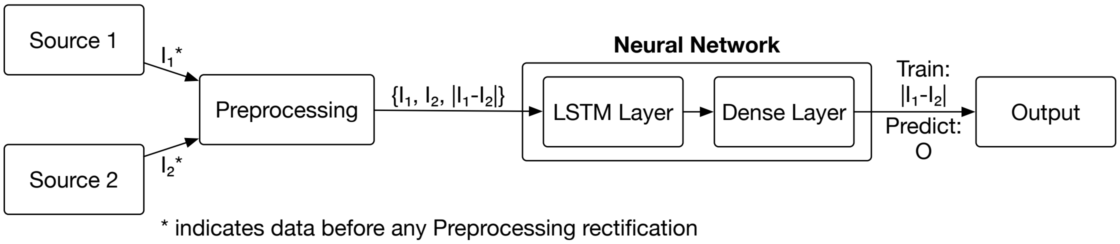


Figure 4.2: LSTM Process Overview

	Training	**Prediction**
Input	$\{I_1, I_2, I_{\lvert I_1-I_2\rvert}\}$	$\{I_1, I_2, I_{\lvert I_1-I_2\rvert}\}$
Output	$\{I_{\lvert I_1-I_2\rvert}\}$	O

Table 4.2: Inputs and Outputs

each sample period of collected data is input, and the l^2-norm of the difference between the collected and predicted differences is the metric value, m, as shown in Equation 4.1. Using the l^2-norm allows for the calculation of a single continuous value to represent the difference between the collected data and the prediction model output, thus facilitating a direct comparison between the difference in performance from one period to another.

$$m = \sqrt{\sum_t (i_{|I_1-I_2|_t} - O_t)^2} \tag{4.1}$$

A threshold, shown in Equation 4.2, is defined as a multiple of the standard deviation above the mean metric value (i.e., the mean l^2-norm). The threshold is tuned based on data quality, and an anomaly is flagged for all periods exceeding the threshold. A period tagged as an anomaly represents a period where the relationship of counts collected by each independent data source varies from the trained prediction model. While a flagged anomaly may represent an equipment error in need of remedy (e.g., reduced receiver sensitivity or increase in background noise), other factors such as a change in vehicle size distribution or weather effects may account for the anomaly tag.

$$Threshold = \bar{m} + q\sigma^2 \tag{4.2}$$

4.4 Results

As highlighted in Section 4.2.1, prior research in traffic data anomaly detection has focused on single-source data. The framework presented in this research is distinct as it requires two independent sources of data. With emerging new traffic data sources, the expectation is that soon, more secondary sources of traffic data will become available.

The case study presented in this section is that of two traffic counting stations in the Portland, Oregon, metro area. This pair of stations proved suitable for collecting comparable data from the same traffic population through an evaluation of hundreds of potential traffic counting stations.

4.4.1 Data Sources

The data were sourced from two independent systems. The first dataset came from Portland State University's PORTAL, a "transportation data archive for the Portland-Vancouver Metropolitan region." One of the datasets archived is "20-second granularity loop detector data from freeways in the Portland-Vancouver metropolitan region." Loop detector data from station 3117 constitutes the first dataset. The second dataset comes from an Automatic Traffic Recorder (ATR) maintained and operated by the Oregon Department of Transportation (ODOT). ODOT has ATRs throughout the state collecting traffic data, including traffic volume and vehicle type. Traffic volume data from ATR 26-016 constitutes the second dataset.

4.4.1.1 Location

The two datasets are collected approximately 0.64 miles apart. Loop detector station 3117 collects data at milepost (MP) 297.6 for northbound Interstate 5 (I-5) traffic. ODOT's ATR 26-016 is located at MP 298.24 and collects both northbound and southbound traffic data. Because the loop detector station only collects northbound data, only northbound data is implemented in this analysis.

Despite the two locations being 0.64 miles apart, they are suitable for use in this analysis because there are no ingress or egress opportunities between the two locations, as shown in Figure 4.3. Therefore, it can be presumed that traffic passing one location also passes the other.

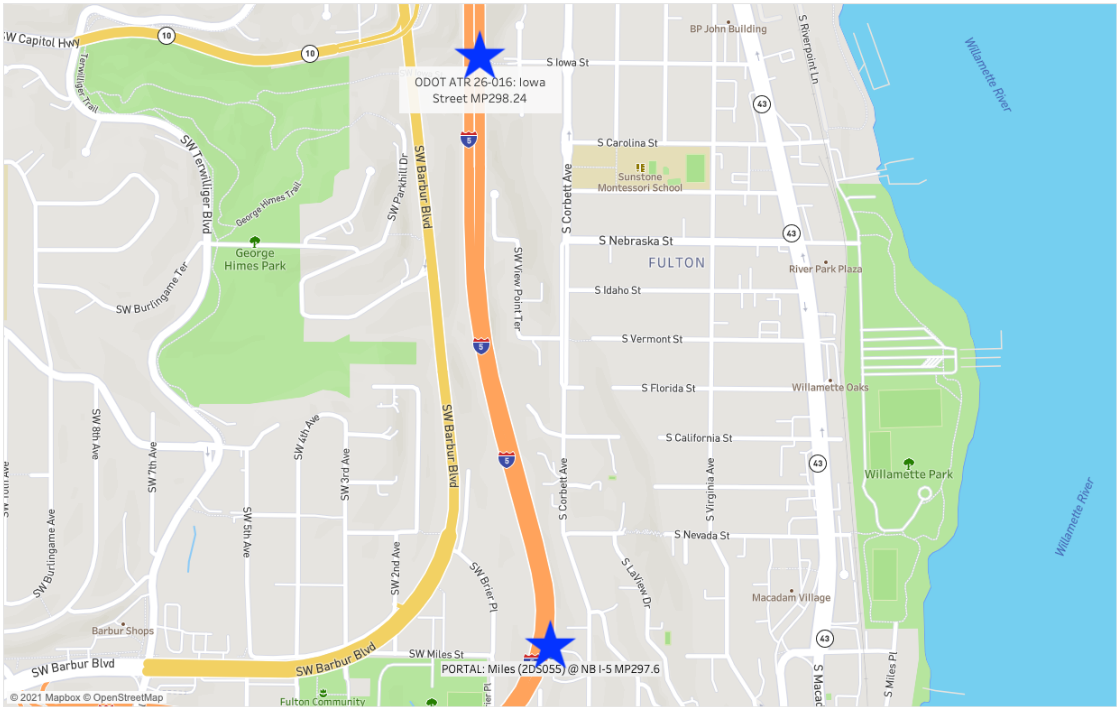


Figure 4.3: Map of Data Collection Locations

To compensate for the geographic separation in the timestamped data, an assessment was made of the time required to travel between the two stations. Using Google's Distance Matrix API, it is possible to get a travel time estimate from one of three available empirical, data-derived traffic models. Using the 'best_guess' model [98], the travel time between the two locations was estimated at 38 seconds. As the loop detector data at station 3117 has 20-second granularity, a 40-second offset was used to align the two datasets (see Step 1 in Figure 4.1).

4.4.1.2 Trends

Plotting the traffic count data aggregated daily suggests that while general trends are shared between the two sources, the loop detector data regularly reports lower aggregated counts. This behavior is depicted in Figure 4.4.

Figure 4.5 depicts the average count by day of week and hour of day, separated into weekday and weekend data. This plot shows consistently lower loop detector counts throughout. However, when comparing the plots in Figure 4.4 and Figure 4.5, the data suggest that difference is not consistent. Across the examined period, the loop detector count is 10.8% lower than ODOT's ATR count.

4.4.2 Framework Analysis

A sample unit of one week is used for this case study. To maintain sample consistency, only weeks with data for all seven days, considered Sunday to Saturday, inclusive, are included in further analysis. The resulting period of analysis is 5 August 2019 to 29 September 2019, continuous.

4.4.2.1 Difference Feature Engineering

For use as both one of the inputs and the output of the autoencoder, the absolute difference of the two datasets is calculated. The calculated absolute difference grouped by hour is plotted in Figure 4.6. While Figure 4.6 shows periods of greater differences between the two sources, Figure 4.7 shows that the majority of hours result in counts with a difference of less than 200.

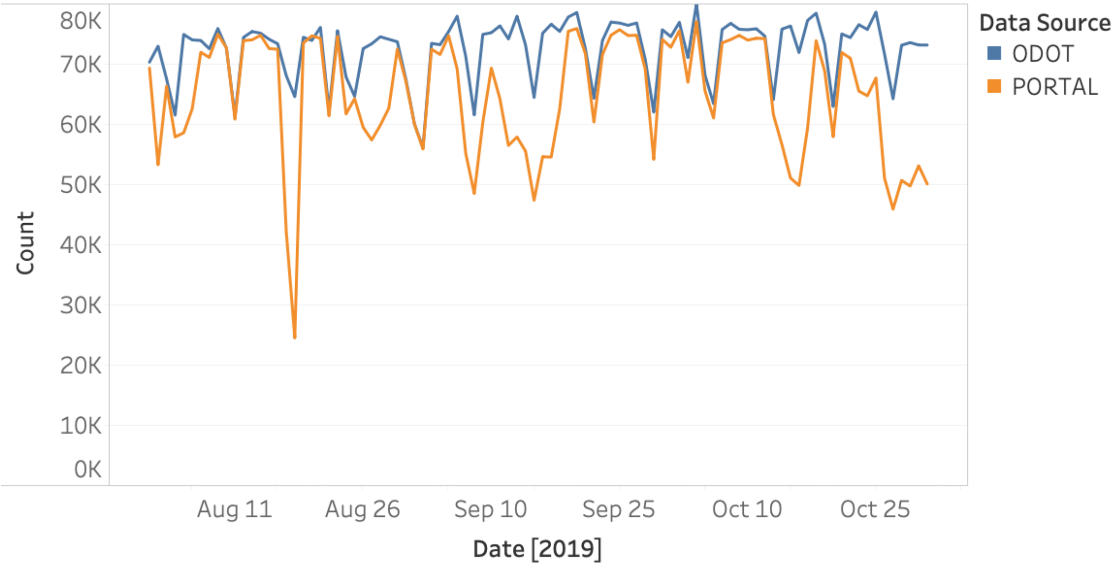


Figure 4.4: Daily Aggregated Counts of Traffic

4.4.2.2 Predicted Difference

Following the training of the LSTM network, the three input features are fed into the network, with the output being the predicted difference between the two data sources. The resulting predicted difference is plotted in Figure 4.8. The predicted differences follow the pattern of lower count differences overnight (when vehicle counts are typically lower) and higher count differences during the day as shown in Figure 4.9.

4.4.2.3 Threshold

For this case study, a daily anomaly threshold was used. The absolute value of the hourly difference between collected and predicted counts is summed for each day. As discussed in Section 4.3.2, the threshold is a mean value plus a multiple of the standard deviation. In this instance, one standard deviation is used as the multiplier. For the analyzed data, the daily threshold is 3,520.6. The daily differences (both predicted and collected), the difference of daily differences, and the threshold are plotted in Figure 4.11. Out of the 56 days plotted in Figure 4.11, 11 qualify as anomalous using a multiplier of one standard deviation. To determine a suitable multiplier and resulting threshold, SME input would

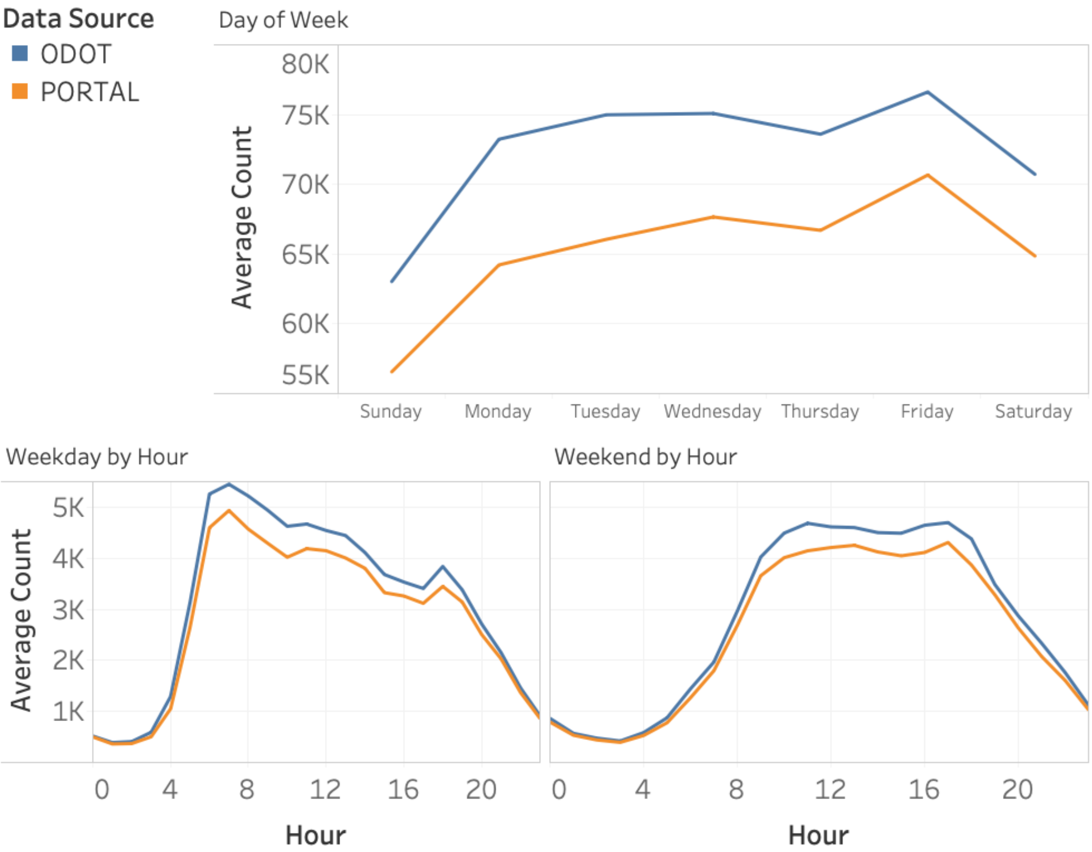


Figure 4.5: Day of Week and Hour of Day Aggregated Counts of Traffic

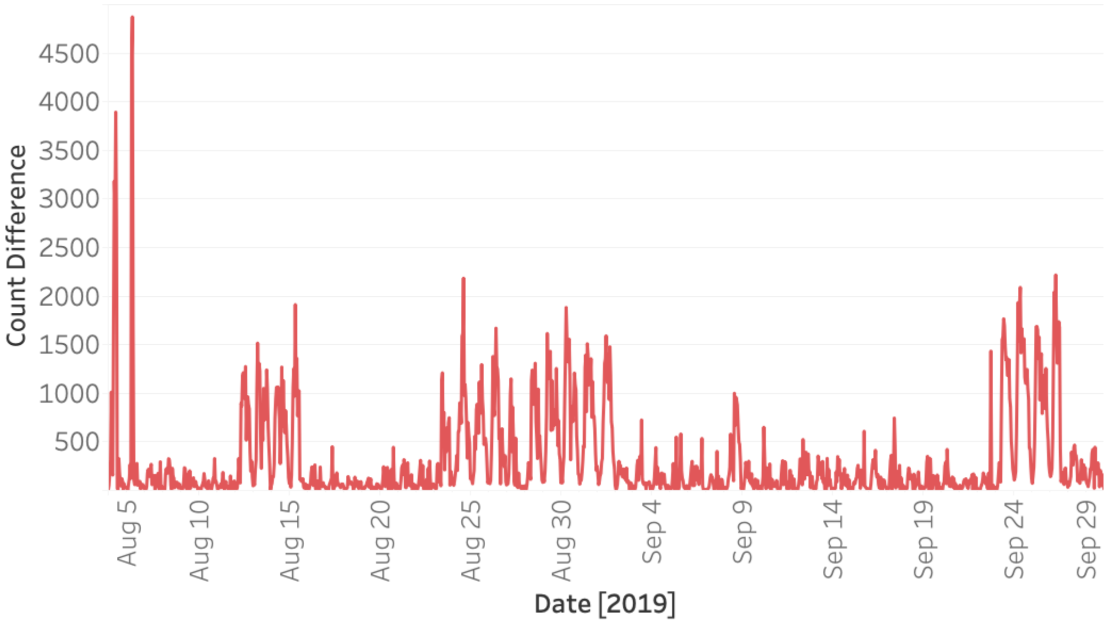


Figure 4.6: Collected Count Difference

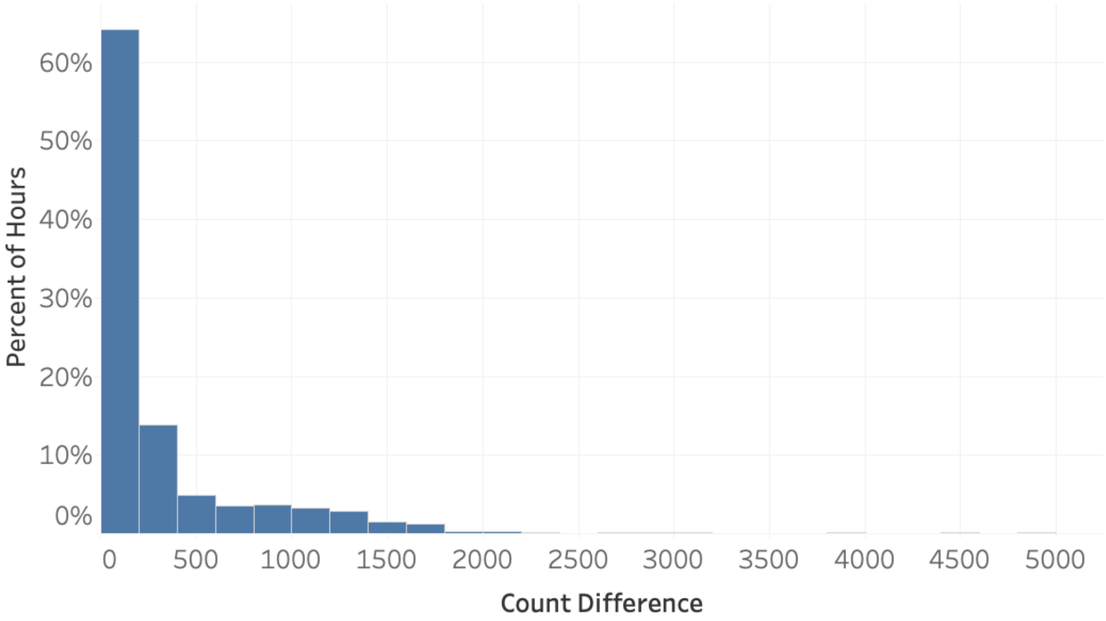


Figure 4.7: Histogram of Differences Between Collected Counts

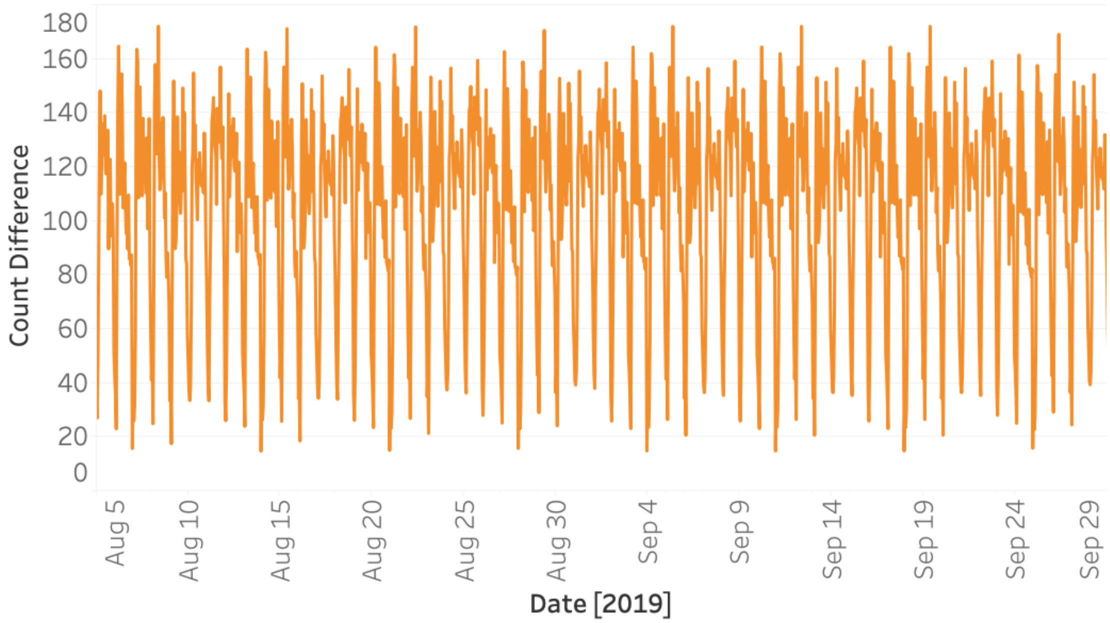


Figure 4.8: Predicted Count Difference

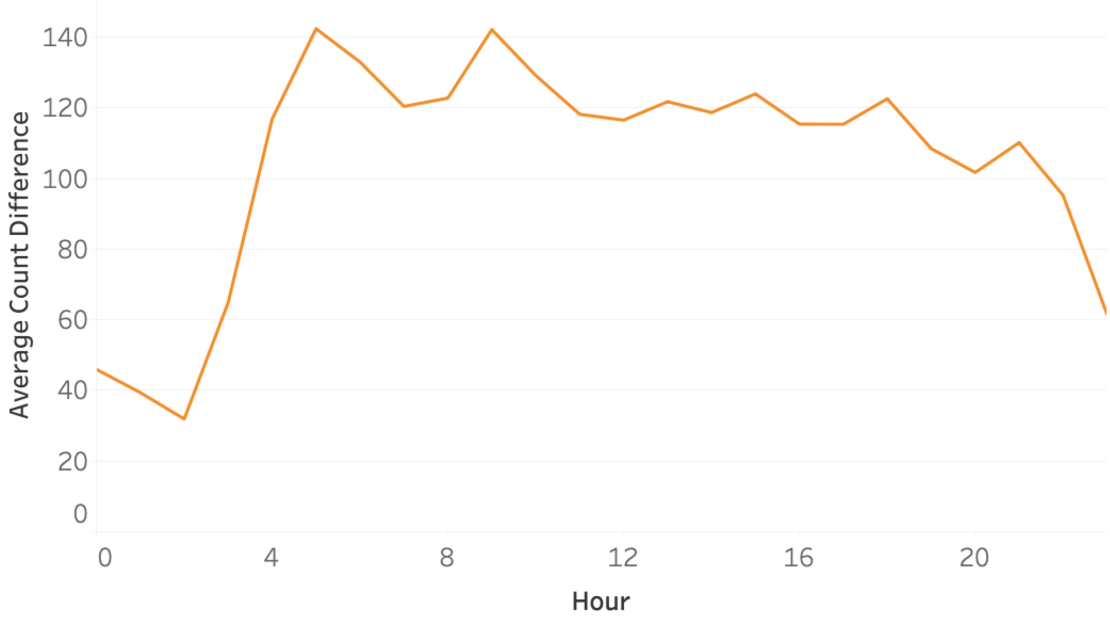


Figure 4.9: Predicted Count Difference by Hour

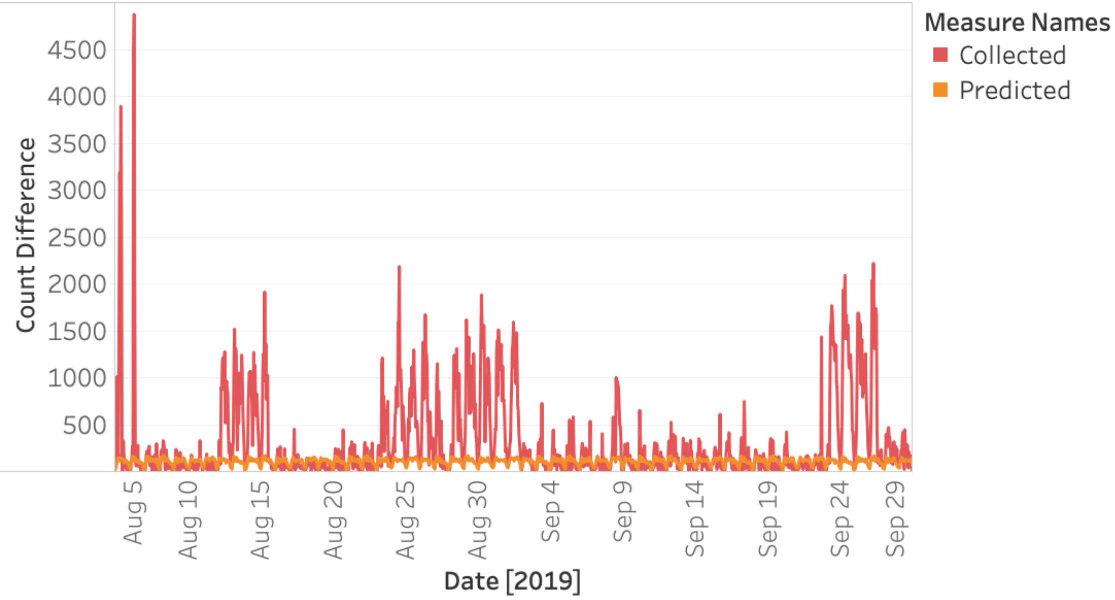


Figure 4.10: Collected and Predicted Count Difference

be necessary [89, 90].

4.4.3 Possible Sources of Anomalous Days

With different methods of data collection producing the two data sources, other explanations for the differences in counts were considered before assuming an equipment malfunction. For example, weather effects can be a source of data discrepancies. However, after reviewing the historical data from the National Weather Service [99] shown in Figure 4.12, no apparent link exists.

Consulting with ODOT engineers revealed additional possible sources of anomalies. The differing collection methods can experience distinct impacts. For example, a large vehicle could block a smaller vehicle from the radar, but such is explicitly accommodated in the framework during the training phase. Events not present in the training data, such as a truckers' convention leading to an atypical representation of large vehicles occluding smaller vehicles, could occur that impact data collection for a period of time. These unforeseen events may present as anomalies and would require an SME to make the

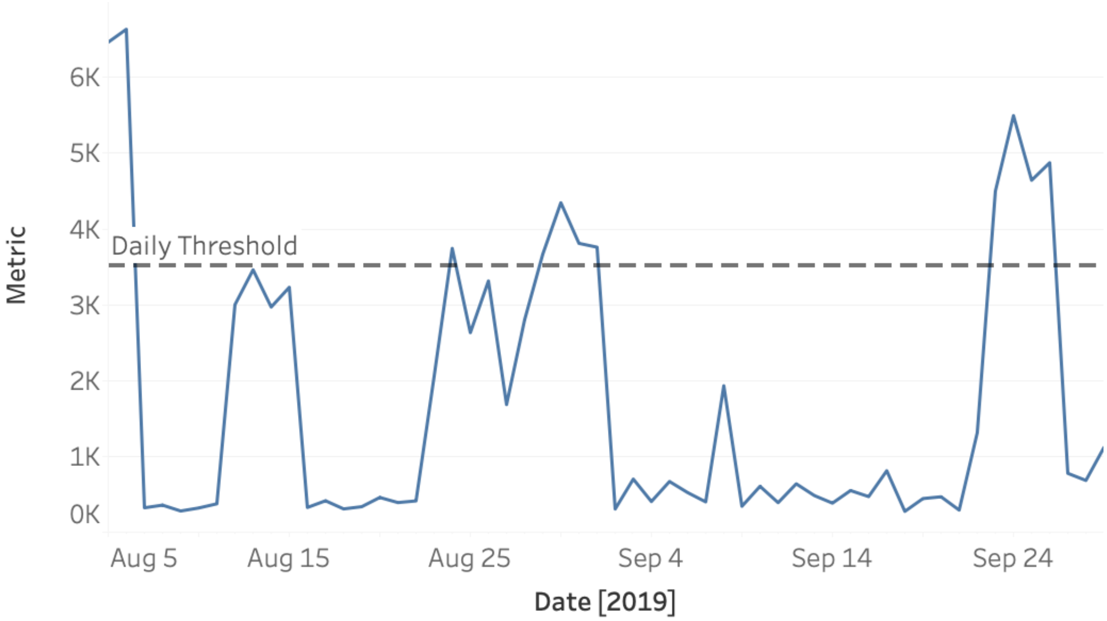


Figure 4.11: Threshold Evaluation

association.

Communication issues appear to be the most likely source of the anomalous days. Although a physical fiber connects the collection stations, some polling intervals are missing data. Additionally, it does not appear that the absent counts were included in later polling periods, as demonstrated by the anomaly being raised due to count differences.

ODOT has a fault detection mechanism in place whereupon an alert is generated after six missed polling events. Many of the missed events observed were intermittent, so they would not have triggered the currently configured fault detection mechanism. One option would be to lower the threshold for an alert to be generated. However, the potential increase in the volume of alerts for missed polling events that are not impactful over the system the size of ODOT's could cloud problems deserving of further investigation. Looking at trends rather than individual events is a primary strength of the framework proposed here. By alerting to periods anomalous to the trained norm, this framework notifies monitoring agencies of equipment issues that may occur in a manner not to be concerning on their own, but in aggregate, represent impactful effects on the data.

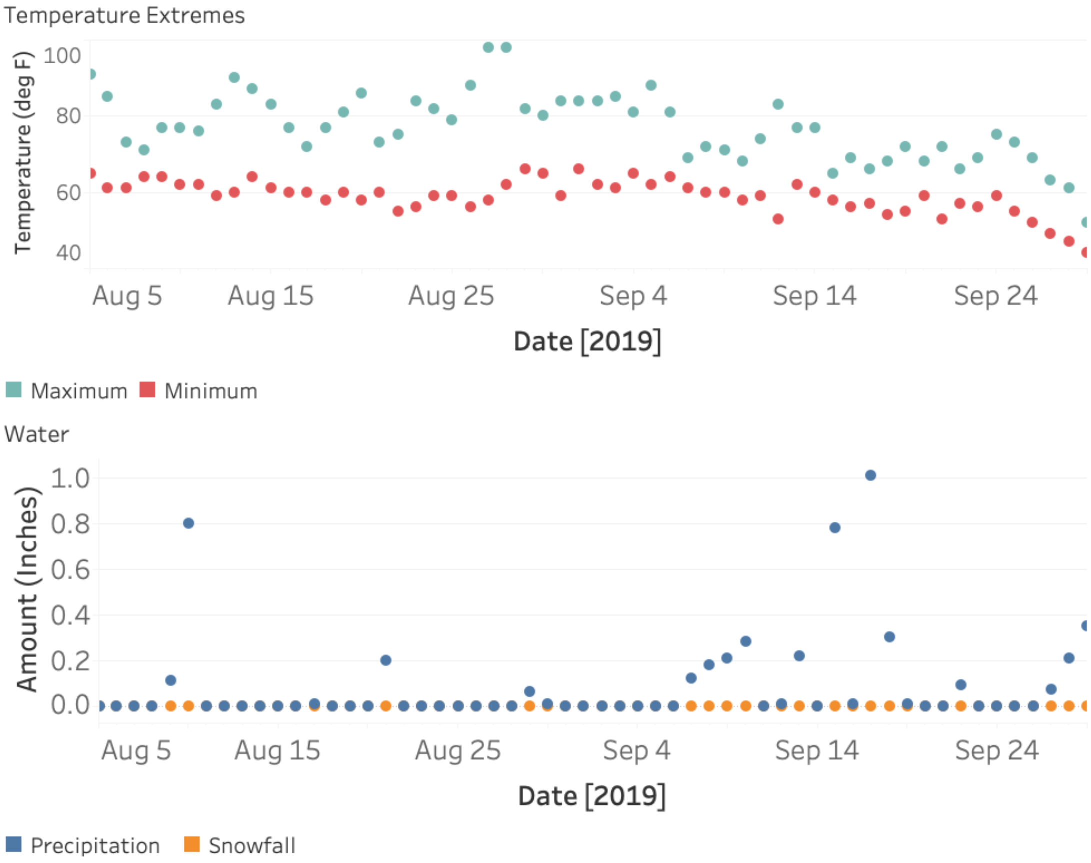


Figure 4.12: Portland, Oregon, Airport Weather [99]

4.5 Conclusion

This research presents a traffic data anomaly detection framework. Unique to this framework is the use of two or more independent traffic data sources as inputs and the method to process the data to indicate anomalous periods of time. A case study that considers data from ODOT was conducted using two collection methods sampling the same traffic population. Of the 56 days processed in the case study, 11 qualified as anomalous. In consultation with ODOT engineers, intermittent missing data resulting in loss of traffic counts was attributed to communication problems. The fault detection mechanism ODOT currently employs was likely not triggered due to the intermittent nature of the missing data.

The case study demonstrates the need for frameworks similar to that proposed here to be implemented to identify anomalous data. Current systems that rely on abrupt disruptions may miss intermittent issues that compound to degrade data quality severely. Using two or more independent data sources and training a system to recognize the nominal sampling relationship between the two or more data sources, transportation agencies will be able to identify anomalous data patterns that can impact decision-making processes and distinguish between actual changes in traffic and equipment malfunctions.

4.5.1 Future Work

This work was originally developed for ridership anomaly detection in public transit. Research by Nasir, Gharib, and Jaafar [100], Yang and Wu [16], Dunlap, Li, Henrickson, *et al.* [101], and others have shown the potential of new sources of ridership data. With inaccuracies acknowledged in many automated passenger counting (APC) systems [102, 103], transit agencies could experience increased confidence in the quality of their ridership data using the newly developed anomaly detection methods. Work showing the potential for a public transit implementation was presented at the Transportation Research Board's 2019 Conference on Performance and Data in Transportation Decision Making [88].

Implementing this method in a public transit environment for ridership requires different considerations than what is necessary for traffic data. Depending on the respective data sources, the spatial and temporal rectification, as conducted in Section 4.3.1, may require additional analysis to ensure the same representations are being compared. For example, akin to vehicle counts in traffic data, boarding and alighting data represent ridership passing a defined threshold. While much public transit ridership data is collected as boardings and alightings, some, as in the work of Nasir, Gharib, and Jaafar [100], represents the current passenger load. Reconciliation would be required to facilitate the use of boarding and alighting data alongside passenger load data.

Public transit vehicles also present a moving rather than fixed point of collection. This difference manifests when seeking to compare fixed-point counts, as might be seen from passively collected WiFi packets [16], and on-vehicle counts from sources like APC or Automated Fare Collection (AFC) systems. Having both mobile and multiple fixed counts would enable an additional benefit not possible with two fixed independent sources. With a network being trained for comparing each fixed data source with the mobile data source, rather than an anomaly being flagged for an unknown shift in the comparative data of two sources, the anomaly source can be isolated as either the mobile source, if an anomaly is identified in all pairwise comparison, or a specific fixed source if an anomaly is identified in just one comparison. This anomaly source isolation would hasten a resolution to potentially malfunctioning equipment by directly identifying the source of the change in data collection performance.

While such data sources exist to enact the proposed work, currently, access to such data proved infeasible. Privacy concerns surrounding potentially individually-identifiable AFC data restrict the availability of the data for research. Production environment concerns of large WiFi network deployments make the additional system demands of more granular packet data a barrier to such data availability. The privatization of call detail records makes their use prohibitively expensive for research use. Additional sources of mobility data continue to be developed, and each new source has potential for comparison against existing sources.

Chapter 5: Conclusion

Transportation planners have used modeling to fill gaps in empirical data about how people use their transportation networks. In this research, several frameworks were developed that employ data already being collected to facilitate the analysis and improve the utilization of transportation networks. Techniques like evolutionary optimization heuristics and long short-term memory (LSTM) networks are at the core of these frameworks, resulting in improvements to transportation network metrics and helping planners gain confidence in the quality of data they use.

The first demonstration of the work is presented in Chapter 2 and introduces two new metrics as potential objectives for finding solutions to a type of Urban Transit Routing Problem (UTRP). Unlike prior UTRP research, a framework is developed that serves to improve the social experience of riders through route design. While improvements such as service frequency and extended hours can increase ridership, they come with ongoing added costs, which are not readily recouped by a ridership increase. The route solutions presented by the socially-aware framework do not inherently have costs associated with their implementation beyond what agencies may normally consider with routine route reassignments. The work is presented as an introduction to expanding the scope of what a UTRP framework can include when evaluating potential solutions.

After initially developing new approaches to the UTRP, additional metrics were developed, and the framework expanded to a multi-objective heuristic optimization. Responding to global events, Chapter 3 presented a framework to produce solutions that allow transit planners to balance the need to reduce the susceptibility of disease transmission in their transit vehicles while maintaining transit network utility for potential riders. This work further moves in the direction where UTRP solutions balance transit demands that include not only coverage and efficiency but also areas like rider experience, public health impacts, equity, and others.

Developing better uses for existing data in transportation extends beyond mass transit environments. Organizations responsible for roadways collect data about traffic volume to support operational and planning needs. With limited continual validation, changes in equipment behavior may be incorrectly attributed to a change in traffic volume rather than an equipment malfunction. Chapter 4 develops an anomaly detection framework that leverages redundancies in sampling populations that will arise as additional sources of data are identified. The framework uses an LSTM network to model the nominal sampling relationship between data sources. When the collected data from one or more sources deviate from the nominal model, the data is marked as potentially anomalous. Data provided to conduct a case study happened to demonstrate just such a case where one source would occasionally collect approximately 11% fewer vehicles than the alternative source. The anomaly was attributed to likely communication issues and had not been previously identified because the fault was too intermittent to be detected by the existing anomaly detection mechanism.

The three chapters developed modeling approaches and solution methods and demonstrated their employment to show new potential to apply data to make better transportation planning decisions. Each research effort faced limitations surround access to data and computational resources. While constraining some potential development, such limitations can be acknowledged as issues regularly faced by transportation organizations. By developing methods that can both function and grow under such constraints, the proposed work maintains accessibility and feasibility for a wide breadth of transportation organizations.

5.1 Future Work

Chapter 2, Chapter 3, and Chapter 4 individually outline specific opportunities for future work. Overarching all of the chapters is the expanding availability of data that may represent aspects of human mobility. Companies like StreetLight Data [104] and Replica [105] have recognized the value of emerging data sources and have monetized their collection and analysis. Other companies, like Remix [106], acknowledge the gap between predominant skills in organizations that operate transportation systems and what resources are needed to leverage emerging data sources.

Resources matter. Be it private companies developing projects, or public agencies focusing on hiring and making investments to evolve their data capabilities, resources need to be dedicated to enable transportation planners to make data-informed decisions. For example, in Chapter 4, a systemic evaluation and validation of the method proposed proved beyond a feasible scope. Similar to what Alam, Gerostathopoulos, Amini, *et al.* [89] experienced, labeled traffic anomaly data is all but non-existent, and the measure of what makes traffic data anomalous is often subjectively defined. The experience of executing what became Chapter 4 was one of having the framework identify anomalies followed by a lengthy process of acknowledgment that identified periods of anomalous data that were indeed anomalous. To make progress in refining the current framework or evaluating performance comparing other approaches, more data and ample labeling (either discrete or continuous) will be necessary; both require significant resources to be dedicated.

In the introduction, presented are the worsening trends of the United States transportation systems. The subsequent work presented in this dissertation and the work of companies mentioned above are striving towards a data-centric and engineering-centric solution, but, as mentioned, resources are central to making progress. In transportation, those resources primarily come from political entities, and those entities are driven by policy. In pursuing these topics, what appeared often was the unclear boundary between engineering and policy. For example, there is often a conflict in public transit between maximizing ridership and providing service to communities assessed to be most in need, a question that would require a definitive decision when implementing a multi-objective UTRP. With improved traffic data, is the result of increasing volume cause for expanding roadways? Or is it better to implement congestion pricing strategies that can leverage the improved data to manage traffic volumes dynamically? Do rising hours of commuter delay mean more road capacity is inevitably needed? Or that alternatives to personal vehicles and restrictive zoning that pushes places of employment further away from where workers live need to be reevaluated? Answers to such questions stray outside of a purely technical engineering response but are central to how, if any, benefits are realized from emerging data.

Where does that leave future work to further the problem presented in the introduction (i.e., the need to improve transportation system planning to meet demand efficiently)? It is easy to look to European equivalents that have achieved outcomes desirable in the United States context and hope that a direct translation would achieve similar results. However, cultural and structural differences between how Europe developed and how United States cities developed suggest caution in making any assumptions about the success of any implementations without significant adaptation. What is needed from both further work in engineering and policy are data-backed solutions tailored to the realities of how the United States developed.

Bibliography

[1] JL Toole, S Colak, B Sturt, LP Alexander, A Evsukoff, and MC González, “The path most traveled: Travel demand estimation using big data resources”, *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 162–177, Sep. 2015, ISSN: 0968090X. DOI: 10.1016/j.trc.2015.04.022. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0968090X15001631 (visited on 07/25/2018).

[2] David Schrank, Bill Eisele, and Tim Lomax, “2019 Urban Mobility Report”, Texas A&M Transportation Institute, Tech. Rep., Aug. 2019. [Online]. Available: https://mobility.tamu.edu/umr/report/.

[3] Jeff Davis, *Trends in Per Capita VMT*, Jun. 2019. [Online]. Available: https://www.enotrans.org/article/trends-in-per-capita-vmt/ (visited on 08/24/2020).

[4] N Bogel-Burroughs, “Deadliest Year for Pedestrians and Cyclists in U.S. Since 1990”, *The New York Times*, Oct. 2019, ISSN: 0362-4331. [Online]. Available: https://www.nytimes.com/2019/10/22/us/pedestrian-cyclist-deaths-traffic.html (visited on 08/24/2020).

[5] Pew Research Center, *Demographics of Mobile Device Ownership and Adoption in the United States*, Jun. 2019. [Online]. Available: https://www.pewresearch.org/internet/fact-sheet/mobile/ (visited on 08/20/2020).

[6] P Ryus, E Ferguson, KM Laustsen, RJ Schneider, FR Proulx, T Hull, L Miranda-Moreno, National Cooperative Highway Research Program, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine, *Guidebook on Pedestrian and Bicycle Volume Data Collection*. Washington, D.C.: Transportation Research Board, Jan. 2014, ISBN: 978-0-309-30826-7. DOI: 10.17226/22223. [Online]. Available: https://www.nap.edu/catalog/22223 (visited on 02/19/2020).

[7] P Ryus, E Ferguson, KM Laustsen, RJ Schneider, FR Proulx, T Hull, and L Miranda-Moreno, *Methods and Technologies for Pedestrian and Bicycle Volume Data Collection*. Washington, D.C.: Transportation Research Board, Jan. 2014, ISBN: 978-0-309-37517-7. DOI: 10.17226/23429. [Online]. Available: https://www.nap.edu/catalog/23429 (visited on 02/19/2020).

[8] P Ryus, A Butsick, FR Proulx, RJ Schneider, and T Hull, *Methods and Technologies for Pedestrian and Bicycle Volume Data Collection: Phase 2*. Washington, D.C.: Transportation Research Board, Mar. 2017, ISBN: 978-0-309-45737-8. DOI: 10.17226/24732. [Online]. Available: https://www.nap.edu/catalog/24732 (visited on 02/19/2020).

[9] K Nordback, S Kothuri, T Phillips, C Gorecki, and M Figliozzi, "Accuracy of Bicycle Counting with Pneumatic Tubes in Oregon", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2593, no. 1, pp. 8–17, Jan. 2016, ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2593-02. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2593-02 (visited on 02/19/2020).

[10] G Griffin, K Nordback, T Götschi, E Stolz, S Kothuri, Technical Activities Division, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine, *Monitoring Bicyclist and Pedestrian Travel and Behavior*. Washington, D.C.: Transportation Research Board, Mar. 2014, ISBN: 978-0-309-43366-2. DOI: 10.17226/22420. [Online]. Available: https://www.nap.edu/catalog/22420 (visited on 02/19/2020).

[11] S Kothuri, K Nordback, A Schrope, T Phillips, and M Figliozzi, "Bicycle and Pedestrian Counts at Signalized Intersections Using Existing Infrastructure: Opportunities and Challenges", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2644, no. 1, pp. 11–18, Jan. 2017, ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2644-02. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2644-02 (visited on 02/19/2020).

[12] B Namaki Araghi, R Krishnan, and H Lahrmann, "Mode-Specific Travel Time Estimation Using Bluetooth Technology", *Journal of Intelligent Transportation Systems*, vol. 20, no. 3, pp. 219–228, May 2016, ISSN: 1547-2450, 1547-2442. DOI: 10.1080/15472450.2015.1052906. [Online]. Available: https://www.

tandfonline.com/doi/full/10.1080/15472450.2015.1052906 (visited on 02/19/2020).

[13] A Kurkcu and K Ozbay, "Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2644, no. 1, pp. 72–82, Jan. 2017, ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2644-09. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2644-09 (visited on 02/19/2020).

[14] EO Ryeng, T Haugen, H Grønlund, and SB Overå, "Evaluating Bluetooth and Wi-Fi Sensors as a Tool for Collecting Bicycle Speed at Varying Gradients", *Transportation Research Procedia*, vol. 14, pp. 2289–2296, 2016, ISSN: 23521465. DOI: 10.1016/j.trpro.2016.05.245. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352146516302514 (visited on 02/19/2020).

[15] L Schauer and M Werner, "Analyzing Pedestrian Flows Based on Wi-Fi and Bluetooth Captures", *ICST Transactions on Ubiquitous Environments*, vol. 1, no. 4, e4, May 2015, ISSN: 2032-9377. DOI: 10.4108/ue.1.4.e4. [Online]. Available: http://eudl.eu/doi/10.4108/ue.1.4.e4 (visited on 02/19/2020).

[16] S Yang and YJ Wu, "Travel mode identification using bluetooth technology", *Journal of Intelligent Transportation Systems*, pp. 1–15, Sep. 2017, ISSN: 1547-2450. DOI: 10.1080/15472450.2017.1384698. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15472450.2017.1384698 (visited on 07/25/2018).

[17] Bloom Communications, "Portland Bureau of Transportation Final Research Report & Recommendations", Tech. Rep., 2018.

[18] Matthew Dickens, "Public Transportation Ridership Report", American Public Transportation Association, Tech. Rep., May 2020. [Online]. Available: https://www.apta.com/research-technical-resources/transit-statistics/ridership-report/.

[19] K Watkins, S Berrebi, C Diffee, B Kiriazes, and D Ederer, "Analysis of Recent Public Transit Ridership Trends", TCRP Research Report 209, 2020, p. 111.

[20] S Berrebi, S Joshi, T Gibbs, and K Watkins, *Modeling Bus Ridership Trends on a Hyper-Local Scale Between 2012 and 2017*, Conference on Performance and Data in Transportation Decision Making, Atlanta, Georgia, Sep. 2019.

[21] X Wang, DA Rodríguez, OL Sarmiento, and O Guaje, "Commute patterns and depression: Evidence from eleven Latin American cities", *Journal of Transport & Health*, vol. 14, p. 100 607, Sep. 2019, ISSN: 22141405. DOI: 10.1016/j.jth.2019.100607. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214140518306169 (visited on 11/02/2020).

[22] J Vallée, E Cadot, C Roustit, I Parizot, and P Chauvin, "The role of daily mobility in mental health inequalities: The interactive influence of activity space and neighbourhood of residence on depression", *Social Science & Medicine*, vol. 73, no. 8, pp. 1133–1144, Oct. 2011, ISSN: 02779536. DOI: 10.1016/j.socscimed.2011.08.009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0277953611004977 (visited on 11/02/2020).

[23] T Litman, "A New Transit Safety Narrative", *Journal of Public Transportation*, vol. 17, no. 4, pp. 114–135, Dec. 2014, ISSN: 1077-291X, 2375-0901. DOI: 10.5038/2375-0901.17.4.7. [Online]. Available: http://scholarcommons.usf.edu/jpt/vol17/iss4/8/ (visited on 02/19/2020).

[24] E Paulos and E Goodman, "The familiar stranger: Anxiety, comfort, and play in public places", in *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, Vienna, Austria: ACM Press, 2004, pp. 223–230, ISBN: 978-1-58113-702-6. DOI: 10.1145/985692.985721. [Online]. Available: http://portal.acm.org/citation.cfm?doid=985692.985721 (visited on 11/16/2020).

[25] V Guihaire and JK Hao, "Transit network design and scheduling: A global review", *Transportation Research Part A: Policy and Practice*, vol. 42, no. 10, pp. 1251–1273, 2008, ISSN: 0965-8564. DOI: 10.1016/j.tra.2008.03.011.

[26] Y Xiong and JB Schneider, "Transportation network design using a cumulative genetic algorithm and neural network", *Transportation Research Record: Journal of the Transportation Research Board*, vol. p. 37-44, 1992.

[27] CL Mumford, “New Heuristic and Evolutionary Operators for the Multi-Objective Urban Transit Routing Problem”, *2013 IEEE Congress on Evolutionary Computation*, pp. 939–946, 2013. DOI: 10.1109/cec.2013.6557668.

[28] A Mahmoudzadeh and XB Wang, “Cluster Based Methodology for Scheduling a University Shuttle System”, *Transportation Research Record: Journal of the Transportation Research Board*, 2020, ISSN: 0361-1981. DOI: 10.1177/0361198119900636.

[29] D Bertsimas, A Delarue, and S Martin, “Optimizing schools’ start time and bus routes”, *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, p. 201 811 462, 2019, ISSN: 0027-8424. DOI: 10.1073/pnas.1811462116.

[30] JY Lemos, AR Joshi, MM D’souza, and AD D’souza, “Design a Smart and Intelligent Routing Network using Optimization Techniques”, pp. 297–310, 2018, ISSN: 2194-5357. DOI: 10.1007/978-981-13-1274-8_23.

[31] K Leksakul, U Smutkupt, R Jintawiwat, and S Phongmoo, “Heuristic approach for solving employee bus routes in a large-scale industrial factory”, *Advanced Engineering Informatics*, vol. 32, pp. 176–187, 2017, ISSN: 1474-0346. DOI: 10.1016/j.aei.2017.02.006.

[32] Y Liu, C Liu, NJ Yuan, L Duan, Y Fu, H Xiong, S Xu, and J Wu, “Intelligent bus routing with heterogeneous human mobility patterns”, *Knowledge and Information Systems*, vol. 50, no. 2, pp. 383–415, 2017, ISSN: 0219-1377. DOI: 10.1007/s10115-016-0948-6.

[33] C Iliopoulou, K Kepaptsoglou, and E Vlahogianni, “Metaheuristics for the transit route network design problem: A review and comparative analysis”, *Public Transport*, vol. 11, no. 3, pp. 487–521, Oct. 2019, ISSN: 1866-749X, 1613-7159. DOI: 10.1007/s12469-019-00211-2. [Online]. Available: http://link.springer.com/10.1007/s12469-019-00211-2 (visited on 04/24/2020).

[34] K Kepaptsoglou and M Karlaftis, “Transit Route Network Design Problem: Review”, *Journal of Transportation Engineering*, vol. 135, no. 8, pp. 491–505, Aug. 2009, ISSN: 0733-947X, 1943-5436. DOI: 10.1061/(ASCE)0733-947X(2009)135:8(491). [Online]. Available: http://ascelibrary.org/doi/10.1061/%28ASCE%290733-947X%282009%29135%3A8%28491%29 (visited on 04/24/2020).

[35] S Milgram, *The individual in a social world: essays and experiments*, ser. Addison-Wesley series in social psychology. Reading, Mass: Addison-Wesley Pub. Co, 1977, ISBN: 978-0-201-04382-2.

[36] D Liang, X Li, and YQ Zhang, "Identifying familiar strangers in human encounter networks", *EPL (Europhysics Letters)*, vol. 116, no. 1, p. 18 006, 2016, ISSN: 0295-5075. DOI: 10.1209/0295-5075/116/18006.

[37] K Liu, L Yin, Z Ma, F Zhang, and J Zhao, "Investigating physical encounters of individuals in urban metro systems with large-scale smart card data", *Physica A: Statistical Mechanics and its Applications*, 2019, ISSN: 0378-4371. DOI: 10.1016/j.physa.2019.123398.

[38] L Sun, KW Axhausen, DH Lee, and M Cebrian, "Efficient detection of contagious outbreaks in massive metropolitan encounter networks", *Scientific Reports*, vol. 4, no. 1, p. 5099, 2014. DOI: 10.1038/srep05099.

[39] K Asatani, F Toriumi, J Mori, M Ochi, and I Sakata, "Detecting interpersonal relationships in large-scale railway trip data", *Journal of Computational Social Science*, vol. 1, no. 2, pp. 313–326, 2018, ISSN: 2432-2717. DOI: 10.1007/s42001-018-0021-1.

[40] Transit, *Finding The Hidden Social Networks of Public Transit: Montreal's Familiar Strangers*. 2015. [Online]. Available: https://medium.com/transit-app/montreal-s-familiar-strangers-6b224f3cc9c3.

[41] F Zhang, B Jin, T Ge, Q Ji, and Y Cui, "Who are My Familiar Strangers?: Revealing Hidden Friend Relations and Common Interests from Smart Card Data", pp. 619–628, 2016. DOI: 10.1145/2983323.2983804.

[42] L Sun, KW Axhausen, DH Lee, and X Huang, "Understanding metropolitan patterns of daily encounters", *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, pp. 13 774–13 779, 2013, ISSN: 0027-8424. DOI: 10.1073/pnas.1306440110.

[43] DJ Watts and SH Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, vol. 393, no. 6684, pp. 440–442, 1998, ISSN: 0028-0836. DOI: 10.1038/30918.

[44] N Agarwal, H Liu, S Murthy, A Sen, and X Wang, “A Social Identity Approach to Identify Familiar Strangers in a Social Network”, in *Proceedings of the Third International ICWSM Conference*, 2009. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/184/565.

[45] *WiGLE: Wireless Network Mapping*. [Online]. Available: https://wigle.net/.

[46] G Boeing, “OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks”, *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, Sep. 2017, ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2017.05.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0198971516303970 (visited on 06/11/2019).

[47] Y Fan, T Ormsby, P Wiringa, CF Liao, and J Wolfson, “Visualizing Transportation Happiness in the Minneapolis-St. Paul Region”, Tech. Rep. CTS 20-05, Mar. 2020, p. 13.

[48] Dask Development Team, *Dask: Library for dynamic task scheduling*. 2016. [Online]. Available: https://dask.org.

[49] JP Onnela, J Saramäki, J Kertész, and K Kaski, “Intensity and coherence of motifs in weighted complex networks”, *Physical Review E*, vol. 71, no. 6, p. 065 103, Jun. 2005, ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.71.065103. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.71.065103 (visited on 05/08/2020).

[50] G Fagiolo, “Clustering in complex directed networks”, *Physical Review E*, vol. 76, no. 2, p. 026 107, Aug. 2007, ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.76.026107. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.76.026107 (visited on 03/01/2020).

[51] AA Hagberg, DA Schult, and PJ Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX”, in *Proceedings of the 7th Python in Science Conference*, G Varoquaux, T Vaught, and J Millman, Eds., Pasadena, CA USA, 2008, pp. 11–15.

[52] TP Peixoto, *The graph-tool python library*, 2017. DOI: 10.6084/M9.FIGSHARE.1164194. [Online]. Available: https://figshare.com/articles/graph_tool/1164194 (visited on 03/01/2020).

[53] L Fan, CL Mumford, and D Evans, “A simple multi-objective optimization algorithm for the urban transit routing problem”, in *2009 IEEE Congress on Evolutionary Computation*, Trondheim, Norway: IEEE, May 2009, pp. 1–7. DOI: 10.1109/CEC.2009.4982923. [Online]. Available: http://ieeexplore.ieee.org/document/4982923/ (visited on 02/19/2020).

[54] JSC Chew, LS Lee, and HV Seow, “Genetic Algorithm for Biobjective Urban Transit Routing Problem”, *Journal of Applied Mathematics*, vol. 2013, pp. 1–15, 2013, ISSN: 1110-757X. DOI: 10.1155/2013/698645.

[55] MP John, CL Mumford, and R Lewis, “An Improved Multi-Objective Algorithm for the Urban Transit Routing Problem”, pp. 49–60, 2014, ISSN: 0302-9743. DOI: 10.1007/978-3-662-44320-0_5.

[56] IM Cooper, MP John, R Lewis, CL Mumford, and A Olden, “Optimising Large Scale Public Transport Network Design Problems using Mixed-Mode Parallel Multi-Objective Evolutionary Algorithms”, *2014 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2841–2848, 2014. DOI: 10.1109/cec.2014.6900362.

[57] B Manderick and P Spiessens, “Fine-Grained Parallel Genetic Algorithms”, in *ICGA 1989: 428-433*, 1989.

[58] TC Belding, “The distributed genetic algorithm revisited”, in *Proceedings of the 6th International Conference on Genetic Algorithms*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 114–121, ISBN: 1558603700.

[59] M Ruciński, D Izzo, and F Biscani, “On the impact of the migration topology on the Island Model”, *Parallel Computing*, vol. 36, no. 10-11, pp. 555–571, 2010, ISSN: 0167-8191. DOI: 10.1016/j.parco.2010.04.002.

[60] F Biscani and D Izzo, “A parallel global multiobjective framework for optimization: Pagmo”, *Journal of Open Source Software*, vol. 5, no. 53, p. 2338, 2020. DOI: 10.21105/joss.02338. [Online]. Available: https://doi.org/10.21105/joss.02338.

[61] M Mernik, SH Liu, D Karaboga, and M Črepinšek, “On clarifying misconceptions when comparing variants of the Artificial Bee Colony Algorithm by offering a new implementation”, *Information Sciences*, vol. 291, pp. 115–127, Jan. 2015, ISSN: 00200255. DOI: 10.1016/j.ins.2014.08.040. [Online]. Available: https:

//linkinghub.elsevier.com/retrieve/pii/S0020025514008378 (visited on 03/05/2020).

[62] M Mahdavi, M Fesanghary, and E Damangir, “An improved harmony search algorithm for solving optimization problems”, *Applied Mathematics and Computation*, vol. 188, no. 2, pp. 1567–1579, May 2007, ISSN: 0096-3003. DOI: 10.1016/j.amc.2006.11.033. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0096300306015098 (visited on 04/08/2020).

[63] *PyGMO List of algorithms — pygmo 2.14.0 documentation*. [Online]. Available: https://esa.github.io/pygmo2/algorithms.html%5C (visited on 03/05/2020).

[64] AL Barabási and M Pósfai, *Network science*. Cambridge, United Kingdom: Cambridge University Press, 2016, ISBN: 978-1-107-07626-6.

[65] Z Zhang, H Wang, C Wang, and H Fang, “Modeling Epidemics Spreading on Social Contact Networks”, *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 410–419, Sep. 2015, ISSN: 2168-6750. DOI: 10.1109/TETC.2015.2398353. [Online]. Available: http://ieeexplore.ieee.org/document/7029011/ (visited on 06/30/2020).

[66] AD Broido and A Clauset, “Scale-free networks are rare”, *Nature Communications*, vol. 10, no. 1, p. 1017, Dec. 2019, ISSN: 2041-1723. DOI: 10.1038/s41467-019-08746-5. [Online]. Available: http://www.nature.com/articles/s41467-019-08746-5 (visited on 12/29/2020).

[67] S Hoover, JD Porter, and C Fuentes, “Strategic Route Planning to Manage Transit’s Susceptibility to Disease Transmission”, *Transportation Research Record: Journal of the Transportation Research Board*, p. 036 119 812 199 781, Mar. 2021, ISSN: 0361-1981, 2169-4052. DOI: 10.1177/0361198121997815. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0361198121997815 (visited on 05/11/2021).

[68] Max Larkin, *2 MIT Engineers Use Math To Plot A Path For Boston’s School Buses*, Jul. 2017. [Online]. Available: https://www.wbur.org/edify/2017/07/27/mit-quantum-boston-bus-routes.

[69] Joi Ito, *What the Boston School Bus Schedule Can Teach Us About AI*, May 2018. [Online]. Available: https://www.wired.com/story/joi-ito-ai-and-bus-routes/.

[70] David Grossman, *How an Algorithm Made the Buses in Boston Better*, Aug. 2019. [Online]. Available: https://www.popularmechanics.com/technology/infrastructure/a28689713/algorithm-boston-buses/.

[71] JY Lemos, AR Joshi, MM D'souza, and AD D'souza, "Design a Smart and Intelligent Routing Network Using Optimization Techniques", in *Data Management, Analytics and Innovation*, VE Balas, N Sharma, and A Chakrabarti, Eds., Singapore: Springer Singapore, 2019, pp. 297–310.

[72] MJ Keeling and KT Eames, "Networks and epidemic models", *Journal of The Royal Society Interface*, vol. 2, no. 4, pp. 295–307, Sep. 2005, ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2005.0051. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsif.2005.0051 (visited on 06/30/2020).

[73] SM Firestone, MP Ward, RM Christley, and NK Dhand, "The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis", *Preventive Veterinary Medicine*, vol. 102, no. 3, pp. 185–195, Dec. 2011, ISSN: 01675877. DOI: 10.1016/j.prevetmed.2011.07.006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167587711002327 (visited on 07/13/2020).

[74] F Xu, C Connell McCluskey, and R Cressman, "Spatial spread of an epidemic through public transportation systems with a hub", *Mathematical Biosciences*, vol. 246, no. 1, pp. 164–175, Nov. 2013, ISSN: 00255564. DOI: 10.1016/j.mbs.2013.08.014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0025556413002162 (visited on 06/30/2020).

[75] L Sun, KW Axhausen, DH Lee, and M Cebrian, "Efficient detection of contagious outbreaks in massive metropolitan encounter networks", *Scientific Reports*, vol. 4, no. 1, p. 5099, May 2015, ISSN: 2045-2322. DOI: 10.1038/srep05099. [Online]. Available: http://www.nature.com/articles/srep05099 (visited on 06/30/2020).

[76] Bo Song, Yu-Rong Song, and Guo-Ping Jiang, “How clustering affects epidemics in complex networks”, in *2017 International Conference on Computing, Networking and Communications (ICNC)*, Silicon Valley, CA, USA: IEEE, Jan. 2017, pp. 178–183, ISBN: 978-1-5090-4588-4. DOI: 10.1109/ICCNC.2017.7876123. [Online]. Available: http://ieeexplore.ieee.org/document/7876123/ (visited on 06/30/2020).

[77] National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases, *What Bus Transit Operators Need to Know About COVID-19*, Apr. 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/community/organizations/bus-transit-operator.html (visited on 06/19/2020).

[78] L Goscé and A Johansson, “Analysing the link between public transport use and airborne transmission: Mobility and contagion in the London underground”, *Environmental Health*, vol. 17, no. 1, p. 84, Dec. 2018, ISSN: 1476-069X. DOI: 10.1186/s12940-018-0427-5. [Online]. Available: https://ehjournal.biomedcentral.com/articles/10.1186/s12940-018-0427-5 (visited on 06/30/2020).

[79] A Bota, LM Gardner, and A Khani, “Modeling the spread of infection in public transit networks: A decision-support tool for outbreak planning and control”, in *TRB 96th Annual Meeting Compendium of Papers*, Washington DC, United States, Jan. 2017.

[80] A Bóta, LM Gardner, and A Khani, “Identifying Critical Components of a Public Transit System for Outbreak Control”, *Networks and Spatial Economics*, vol. 17, no. 4, pp. 1137–1159, Dec. 2017, ISSN: 1566-113X, 1572-9427. DOI: 10.1007/s11067-017-9361-2. [Online]. Available: http://link.springer.com/10.1007/s11067-017-9361-2 (visited on 06/30/2020).

[81] AE Shoghri, J Liebig, L Gardner, R Jurdak, and S Kanhere, “How Mobility Patterns Drive Disease Spread: A Case Study Using Public Transit Passenger Card Travel Data”, in *2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Washington, DC, USA: IEEE, Jun. 2019, pp. 1–6, ISBN: 978-1-72810-270-2. DOI: 10.1109/WoWMoM.

2019.8793018. [Online]. Available: https://ieeexplore.ieee.org/document/8793018/ (visited on 06/17/2020).

[82] B Mo, K Feng, Y Shen, C Tam, D Li, Y Yin, and J Zhao, "Modeling Epidemic Spreading through Public Transit using Time-Varying Encounter Network", *arXiv:2004.04602 [physics, q-bio]*, Apr. 2020. [Online]. Available: http://arxiv.org/abs/2004.04602 (visited on 06/30/2020).

[83] V Amati, A Lomi, and A Mira, "Social Network Modeling", *Annual Review of Statistics and Its Application*, vol. 5, no. 1, pp. 343–369, Mar. 2018, ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-031017-100746. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-statistics-031017-100746 (visited on 07/31/2020).

[84] R Toivonen, L Kovanen, M Kivelä, JP Onnela, J Saramäki, and K Kaski, "A comparative study of social network models: Network evolution models and nodal attribute models", *Social Networks*, vol. 31, no. 4, pp. 240–254, Oct. 2009, ISSN: 03788733. DOI: 10.1016/j.socnet.2009.06.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378873309000331 (visited on 07/31/2020).

[85] K Deb, A Pratap, S Agarwal, and T Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, Apr. 2002, ISSN: 1089778X. DOI: 10.1109/4235.996017. [Online]. Available: http://ieeexplore.ieee.org/document/996017/ (visited on 06/16/2020).

[86] R Pastor-Satorras and A Vespignani, "Epidemic dynamics in finite size scale-free networks", *Physical Review E*, vol. 65, no. 3, p. 035 108, Mar. 2002, ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.65.035108. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.65.035108 (visited on 07/21/2020).

[87] JCT Gutierrez, DS Adamatti, and JM Bravo, "A new stopping criterion for multi-objective evolutionary algorithms: Application in the calibration of a hydrologic model", *Computational Geosciences*, vol. 23, no. 6, pp. 1219–1235, Dec. 2019, ISSN: 1420-0597, 1573-1499. DOI: 10.1007/s10596-019-09870-3. [Online].

Available: http://link.springer.com/10.1007/s10596-019-09870-3 (visited on 06/30/2020).

[88] S Hoover, *Methodology to Validate Accuracy of Automated Systems Using Cell Phones*, Conference on Performance and Data in Transportation Decision Making, Atlanta, Georgia, Sep. 2019.

[89] MR Alam, I Gerostathopoulos, S Amini, C Prehofer, and A Attanasi, "Adaptable Anomaly Detection in Traffic Flow Time Series", in *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Cracow, Poland: IEEE, Jun. 2019, pp. 1–9, ISBN: 978-1-5386-9484-8. DOI: 10.1109/MTITS.2019.8883338. [Online]. Available: https://ieeexplore.ieee.org/document/8883338/ (visited on 02/17/2021).

[90] VM Megler, K Tufte, and D Maier, "Improving Data Quality in Intelligent Transportation Systems", *arXiv:1602.03100 [cs]*, Feb. 2016, arXiv: 1602.03100. [Online]. Available: http://arxiv.org/abs/1602.03100 (visited on 02/17/2021).

[91] M Riveiro, M Lebram, and M Elmer, "Anomaly Detection for Road Traffic: A Visual Analytics Framework", *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2260–2270, Aug. 2017, ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2017.2675710. [Online]. Available: http://ieeexplore.ieee.org/document/7887700/ (visited on 02/17/2021).

[92] C Liu, M Zhao, A Sharma, and S Sarkar, "Traffic Dynamics Exploration and Incident Detection Using Spatiotemporal Graphical Modeling", *Journal of Big Data Analytics in Transportation*, vol. 1, no. 1, pp. 37–55, Jun. 2019, ISSN: 2523-3556, 2523-3564. DOI: 10.1007/s42421-019-00003-x. [Online]. Available: http://link.springer.com/10.1007/s42421-019-00003-x (visited on 02/17/2021).

[93] Y Feng, Y Yuan, and X Lu, "Deep Representation for Abnormal Event Detection in Crowded Scenes", in *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, Amsterdam, The Netherlands: ACM Press, 2016, pp. 591–595, ISBN: 978-1-4503-3603-1. DOI: 10.1145/2964284.2967290. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2964284.2967290 (visited on 06/11/2019).

[94] A Taylor, S Leblanc, and N Japkowicz, "Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks", in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Montreal, QC, Canada: IEEE, Oct. 2016, pp. 130–139, ISBN: 978-1-5090-5206-6. DOI: 10.1109/DSAA.2016.20. [Online]. Available: http://ieeexplore.ieee.org/document/7796898/ (visited on 06/11/2019).

[95] R Vinayakumar, K Soman, and P Poornachandran, "Long short-term memory based operation log anomaly detection", pp. 236–242, 2017. DOI: 10.1109/icacci.2017.8125846.

[96] J Chung, C Gulcehre, K Cho, and Y Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", *arXiv:1412.3555 [cs]*, Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.3555 (visited on 09/01/2019).

[97] L Böcker, M Dijst, and J Prillwitz, "Impact of Everyday Weather on Individual Daily Travel Behaviours in Perspective: A Literature Review", *Transport Reviews*, vol. 33, no. 1, pp. 71–91, Jan. 2013, ISSN: 0144-1647, 1464-5327. DOI: 10.1080/01441647.2012.747114. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01441647.2012.747114 (visited on 09/01/2019).

[98] *Overview | Distance Matrix API*, Feb. 2021. [Online]. Available: https://developers.google.com/maps/documentation/distance-matrix/overview (visited on 03/08/2021).

[99] National Weather Service, *Daily Temps/Precip/Snowfall: Portland (1940-2019, csv file)*. [Online]. Available: http://www.weather.gov/source/pqr/climate/webdata/Portland_dailyclimatedata.csv (visited on 02/17/2021).

[100] A Nasir, N Gharib, and H Jaafar, "Automatic Passenger Counting System Using Image Processing Based on Skin Colour Detection Approach", *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, pp. 1–8, 2018. DOI: 10.1109/icassda.2018.8477628.

[101] M Dunlap, Z Li, K Henrickson, and Y Wang, "Estimation of Origin and Destination Information from Bluetooth and Wi-Fi Sensing for Transit", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2595, no. 1,

pp. 11–17, Jan. 2016, ISSN: 0361-1981, 2169-4052. DOI: 10.3141/2595-02. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2595-02 (visited on 02/19/2020).

[102] Transit Cooperative Research Program, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine, *Using Archived AVL-APC Data to Improve Transit Performance and Management*. Washington, D.C.: Transportation Research Board, Sep. 2006, ISBN: 978-0-309-09861-8. DOI: 10.17226/13907. [Online]. Available: http://www.nap.edu/catalog/13907 (visited on 06/13/2019).

[103] DK Boyle, Transportation Research Board, Transit Cooperative Research Program Synthesis Program, and Transportation Research Board, *Passenger Counting Systems*. Washington, D.C.: National Academies Press, Jan. 2009, ISBN: 978-0-309-27974-1. DOI: 10.17226/14207. [Online]. Available: http://www.nap.edu/catalog/14207 (visited on 06/13/2019).

[104] *StreetLight Data: Who We Are*. [Online]. Available: https://www.streetlightdata.com/who-we-are-streetlight-data/ (visited on 05/10/2021).

[105] *Replica: Methodology*. [Online]. Available: https://replicahq.com/methodology (visited on 05/10/2021).

[106] *Remix | About us*. [Online]. Available: https://www.remix.com/about-us (visited on 05/10/2021).